Deep Learning–Based Building Attribute Estimation from Google Street View Images for Flood Risk Assessment Using Feature Fusion and Task Relation Encoding

Fu-Chen Chen[1]; Abhishek Subedi[2]; Mohammad R. Jahanshahi, A.M.ASCE[3]; David R. Johnson[4]; and Edward J. Delp[5]

Abstract: Floods are the most common and damaging natural disaster worldwide in terms of both economic losses and human casualties. Currently, policymakers rely on data collected through labor-intensive and, consequently, expensive street-level surveys to assess flood risks. We propose a laborless and financially feasible alternative: a framework that can effectively and efficiently collect building attribute data without manual street surveys. By utilizing deep learning, the proposed framework analyzes Google Street View (GSV) images to estimate multiple attributes of buildings simultaneously—including foundation height, foundation type, building type, and number of stories—that are necessary for assessing flood risks. The proposed framework achieves a 0.177-m mean absolute error (MAE) for foundation height estimation and classification F1 scores of 77.96% for foundation type, 83.12% for building type, and 94.60% for building stories, and requires less than five days to predict the attributes of 0.8 million buildings in coastal Louisiana. **DOI: 10.1061/(ASCE)CP.1943-5487.0001025.**

Author keywords: Deep learning; Object recognition; Feature fusion; Multitask learning; Flood risk assessment.

Introduction and Motivation

Floods are a very common natural disaster, occurring worldwide and causing economic losses and human casualties. Expected climate changes over the next century, including sea level rise (Michael 2007; Neumann et al. 2015), more frequent extreme precipitation events (Meehl et al. 2005; Lehmann et al. 2015), and more intense cyclone activity (Emanuel 2005; Landsea et al. 2006), pose existential threats to coastal cities hosting the large majority of human life and activity (Hallegatte et al. 2013). Governments are working to adapt to the changing environment by investing in large-scale risk mitigation. For example, the State of Louisiana in the United States has produced its *Comprehensive Master Plan for a Sustainable Coast*, a fifty-year, legislatively-mandated plan consisting of approximately $50 billion in coastal protection and restoration projects (Louisiana Coastal Protection and Restoration Authority 2012a, b). In coastal areas, governments, individual homeowners, landlords, and businesses all need accurate information about current and future flood risk to make effective decisions about risk mitigation. Decision makers such as state and federal governments may have the resources to invest in large-scale protection projects (e.g., levees, floodwalls, and pumps) that alter the local probability distribution of flood depths, but others generally do not. However, homeowners can still reduce their risk through measures such as elevation-in-place that raise the home's foundation. In this paper, we focus on identifying building attributes that provide individuals with the decision support they need to make informed, cost-effective decisions about risk mitigation of their own properties (Kellens et al. 2013).

Damage calculations in the Coastal Louisiana Risk Assessment (CLARA) model primarily follow methods developed for the FEMA Hazus Multi-Hazard model (Hazus-MH) (Scawthorn et al. 2006; Johnson et al. 2013; Fischbach et al. 2017). Direct economic losses associated with flooding are calculated as a function of a multitude of variables, such as the building's replacement cost, depth of flooding relative to the building's first-floor elevation above grade, number of stories, foundation type, and building type. The last three characteristics can be collectively referred to as building characteristics.

Assume a structure of building characteristics i with size s and construction quality q is being retrofitted to elevate its foundation to a height of h feet above the current level. Moreover, assume that the elevation of flooding relative to the top of the building's foundation is e. Then, the depth damage function for the building can be denoted as $Di(e)$, the replacement cost can be denoted as $V(s, q)$, and the probability distribution function of flood elevations occurring in a given year can be denoted as $f(e)$, where each of these depends on the variables inside their respective parenthesis. In the CLARA model, the depth damage function is expressed as the proportion of the structure's replacement cost incurred as damage to repair or reconstruct the building after a flood event. Elevating the structure directly reduces the effective flood depth experienced in comparison. Therefore, the annual losses are

[1]School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN 47907. Email: chen1623@purdue.edu

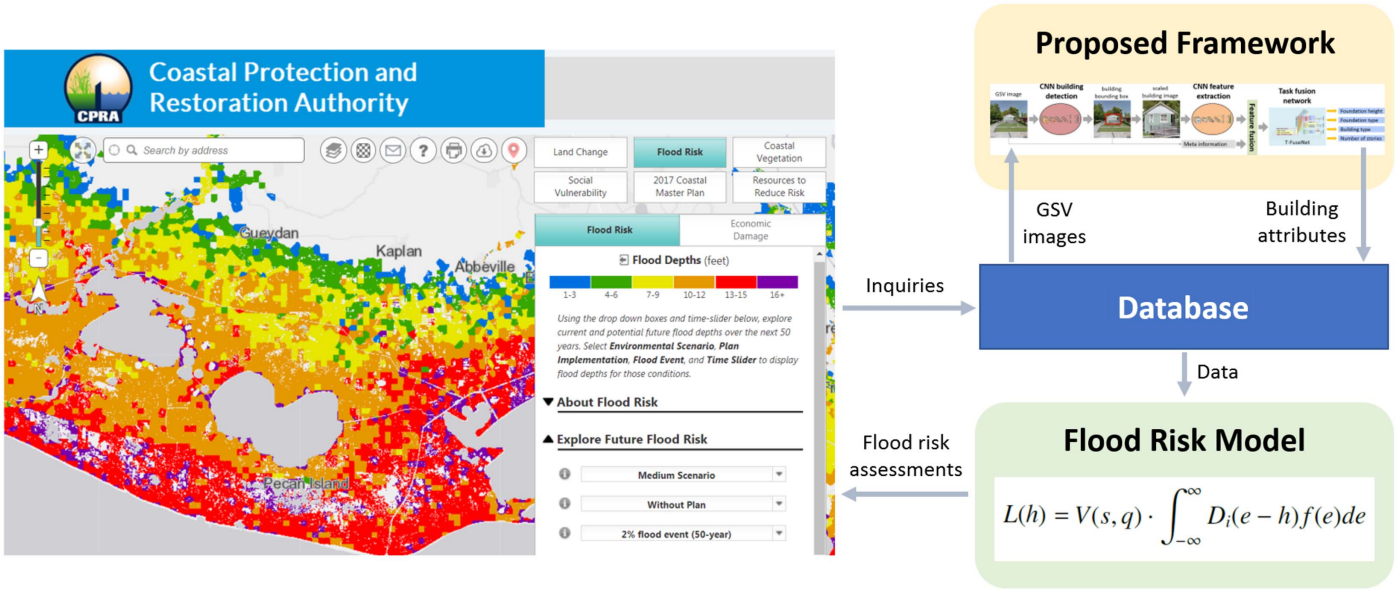
[2]Lyles School of Civil Engineering, Purdue Univ., West Lafayette, IN 47907. ORCID: https://orcid.org/0000-0003-1543-9633. Email: asubedi@purdue.edu

[3]Associate Professor, Lyles School of Civil Engineering, and the School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN 47907 (corresponding author). ORCID: https://orcid.org/0000-0001-6583-3087. Email: jahansha@purdue.edu

[4]Assistant Professor, School of Industrial Engineering, and Dept. of Political Science, Purdue Univ., West Lafayette, IN 47907. ORCID: https://orcid.org/0000-0002-2364-340X. Email: davidjohnson@purdue.edu

[5]Professor, School of Electrical and Computer Engineering, Purdue Univ., West Lafayette, IN 47907.

Note. This manuscript was submitted on April 30, 2021; approved on January 29, 2022; published online on August 25, 2022. Discussion period open until January 25, 2023; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Computing in Civil Engineering***, © ASCE, ISSN 0887-3801.



Images of buildings are from the Google Street View database courtesy of Google Inc.

Fig. 1. Schematic for integrating proposed deep learning framework for building attribute recognition into CPRA's existing online flood risk decision support system. (Reprinted with permission from USGS.)

$$L(h) = V(s,q) \cdot \int_{-\infty}^{\infty} D_i(e) f(e+h) de = V(s,q) \cdot \int_{-\infty}^{\infty} D_i(e-h) f(e) de \quad (1)$$

The damage curve $D_i(e)$ can vary substantially depending on the building characteristics, with building type being the most important driver of variability. For example, 0.3-m of inundation above the first-floor elevation is estimated to cause structural damage equal to 48% of the replacement cost for an average commercial building but 68% for a manufactured home. Damage to a single-family home on a slab foundation is 56% if one story and 44% if two stories; the respective values are 62% or 54% if the home is on a pier foundation (Scawthorn et al. 2006).

The set of structural attributes relevant to this calculation guided the selection of features to estimate in this study. Such data are expensive to collect manually and, as a result, seldom up to date. For instance, in large parts of the Louisiana coast, where significant efforts have been made toward studying flood risk since Hurricane Katrina struck in 2005, the most recent data on the height of building foundations above grade were from street-level surveys performed by the USACE in 1991 (USACE 2009). Frequent reconstructions and retrofits in the three decades since then have made it obsolete, but the state has no better estimates to rely on for making investment decisions about coastal protection measures in some areas. A small number of building-level data collection efforts utilizing post-Katrina reconstruction and tax records have somewhat improved the estimates of structural features in several Louisiana parishes; however, the coverage of high quality data is far from complete. In other states and, in particular, developing countries, individual structure-level data either do not exist or are scattered across multiple agencies and jurisdictions, making them prohibitively expensive and time-consuming to collect.

In recent years, deep learning (LeCun et al. 2015) techniques have helped engineering researchers improve solutions for a variety of problems, and the improvements have usually been significant (Cheng et al. 2021; Bao et al. 2019). To better tackle the grand challenge of flood risk management, this study proposes a deep-learning-based framework that can collect structure-level building attribute data in a more effective and efficient manner than the current approach of performing manual street-level surveys. First, the Google Street View (GSV) images of buildings in the areas of interest are gathered. Then, the proposed framework uses them to simultaneously estimate multiple structural attributes that are crucial for assessing flood risk. Consequently, combining the estimated structural attributes with geospatial data and flood risk models (e.g., CLARA model) will directly improve flood risk assessments for the areas of interest. Fig. 1 shows one view of an online flood risk decision support system developed by Louisiana's Coastal Protection and Restoration Authority (CPRA) that visualizes flood risk information for public individuals and businesses in the coastal zone in which the proposed framework will analyze GSV images and estimate risk-relevant structural attributes.

The estimated attributes include each building's foundation height, foundation type (pier, slab-on-grade, mobile home, or other), building type (commercial, residential, or mobile home), and number of stories (one story or more). Fig. 2 shows sample GSV images of typical buildings displaying these attributes.

By integrating the proposed framework into Louisiana's master planning and community resilience programs, individual homeowners can directly benefit from the "hyper-local" flood risk assessments by informing risk mitigation decisions. This will be especially true in vulnerable communities that do not currently have structural protection provided by levee and floodwall systems. Although this study focuses on predicting building attributes for managing flood risk, the proposed framework can be extended to predict different attributes for other hazards, including hurricane, tornado, or seismic hazards.

This study strives to first model the building attribute estimation problem as a multiobjective one, that is, by using one model for predicting multiple attributes at once. Then, to improve the performance of this multiobjective model, this study proposes the addition of metadata and the encoding of known relationships between the different building attributes. Not all of the building attributes may have known relationships or relevant metadata readily available; however, those necessary for assessing flood risks usually do.

Fig. 2. Sample GSV images of typical buildings displaying: (a) high foundation heights; (b) pier foundations; (c) slab foundations; (d) mobile homes; (e) commercial buildings; and (f) two or more stories. (Images © 2017 Google.)

For instance, piers are one of the most prevalent types of foundations in coastal regions, and we know that they are usually taller than slab foundations. Because both foundation height and foundation type are extremely important for the assessment and mitigation of flood risks, the knowledge that pier foundations are usually taller can be included in an automated building attribute extraction scheme to allow the modeled probabilities of the foundation type and height estimates to inform each other and improve overall performance.

Similarly, some information relevant to the estimation of attributes for flood risk management is not present in an image. For instance, a building's actual scale is vital when predicting foundation height. However, using an object detector followed by scaling the image to a fixed size masks some of this information from the models. Meta information, such as the size of bounding boxes, distance between the camera and the building, building aspect ratio, and others, reintroduces some of the lost information and helps predict the attributes that are dependent on them. Adding meta information when predicting building attributes also has other advantages. Sometimes, they directly capture specific relationships between the tasks. For instance, commercial buildings in some neighborhoods tend to have shorter building-camera distances than residential buildings because the latter are usually farther from the street, and buildings with lower aspect ratios have a higher probability of being multistoried than those with higher aspect ratios. Being judicious in the selection of meta information that either characterizes or differentiates the relevant classes can significantly improve the estimation of building attributes relevant to flood risk estimation. Our study shows that this combination of multitask learning (MTL), relationship encoding, and feature fusion can improve the performance of building attribute prediction models.

Related Works

Image Classification and Object Detection

Image classification (Haralick et al. 1973) and object detection (Viola et al. 2001) have been two popular computer vision research topics in recent decades. The former focuses on classifying the content of images, whereas the latter identifies, localizes, and categorizes the objects in images. Recently, deep learning (LeCun et al. 2015) has dominated computer vision research fields by using convolutional neural networks (CNNs) (Krizhevsky et al. 2012; Szegedy et al. 2015). Unlike traditional approaches that extract "engineered" features from images, CNNs can learn representative features from training data and achieve improved accuracies.

Several CNN architectures have been developed to improve classification accuracies for the ImageNet dataset (Deng et al. 2009). The most popular among them have been AlexNet (Krizhevsky et al. 2012), VGG-16/VGG-19 (Simonyan and Zisserman 2014), ResNets (He et al. 2016b, a), and Inception style networks (Szegedy et al. 2015; Ioffe and Szegedy 2015; Szegedy et al. 2017, 2016). After a CNN is well-trained on the ImageNet dataset, it can be reused for other applications via transfer learning, for which the variable weights in CNN are fine-tuned from the original network on a different dataset. For instance, ImageNet pretrained VGG networks (Simonyan and Zisserman 2014) were fine-tuned on relatively small datasets for pavement distress detection (Gopalakrishnan et al. 2017) and structural damage recognition (Gao and Mosalam 2018). Although these CNNs have achieved human-level performances in a wide variety of classification and recognition studies, estimating or quantifying objects' physical attributes (e.g., building characteristics or foundation heights in this study) has not been well explored. In this case, the problem is not only classification but also regression to predict attribute values from objects embedded within images.

Before estimating buildings' attributes, they first need to be detected. Several approaches have been proposed to improve object detection precision and processing speed for the COCO and ILSVRC datasets. In general, every approach can pair up with any CNN architecture. Bounding box regression for detection and class probabilities for categorization have been trained simultaneously by MTL (Ruder 2017). Approaches such as faster-RCNN (Ren et al. 2015), R-FCN (Dai et al. 2016), and particularly SSD (Liu et al. 2016) and YOLO (Redmon et al. 2016) for real-time detection have been used on the COCO and the ILSVRC datasets. Unfortunately, the object categories in the COCO or ILSVRC datasets do not include buildings that could be used in this study. Thus, the CNNs pretrained from these datasets could not be directly used, and the detection precision and processing speed of different approaches need to be analyzed for detecting buildings.

Multitask Learning

MTL (Ruder 2017) is a technique used to train a machine learning model on multiple tasks (e.g., predicting multiple building attributes) at the same time, for which the tasks share the same feature layers. Over the years, several civil engineering works have used MTL to improve performance, training time, and inference time. Most recently, Cai et al. (2021) estimated the visual focus of attention of construction workers through noisy low-resolution photographs. They reported that formulating their solution as an MTL increased prediction accuracies on the two harder categories (body orientation and head yaw) by 2% each. Furthermore, they claimed the deduction of training and inference times by approximately 50% and 35%, respectively. Zhang et al. (2020) compared the performances of seven state-of-the-art single task learning (STL) models and one MTL model to estimate traffic speeds at twenty-four links in a road network. Their rationale for employing MTL was that traffic speeds at different road links within the same road network share intricate inter-dependencies, which an MTL setup is more suitable to exploit, as they proved through their experiments.

Wan and Ni (2019) applied MTL using Bayesian modeling with Gaussian prior for reconstructing missing structural health monitoring data and tested their approach on the acceleration and temperature measurements taken at Canton Tower in China. Through a quantitative study, they found that a higher correlation between input variables resulted in the better performance of an MTL model. Furthermore, they compared the performance of the MTL model with an identical STL model and found that the MTL model outperformed its counterpart by an overwhelming margin. Hoskere et al. (2020) used a segmentation MTL model developed by Kendall et al. (2018) to semantically segment material and fine and coarse damage types. They found that MTL performed better than STL for the material category but did not find any noticeable difference between the performance of the damage types. Their conclusion was that the effective number of data samples in an MTL setting is increased for the more difficult categories because of information sharing, which further improves their results.

As is shown, MTL research in civil engineering has mostly been about understanding the phenomenon of MTL, identifying where it works and where it does not and determining its advantages and disadvantages. Our work differs from these multitask studies in two respects: (1) we primarily start with the assumption, which we subsequently validate, that MTL is better for our purpose than single-task learning, and (2) we move two steps beyond MTL by studying the effect of additional types of data and relationship encoding in a MTL setup.

Use of GSV Imagery for Urban Analysis

GSV images have been previously used to extract information from the built environment. In 2011, Rundle et al. (2011) estimated the efficacy of auditing the physical environment through GSV images, albeit manually. More recently, Zou and Wang (2021) detected abandoned houses in rust belt cities using GSV images and a hierarchical deep learning approach—they extracted both local and global details using three CNNs and used a decision tree to process the results. Their performance (F1 score of 0.84) was quite good considering the difficulty of the task. Yu et al. (2020) devised a deep learning-based method to identify soft-story buildings using GSV images. Although their results showed the promise of GSV combined with deep learning for soft story detection, one of the main contributions of their work was their breakdown of the problems that their model faced, especially that GSV images can be noisy, and an object detector can be used to filter out these noises.

Li et al. (2018) used a combination of deep learning and support vector regression to estimate the building age from GSV images collected in Victoria, Australia. They found that deeper networks usually perform better than shallow ones. Maniat et al. (2021) collected pavement images using GSV and trained a two-step CNN classifier for pavement assessment. The first classifier was used to differentiate image patches with cracks from image patches without cracks, and the second was used to classify cracks into different categories.

Alipour and Harris (2020) used GSV images (along with images collected from other sources on the Internet) to train a ten-class street damage classifier. They used semisupervised learning to annotate their GSV images, and the increase in data because of it yielded them a 20% boost in accuracy across categories. As is shown, GSV images have been used to automatically extract information about the built environment. GSV imagery databases contain a vast plethora of information-rich images of various types of buildings taken from various perspectives and for absolutely no cost. However, the metadata associated with the GSV imagery database are seldom viewed as capable of informing the images when making building estimations. In this study, we not only use the metadata themselves but also form added functions of the metadata to carefully construct features that may further improve performance in our feature fusion pipeline.

Building Attribute Estimation via Machine and Deep Learning Methods

Several studies have attempted the estimation of building attributes through machine and deep learning methods. Qi et al. (2016) used Google Earth images and celestial geometry to estimate building heights. Mou and Zhu (2018) proposed a convolutional-deconvolutional neural network architecture with skip connections to output height maps from aerial images. Liu et al. (2020) proposed another convolutional-deconvolutional neural network architecture for building height estimation and fused image data with Lidar data to improve the quality of their training set. Xie and Zhou (2017) suggested an extended multiresolution segmentation and soft back-propagation classification approach to classify building types from aerial images in urban locations. Huang et al. (2017b) used aerial images and Lidar data to predict building types. However, one major limitation of using aerial images for estimating building types and, to a certain extent, building height is that inferring their value from a building's surrounding and roof is almost always inferior to inferring them from street-level images, which are much richer in information specific to height and type prediction.

Iannelli and Dell'Acqua (2017) used GSV images to predict the number of floors using a CNN-based approach to enable the information to be used as a proxy in exposure models. Gonzalez et al. (2020) collected and annotated images in Medellin to detect lateral load resisting systems of buildings and their materials to build exposure models during earthquakes. They also applied a data fusion scheme by including the number of story data along with image data in a multimodal-styled architecture. Pi et al. (2020) used object detection models to identify objects in aerial images for better disaster response and recovery. They used hurricane videos for this study. One of their conclusions was that pretraining is important for improved performance, and that pretraining done from a different altitude/perspective can alter the performance, which constricts its widespread applicability.

Hoffmann et al. (2019) proposed an ensemble approach to building type classification that classified aerial and street-level images separately using two models, and their respective prediction scores are combined to arrive at the final prediction. Similarly, Li et al. (2017) exploited the vantage point that GSV images provide to classify building-block level land use. Kang et al. (2018) did the same but using a CNN-based approach. Lenjani et al. (2020) used a posthurricane preliminary survey as a test bed for their general approach to postdisaster reconnaissance with two streams of information extraction: postevent and pre-event. The pre-event stream used three classifiers for three building attributes. In fact, the total number of classifiers in their pre-event stream, which could be regarded as comparable with our study, was three.

Customized versions of all of these studies could be easily adapted for estimating any building attribute, including those associated with flood risks. However, most works requiring building data demand more than one attribute. Similarly, that machine/deep learning is a costly endeavor is no secret. If we use as many models as there are relevant building attributes, the costs may become exceptionally high because most agencies, such as local governments, that use building attribute data typically have a dearth of financial resources. We propose a less expensive route—an end-to-end high-performing model for predicting the most important building attributes relevant for an application at once.

Flood Risk Assessment and Response

Flood risk assessment and response consists of two aspects: (1) estimating the frequency and magnitude of floods, and (2) assessing the vulnerability of the built environment against them. The former is a science in itself (Hall and Howell 1963; Hailegeorgis and Alfredsen 2017; Stamataki and Kjeldsen 2021); however, the latter has been met with markedly less enthusiasm (Wright 2015). Our study focuses on the latter.

One of the main components of built environment vulnerability assessment from floods is estimating the potential losses that could be incurred due to their effects on buildings. As briefly touched on in the introduction section, the collection of such data has been traditionally done through manual street-level surveys and has not seen any real progress for several years. For instance, in a 2016 case study (Li et al. 2016) on flood risk assessment at a Chinese town, building data were collected through field surveys. Examples such these are abundant in the literature (Dall'Osso et al. 2009; Laudan et al. 2017; D'Ayala et al. 2020; Sen et al. 2021). Some studies have relied on data retrieved through public and private agencies (McGrath et al. 2014; Pinelli et al. 2018), but the manual labor required to collect such data on a regional scale has resulted in their sparse coverage, even in high-risk places such as Florida. According to Pinelli et al. (2018), building data in the Florida Public Hurricane Loss Model (FPHLM) were taken from the National Flood Insurance Program(NFIP), private insurance records, and county tax appraiser databases. Their study stated that 97% of structures in the NFIP database did not contain attribute-level information, only some private insurance companies were able to provide extensive building attribute data, and the tax appraiser database for the most populated county in Florida (Miami-Dade) contained virtually no information relevant to FPHLM.

Although manual field surveys are still the norm in building attribute data collection for flood vulnerability assessment, some researchers have begun to use more efficient methodologies. In a 2020 study (Arrighi et al. 2020), the authors used GSV images as a virtual environment to quantify several building characteristics; however, their process was still manual and expensive for large areas. Similarly, a mobile- and Internet-based attribute data collection methodology in which building owners act as contributors has been proposed as well (Valenzuela et al. 2016). However, the efficacy of such an approach is questionable because it relies on people volunteering to disclose information without incentive. In addition, cellular and internet reach is not pervasive in many rural areas of developing nations. We collect GSV images containing buildings, use an object detector to identify buildings in those images, and automatically extract relevant building attributes, saving the need to rely on field surveys or potentially incomplete historical records. Similarly, our method can easily scale to large geographical regions and is not based on crowdsourcing.

Contribution

We propose a MTL approach for predicting multiple building attributes at once for flood risk assessment. To the best of our knowledge, this study is the first to do so and the first to estimate foundation type and foundation height as a regression problem from GSV images. MTL, in addition to being cheap, also improves performance by allowing the transference of knowledge across tasks (Thrun 1996; Caruana 1997). Our study shows that this relatively overlooked area of research should be perceived as a viable candidate for building attribute estimation because of its potential to provide accurate predictions at a proverbial penny on the dollar when compared with single-task learning approaches. Similarly, to further increase the performance of MTL models, we propose the two methods of Task Relation Encoding Network (TREncNet) and feature fusion. Traditionally, each task in MTL has independent,

fully connected layers. Recent studies (Meyerson and Miikkulainen 2017; Ma et al. 2018) have shown that the implicit relations among tasks can be learned by soft layer ordering or multiple gating with task-specific parameters. However, the case of tasks having explicit relations (i.e., relationships we have prior knowledge about) has been under-discussed. We encode such known relationships through an architectural modification that we call TREncNet. Moreover, we use another technique called feature fusion, which introduces metadata, such as camera-to-building distances and aspect ratios, into the MTL network. By combining MTL, TREncNet, and feature fusion, we report the best overall prediction scores among three separate single-task learning and four separate plain MTL experiments conducted using state-of-the-art architectures. Given its capabilities, we conclude that the proposed approach can effectively and efficiently process comprehensive data without using street surveys, which will save time and money for flood risk management. Finally, we provide an extensive evaluation of different CNN architectures for building attribute prediction and object detection, which provides practitioners a starting point when implementing this framework.

Scope

The remainder of this paper is organized as follows. The section on "Building Dataset Generation" describes the dataset of buildings collected for training and evaluation. The section on "Proposed Framework" elaborates on the details of the proposed framework. The section on "Experimental Results" discusses the evaluation results. The section on "Error Analysis" discusses why some errors might have occurred. The section on "Key Findings and Potential Implications" provides a detailed breakdown of the results from our proposed approaches and discusses some implications of this study. The section on "Limitations and Future Work" lists some possible future studies that can build on this work, and the "Conclusion" summarizes the paper.

Building Dataset Generation

To train the CNNs in the proposed framework and validate the estimation performance, ground-truth building attributes along with the corresponding GSV images need to be collected. Several field surveys have been conducted in the post-Katrina coastal Louisiana that collected the attributes and the GPS coordinates of 80,109 buildings. For 73,781 buildings, the records had all of the desired attribute information, and the other 6,328 records had only foundation height information. This building-level data primarily originated from three studies performed by the USACE: the Morganza to the Gulf Reformulation study (noa, a), Southwest Coastal Louisiana Feasibility study (noa, b), and West Shore Lake Pontchartrain Feasibility study (noa, c). Coverage includes part or all of Calcasieu, Cameron, Iberia, Jefferson Davis, Lafourche, St. Charles, St. James, St. John, and Terrebonne parishes (i.e., county-level units of governance in Louisiana). Foundation heights from FEMA Elevation Certificates for 2,471 buildings in Jefferson Parish were also obtained from a parish floodplain manager. The buildings' GSV images (640×640 resolution and 75° field of view) were autonomously extracted using GSV's application programming interface (API) and the GPS coordinates. Then, the bounding boxes of buildings in the GSV images were manually annotated, as shown in Fig. 3.

Not all of the GSV images had usable views of buildings. Sometimes, major parts of the buildings were blocked by objects (e.g., fences, trees, or cars), or the buildings appeared too small in the images (e.g., with width or height less than 80 pixels). Given the inconsistencies between the collected building coordinates and the GSV API, some images did not contain a building in the scenes, and some coordinates did not have a GSV image available. Fig. 3 also shows sample GSV images with unusable building views. Images with unacceptable views were removed from the dataset; the remaining dataset contained 42,415 usable GSV images. Fig. 4 shows the distributions of the building attributes in the training dataset. The distributions are imbalanced, and the dominant attributes with the largest prevalence are 0.15-m foundation heights (0.5-ft), concrete slab foundations, residential buildings, and one-story buildings.

Proposed Framework

Fig. 5 shows the overview of the proposed framework, consisting of the following steps. (1) "CNN building detection" detects the bounding box of a building from a GSV image. The pixels inside the bounding box are then scaled to a fixed-sized image of that building. (2) "CNN feature extraction" extracts the image feature vector from the fixed-sized image. (3) "Feature fusion" concatenates the image feature vector with the meta information vector to form a fused feature vector. (4) "Task relation encoding network" simultaneously predicts all of the building attributes from the fused feature vector, including the building's foundation height in feet above grade, foundation type (pier, slab, mobile home, or other), building type (commercial, residential, or mobile home), and number of stories (one story or more). The details of each step are explained in the following subsections, and the training of CNNs is described in the experimental results section.

CNN Building Detection

As mentioned in the literature review, several object detection approaches using deep learning have been proposed and achieved successful results for the COCO (Lin et al. 2014) and ILSVRC (Russakovsky et al. 2015) datasets. However, the object categories do not include buildings. Thus, the CNN architectures cannot be directly used, and the performance of each approach needs to be evaluated for detecting buildings. In this study, the detector's localization accuracy is important because the pixels inside the building bounding box affect the accuracy of the attribute prediction. If the bounding box is too large, the pixels include too much background area. If the bounding box is too small, some of the building's pixels will be missing. Meanwhile, the detector's speed is also a concern because a large number of buildings are in coastal areas: more than 780,000 buildings are in the coastal Louisiana study region with virtually complete GSV coverage; however, only approximately 75% were found to have GSV images usable for feature extraction. After an extensive evaluation of different object detection approaches and CNN architectures to compare the accuracy/speed trade-offs, this study chose Faster R-CNN (Ren et al. 2015) along with Inception-ResNet (Szegedy et al. 2017) to detect building bounding boxes in GSV images. The details of this analysis are explained in the experimental results section. If more than one building bounding box is detected in a GSV image, only the box with the highest detection score was kept.

CNN Feature Extraction

After the building bounding box is detected in a GSV image, the pixels in the bounding box are scaled to generate a fixed-sized image. Then, a CNN takes the scaled image (224×224 or 299×299 pixels depending on the CNN architecture) as input and extracts the image feature vector through the convolution and pooling layers. This study compared several CNNs that achieved high accuracies for the ImageNet dataset (Deng et al. 2009). The CNNs were pretrained on the Imagenet dataset and fine-tuned on the

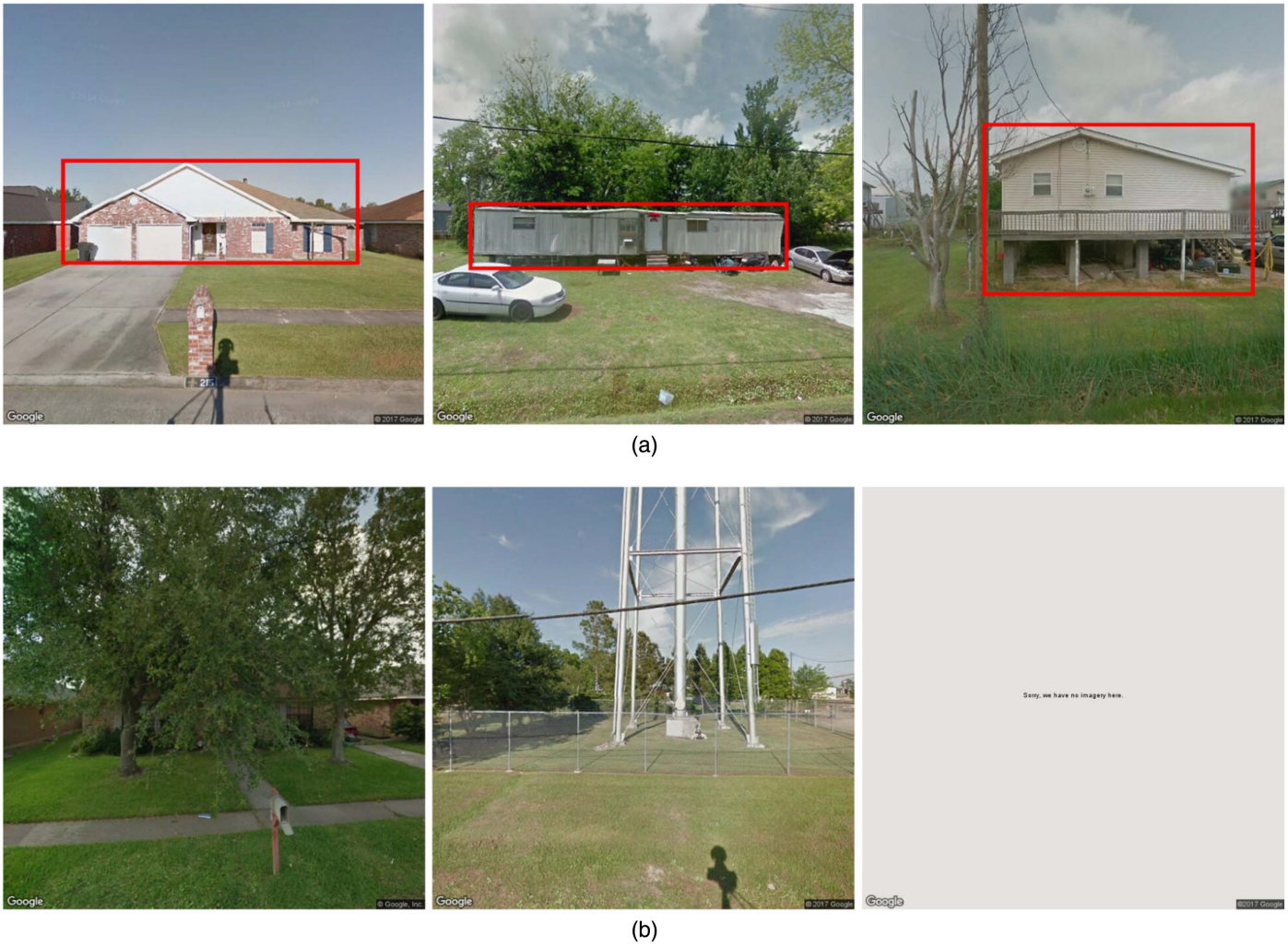


Fig. 3. Sample GSV images of buildings in coastal Louisiana areas: (a) good views with annotated building bounding boxes; and (b) unacceptable views removed from dataset for buildings blocked by front objects, no building present in the scene, buildings too small, or no GSV image available. (Images © 2017 Google; image © Google, Inc.)

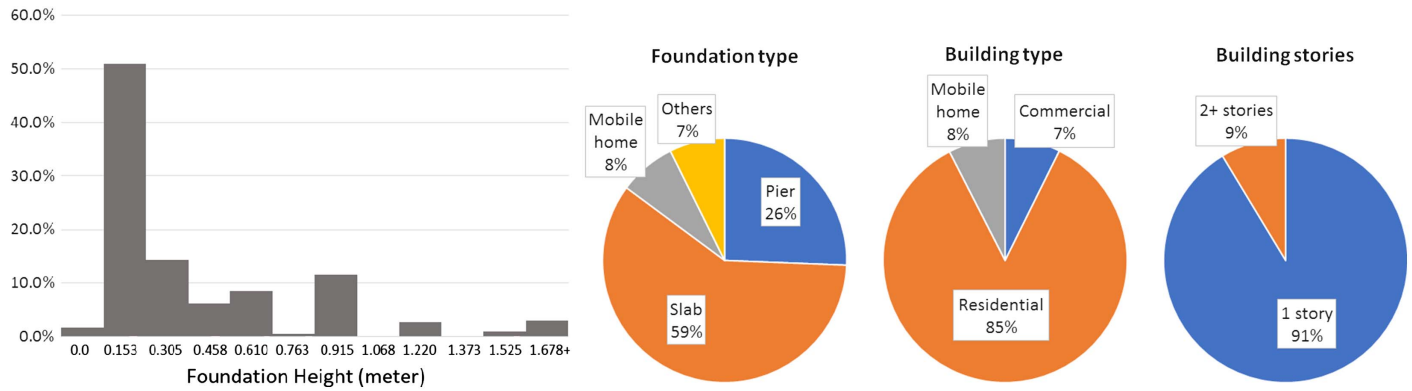


Fig. 4. Distributions of building attributes in dataset.

building attribute dataset in this study. The feature vector of each CNN was extracted from the final global pooling layer and had vector dimension of 1,024 or 1,536, depending on the CNN architecture. In this study, Inception-ResNet (Szegedy et al. 2017) was chosen to extract the image feature vector. Details about the comparison of different CNNs are described in the experimental results section.

Feature Fusion

As described in the literature review, although many studies have achieved successful results in different image classification or recognition tasks, estimating or quantifying objects' physical attributes has seldom been discussed. In addition to the use of image

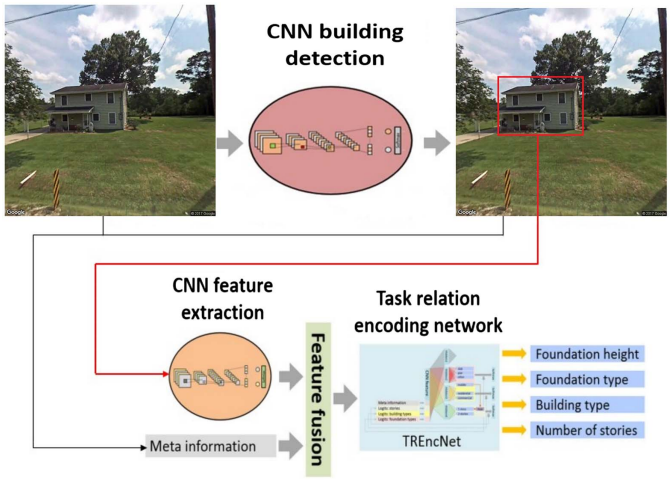


Fig. 5. Overview of proposed framework. (Images © 2017 Google.)

pixels, additional information describing objects' physical properties might improve prediction accuracies. Thus, this study proposes a feature fusion scheme that extracts the meta information of the building and concatenates the meta information vector with the image feature vector to form a fused feature vector for final attribute prediction.

The meta information consists of seven values: d, s, w_p, h_p, w_f, h_f, and r. The first value d represents camera-to-building distance in feet, which can be obtained by using the Haversine formula

$$d = 2R\arcsin \times \left(\sqrt{\sin^2\left(\frac{lat_c - lat_b}{2}\right) + \cos(lat_c)\cos(lat_b)\sin^2\left(\frac{lon_c - lon_b}{2}\right)}\right) \quad (2)$$

where R = radius of earth in feet that varies with latitude; and lat_c, lon_c, lat_b, and lon_b = latitudes and longitudes of the camera and building, respectively, as obtained through the GSV API. The second value s represents pixels per feet for the building in the image, which can be calculated by using

$$s = \frac{W}{2d\tan(\theta/2)} \quad (3)$$

where W = width of GSV image in pixels; and θ = camera's field of view ($W = 640$ and $\theta = 75°$ in this study). The third and fourth values w_p and h_p represent the building bounding box's width and height in pixels, respectively. The fifth and sixth values $w_f = w_p/s$ and $h_f = h_p/s$ represent the building bounding box's width and height in feet. The final value $r = w_p/h_p$ represents the width-to-height ratio. After obtaining all of these seven values, the normalized meta information vector is calculated as

$$\left(\frac{d}{100}, \frac{s}{10}, \frac{w_p}{640}, \frac{h_p}{640}, \frac{w_f}{100}, \frac{h_f}{100}, \frac{r}{5}\right)$$

where the constant divisors approximately normalized each value to a floating range from zero to one. Finally, the image feature vector is concatenated with the normalized meta information vector to form the fused feature vector.

Alternatively, using metadata to directly predict some attributes (e.g., foundation height) is possible. One property that can be considered is camera-building distance. However, due to the inaccuracies of distance calculation and building or camera coordinates, the camera-building distances might not be completely accurate. Fig. 6 shows sample GSV images whose estimated camera-building distances are not accurate. If the camera-building distances are directly used to predict foundation heights (e.g., predict heights in pixels and then convert heights to feet using the distances), those inaccurate distances will result in erroneous predictions. Other possible properties include the building's physical width, height, and width-height ratio. Yet, the estimations for those properties still depend on the bounding box detection or camera-building distances that are not absolutely accurate. However, when used in our feature fusion scheme and with TREncNet, even though the meta information might not be completely accurate, they are not the only source of information. Therefore, the network itself can make

Fig. 6. Samples of GSV images with inaccurate estimated camera-building distances: (a) 2.13 m; (b) 6.14 m; and (c) 43.83 m. Buildings in (a and b) should be farther from the camera, and buildings in (c) should be closer to the camera than the estimated distances. (Images © 2017 Google.)

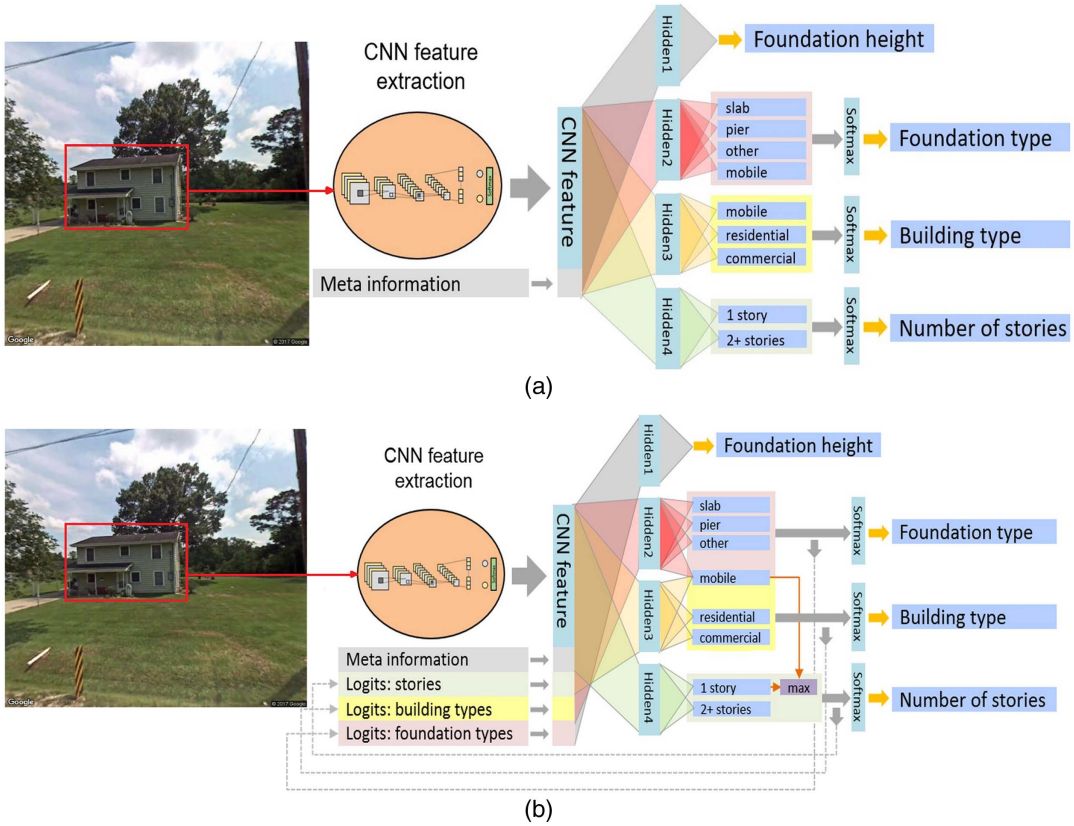


Fig. 7. (a) Traditional multitask learning that treats each task independently with separate fully connected layers; and (b) proposed TREncNet that encodes tasks' explicit and implicit relations. (Images © 2017 Google.)

determinations about their validity and how best to incorporate the information and improve the predictions. In other words, the TREncNet accounts for some of the uncertainties that exist in this problem, to some extent.

Task Relation Encoding Network

After obtaining the fused feature vector from the CNN and meta information, the final step is to simultaneously predict all of the building attributes from the feature vector on the basis of MTL (Ruder 2017). The tasks in this study include one regression for foundation height and three classifications for foundation type, building type, and number of stories. Traditionally, each task has its own fully connected layers, with one hidden layer as illustrated in Fig. 7. This study proposes TREncNet as a means of encoding the explicit and implicit relations of tasks as network connections to improve predictive accuracies.

In a multitask study, some relationships between tasks are obscured from the practitioner's view, whereas others can be more easily identified. For instance, a sophisticated nonlinear function may exist that our human perception fails to detect, but the knowledge that pier foundations usually have greater heights than slab foundations is ubiquitous. The relationships that are transparent to human perception can further be divided into two types: implicit and explicit. Implicit relationships are those that occur with less certain probabilities, whereas explicit relationships are those that occur with absolute probabilities (i.e., probabilities of one). This example that pier foundations are usually taller than slab foundations belongs to the former category, whereas the fact that mobile homes always have mobile foundations belongs to the latter category. In the context of MTL, implicit relationships can be encoded in the probabilistic space, that is, through shared learning in the layer(s) preceding the output layer to ensure that the model can appropriately learn them without us having to tell it to make definite decisions. In contrast, the explicit relationships can be exploited by setting up hard rules, such as if x is true, y is true, and z is false (x, y, and z are hypothetical variables).

As Fig. 7 shows, the knowledge that mobile foundations and mobile homes are the same tasks is encoded through their shared logit value before softmax. Another explicit relation is that a mobile home always has one story. As a result, the final logit value for "one story" is the maximum value of the original "one story" and "mobile home" logit. By doing this, if a building has a small original "one story" logit value but is classified as a "mobile home," it might still ultimately be classified as a one-story building because the logit value for "mobile home" will be large.

In this study, the implicit relationships among tasks are encoded by using the prediction results from one task to help predict other tasks. To do this, the logits from one task are concatenated with the feature vector for other tasks. Specifically, the number of stories task uses the original feature vector, the building type task uses the feature vector plus the logits from the stories task, the foundation type uses the feature vector plus the logits from both previous tasks, and the foundation height task uses the feature vector plus the logit values from all other tasks. This concatenating order is determined by the tasks' prediction difficulty (i.e., the stories task is the easiest, and the foundation height task is the most difficult). To prevent creating loops in the network, the logits for "mobile home" and "one story" are not concatenated with the feature vector.

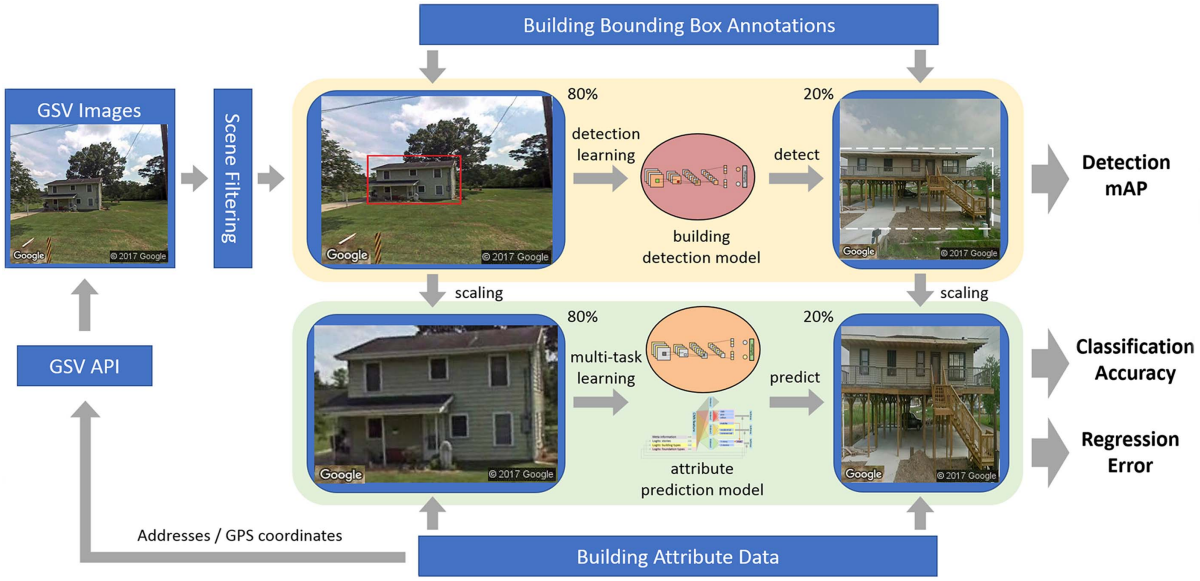


Fig. 8. Evaluation pipeline in this study. Approximately 80% (33,822) of GSV images were used to train (70% of images) and validate (10% of images) building detection and attribution prediction models. The remaining 20% (8,593) of GSV images were used to evaluate their performance. Full boxes: ground truth building bounding boxes; dashed boxes: predicted building bounding boxes. (Images © 2017 Google.)

Experimental Results

Evaluation Pipeline

Fig. 8 illustrates the overall evaluation pipeline. Using the building attribute dataset, the buildings' GSV images were collected and filtered. Then, approximately 80% (33,822) of the GSV images were randomly selected to train (70% of the images) and validate (10% of the images) building detection and attribution prediction models. During training, the validation loss was calculated and compared for every epoch. The models with the least validation losses were chosen whose performances were evaluated using the remaining 20% (8,593) of the GSV images. Mean average precision (mAP) was used to evaluate building detection models, whereas residual error for foundation height and classification accuracies for foundation type, building type, and the number of stories were used to evaluate the attribute prediction models. The evaluation took place on an Exxact deep learning Linux server with Ubuntu 16.04.3 LTS, two Intel Xeon E5-2620 v4 CPUs with a total of 32 cores, 256 GB DDR4 RAM, and four NVIDIA Titan X Pascal GPUs. One GPU at a time was used to train and evaluate.

Evaluation of the Building Detection Scheme

Detection Approaches and Training: A Tensorflow object detection API (Huang et al. 2017a) was utilized to evaluate the accuracy/speed trade-offs for detecting building bounding boxes from GSV images. The detection approaches and CNN architectures that achieve more than 24% mAP for the COCO dataset (Lin et al. 2014) were selected for performance comparison. The detection approaches included Faster R-CNN (Ren et al. 2015), R-FCN (Dai et al. 2016), and SSD (Liu et al. 2016). The CNN architectures included Inception V2 (Ioffe and Szegedy 2015), ResNet 50 and 101 (He et al. 2016a), MobileNet V1 (Howard et al. 2017), Inception-ResNet V2 (Szegedy et al. 2017), and NASNet (Zoph et al. 2018). The variable weights for each CNN were pretrained from the COCO dataset and fine-tuned using the building bounding box annotations described in the section on building dataset generation. Each training took 35 epochs using the optimized training parameters provided by Huang et al. (2017a).

Overall Performance of Building Detection Approaches: Table 1 lists the detection mAP at a different intersection over union (IoU) thresholds, training time, and inference time for different detection approaches and CNN architectures. The IoU threshold defines the size of the minimum IoU between detected and ground-truth bounding boxes. The mAP at [0.5:0.95] thresholds, which takes the average of the mAPs for IoU thresholds equal to 0.5, 0.55, ..., 0.95 (0.05 increments), is the primary challenge metric for the COCO dataset (Lin et al. 2014).

Table 1 shows that most of the detection approaches and CNN architectures achieve more than 98% mAP at 0.5 IoU threshold. Therefore, most of the buildings can be successfully detected. However, to obtain actual pixels of buildings for attribute prediction, a precise detection model that achieves high mAP at a high IoU threshold is preferred. The most precise detection model in Table 1 is Faster R-CNN (Ren et al. 2015) with Inception-ResNet V2 (Szegedy et al. 2017); that combination has the highest values of 79.6% mAP at a 0.75 threshold and 66.6% mAP at [0.5:0.95] thresholds. It takes 0.405 seconds to process a 640 × 640 GSV image with one GPU or less than four days for the 0.8 million GSV images in the study region. Thus, Faster R-CNN (Ren et al. 2015) with Inception-ResNet V2 (Szegedy et al. 2017) has a reasonable processing speed for this study, and its detection results were used to evaluate the building attribute predictions. For computation environments without GPU processing, SSD (Liu et al. 2016) with ResNet 50 (He et al. 2016a) is a feasible choice, considering its accuracy/speed trade-off. SSD achieves the second-highest 78.6% mAP at 0.75 threshold and 65.8% mAP at [0.5:0.95] thresholds. Two CPUs for a total of 32 cores take 0.528 seconds to process a GSV image (less than five days for 0.8 million images).

Evaluation of Attribute Prediction Scheme

CNN Architectures and Training. Because Inception-ResNet V2 (Szegedy et al. 2017) achieved the highest mAP for building detection, it was chosen for the "CNN feature extraction" to evaluate

Table 1. Evaluation results of different detection approaches and CNN architectures for building bounding box detection

Detection approach	CNN architecture	mAP@IoU 0.5	mAP@IoU 0.75	mAP@IoU [0.5:0.95]	Training time (day)	Inference time (s) GPU	Inference time (s) CPU
SSD (Liu et al. 2016)	MobileNet V1 (Howard et al. 2017)	98.2%	77.4%	65.1%	1.2	0.034	0.311
	Inception V2 (Ioffe and Szegedy 2015)	98.2%	77.6%	65.8%	1.1	0.026	0.165
	ResNet 50 (He et al. 2016a)	**98.4%**	78.6%	65.8%	1.4	0.046	0.528
Faster R-CNN (Ren et al. 2015)	Inception V2 (Ioffe and Szegedy 2015)	98.2%	78.5%	65.8%	1.1	0.056	0.391
	ResNet 50 (He et al. 2016a)	98.2%	77.8%	65.2%	1.7	0.109	1.338
	ResNet 101 (He et al. 2016a)	98.2%	77.8%	65.3%	2.4	0.125	1.662
	Inception-ResNet V2 (Szegedy et al. 2017)	98.3%	**79.6%**	**66.6%**	6.9	0.405	6.591
	NASNet (Zoph et al. 2018)	97.9%	78.0%	65.0%	6.9	0.305	2.965
R-FCN (Dai et al. 2016)	ResNet 101 (He et al. 2016a)	98.3%	76.3%	64.7%	2.1	0.072	0.601

Note: Highest value in each column is in bold.

the attribute prediction accuracy of the proposed framework. To show the effectiveness of the proposed feature fusion scheme and TREncNet, three other CNNs were also evaluated, including MobileNet V1 (Howard et al. 2017), Inception V2 (Ioffe and Szegedy 2015), and Inception V4 (Szegedy et al. 2017). MobileNet V1 and Inception V2 take a 224×224 scaled image as input and extract the CNN features of 1,024 dimensions for which Inception V4 and Inception-ResNet V2 take a 299×299 scaled image and extract the CNN features of 1,536 dimensions. The variable weights of CNNs were pretrained using the ImageNet dataset (Deng et al. 2009) with a TensorFlow version 1.4.0 model library (Silberman and Guadarrama 2016) and fine-tuned using the building attribute dataset described in the building dataset generation section. The architecture of TREncNet allows end-to-end (Yang et al. 2018) training for which the variable weights of CNN and TREncNet were fine-tuned together.

During fine-tuning, the loss function to be minimized included a Huber loss (Huber 1992), with $\delta = 15$ [found through a hyperparameter search (Meyer 2021)] for the regression task, and three cross-entropy values for the classification tasks. We further multiplied the regression loss by 0.25 to prevent the overall gradient from overflowing from the classifiers' perspectives because of the imbalance between the scales of the regression and classification losses in our model. If a GSV image did not have certain building attributes (e.g., some images in the training dataset only have foundation height data), the loss weights of the corresponding tasks are zero for that image. The loss function also included a regularization term that equaled the sum of the squared values for all of the variables in TREncNet with a 0.004 loss weight (found through a hyperparameter search). The number of hidden layer nodes in TREncNet was 128 for each task. To prevent overfitting, each hidden layer had a 0.5 dropout rate (Srivastava et al. 2014) during fine-tuning. The learning rate was initially 0.001 for MobileNet V1 and 0.002 for all other CNNs, with a 0.6 decay rate for every 40 epochs. Each CNN was fine-tuned for 160 epochs with a batch size of 32. The training images were randomly augmented in each batch, including horizontal flipping and $\pm 10\%$ brightness, $\pm 20\%$ contrast, $\pm 20\%$ saturation, and $\pm 2.5°$ hue adjustments.

Single versus Multi Task Learning: A detailed comparison between MTL and STL has been shown in Table 2. The evaluation metrics are mean average error (MAE) for foundation height estimation and F1 score for building type, foundation type, and number of stories classification. Of the buildings, 85% are residential, 59% have slab foundations, and 91% are single-storied. To consider the class imbalance issue, the F1 score is the harmonic mean of precision and recall. Furthermore, the percentage differences reported are absolute differences for classification tasks and relative differences for the foundation height task.

The strongest impact of choosing MTL over STL can be observed in foundation height estimation. Both plain MTL and the proposed best MTL architecture (feature fusion and TREncNet combined) perform significantly better than STL. For classification tasks, the F1 scores were averaged to simplify the analysis. There, although plain MTL sometimes results in poor estimates, the proposed MTL can be observed to significantly improve them when compared with STL.

As mentioned in the literature review, a common theme across MTL studies is that it provides significant improvements in

Table 2. Evaluation results of different CNN architectures with STL, MTL, and MTL combined with feature-fusion and TREncNet (i.e., Proposed Best MTL)

CNN architecture	STL	MTL	Proposed Best MTL	MAE (m) F. height	F1 (%) F. type	F1 (%) B. type	F1 (%) B. story	Average (%) Overall	Inference time (s) GPU	Inference time (s) CPU
MobileNet V1 (Howard et al. 2017)	Yes	—	—	0.229	75.84	80.82	94.00	83.55	0.006	0.081
	—	Yes	—	0.201	75.23	79.17	93.90	82.77		
	—	—	Yes	0.180	76.70	81.5	93.91	84.04		
Inception V4 (Szegedy et al. 2017)	Yes	—	—	0.223	77.74	82.58	94.1	84.81	0.029	0.307
	—	Yes	—	0.195	77.49	82.30	94.40	84.73		
	—	—	Yes	0.177	77.31	82.48	94.65	84.81		
Inception-ResNet V2 (Szegedy et al. 2017)	Yes	—	—	0.232	77.77	81.85	94.00	84.54	0.038	0.483
	—	Yes	—	0.192	77.97	82.48	94.09	84.85		
	—	—	Yes	0.177	77.96	83.12	94.60	85.23		

Note: MAE = mean absolute error; F. = foundation; and B. = building.

Table 3. Evaluation results of different CNN architectures without or with proposed feature fusion scheme and TREncNet for building attribute prediction

CNN architecture	Feature fusion	TREncNet	MAE (m) F. height	Precision (%) F. type	Precision (%) B. type	Precision (%) B. story	Recall (%) F. type	Recall (%) B. type	Recall (%) B. story	F1 (%) F. type	F1 (%) B. type	F1 (%) B. story	Average (%) Overall
MobileNet V1 (Howard et al. 2017)	—	—	0.201	78.3	83.7	94.5	72.4	75.1	93.3	75.23	79.17	93.9	82.77
	Yes	—	0.171	76.8	82.5	93.0	73.6	76.6	94.0	75.17	79.44	93.5	82.7
	—	Yes	0.210	78.1	83.8	94.6	76.1	80.1	93.6	77.09	81.91	94.10	84.37
	Yes	Yes	0.180	77.3	83.6	92.0	76.1	79.5	95.9	76.70	81.50	93.91	84.04
Inception V2 (Ioffe and Szegedy 2015)	—	—	0.204	77.5	84.6	93.6	73.6	76.5	93.7	75.5	80.35	93.65	83.17
	Yes	—	0.180	78.3	85.0	95.7	75.3	78.8	92.8	76.77	81.78	94.23	84.26
	—	Yes	0.201	79.7	86.1	95.5	75.4	77.7	90.3	77.49	81.68	92.83	84.00
	Yes	Yes	0.174	78.1	84.5	92.0	74.2	76.6	94.6	76.1	80.36	93.28	83.25
Inception V4 (Szegedy et al. 2017)	—	—	0.195	78.5	85.2	95.0	76.5	79.6	93.8	77.49	82.30	94.4	84.73
	Yes	—	0.177	78.2	84.4	94.8	75.5	79.0	93.2	76.83	81.61	93.99	84.14
	—	Yes	0.192	79.4	85.9	94.7	76.2	79.4	94.5	77.77	82.52	94.60	84.96
	Yes	Yes	0.177	79.1	85.8	95.3	75.6	79.4	94.0	77.31	82.48	94.65	84.81
Inception-ResNet V2 (Szegedy et al. 2017)	—	—	0.192	79.6	85.8	94.9	76.4	79.4	93.3	77.97	82.48	94.09	84.85
	Yes	—	0.177	79.8	86.6	94.1	76.1	78.8	95.0	77.91	82.52	94.55	84.99
	—	Yes	0.192	79.8	86.4	94.8	76.6	79.2	93.3	78.17	82.64	94.04	84.95
	Yes	Yes	0.177	79.7	86.5	94.7	76.3	80.0	94.5	77.96	83.12	94.60	85.23

Note: MAE = mean absolute error; F. = foundation; and B. = building.

performance in categories that are relatively difficult and does so by sharing information from other correlated tasks. In this study, we considered foundation height to be the most difficult task because it concerned the estimation of regression data from images. Using Table 3, the worst performing MTL model had an MAE of 0.210 m, whereas the best performing STL model had an MAE of 0.223 m. Similarly, the best performing MTL model had an MAE of 0.171 m. However, MTL models' performances cannot be evaluated on a per-task basis because they are not used in a real-world application on a per-task basis; therefore, the actual comparison should be done between the 0.177-m MAE of Inception ResNet V2 with feature fusion and TREncNet and the 0.223-m MAE of STL, which is a significant 20.63% reduction. Fig. 9 shows a whisker-and-box plot that provides a visual demonstration, where the boxes are wider and the extensions of the whiskers are denser for STL than for MTL. MTL models produced a much tighter prediction near and around the ground truths, where the number of samples is 4,341, 1,252, 503, 737, 995, and 245, respectively, for each of the foundation height intervals. In other words, MTL leads to smaller standard deviations of MAE, indicating the robustness of MTL compared with STL.

On a NVIDIA Titan X GPU, the inference time added up to 0.125 seconds per image for single task learning but only 0.038 seconds for the multitask model—a 70% reduction. This reduction is significant from a resource budgeting and allocation point-of-view. For instance, to process the 0.8 million images that make up our region of study, using a MTL approach would take less than nine hours but approximately 28 hours using a single-task learning approach. The latter would still yield worse estimates, especially in arguably the most important building attribute with respect to assessing flood risks i.e., foundation height. If only for gains in computational costs and foundation height estimation, MTL should be used over single task learning for the purpose of building attribute prediction in flood management.

Plain MTL versus Feature Fusion versus TREncNet versus Feature-Fusion-Plus-TREncNet: Table 3 lists the performances of MTL with or without TREncNet and feature fusion. Here as well, the metrics used are MAE for foundation height estimation and the

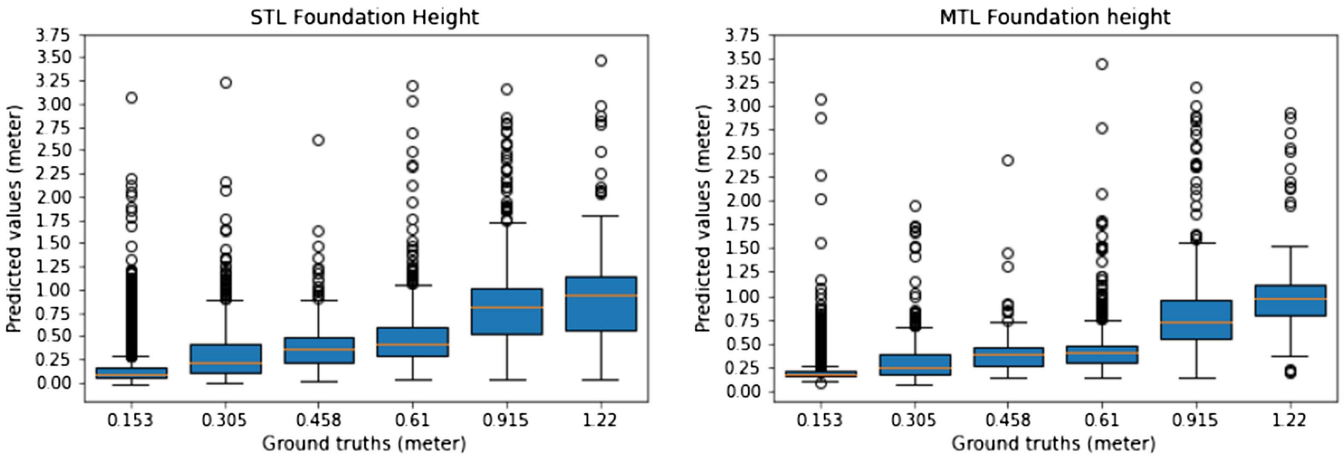


Fig. 9. STL *versus* MTL for foundation height estimation.

F1 score for foundation type, building type, and number of stories classification. Furthermore, we averaged the F1 scores from the three classification tasks.

For foundation height regression, TREncNet did not provide any significant improvement over a plain MTL, whereas feature fusion did. In MobileNet V1, the foundation height estimate of a plain MTL has 0.201-m MAE, whereas the same decreases to 0.171 m when metadata are added. TREncNet's performance is worse than plain MTL because it has an MAE of 0.210 m. Similarly, for Inception V2, feature fusion reduces MAE from 0.204 m to 0.180 m, whereas TREncNet's performance remains almost the same at 0.201 m. Inception V4 and Inception ResNet V2 show similar improvements because of feature fusion (0.195 m to 0.177 m and 0.192 m to 0.177 m, respectively), whereas TREncNet's MAE in both of those architectures is 0.192 m. In percentage terms, feature fusion reduces MAE in MobileNet V1 by 15%, in Inception V2 by 12%, in Inception V4 by 9%, and in Inception ResNet V2 by 8%.

We have treated foundation height separately because its performance cannot be conflated to an F1 score, and it can furthermore be assumed to be of greater importance than other tasks with respect to flood risk assessment. None of the architectures reported plain MTL as providing the best average F1 score, which is strong evidence that plain MTL is not recommended over our proposed alternatives for classification tasks. In fact, the only architecture that reports a significantly higher average F1 score over any of our proposed alternatives is Inception V4, which outperforms its feature fusion version. Some alternative MTLs have almost equal performances as their plain MTL counterparts, such as feature fusion of MobileNet V1, feature-fusion-plus-TREncNet of Inception V2, and feature-fusion-plus-TREncNet of Inception V4.

In summary, we cannot conclude that fusion + TREncNet is necessarily better than fusion only. Having said that, as Table 3 shows, except for MobileNet V1, for all other architectures, fusion + TREncNet provides a similar or better MAE for foundation height than feature fusion alone, whereas the precision, recall, and F1 scores are slightly better for building type, foundation type, and building story estimations. For Inception V2, although the precision, recall, and F1 scores for building type, foundation type, and building story estimations are marginally better when fusion only was used, the MAE foundation height is higher when fusion alone is used relative to fusion + TREncNet. Similarly, although the MAE foundation height estimations for Inspection-Resnet V2 with fusion + TREncNet are comparable to other combinations of architectures, fusion, and TREcNet, the F1 scores were higher for building type, foundation type, and building story estimations when Inspection-Resnet V2 with fusion + TREncNet is used. Similarly, although the MAE foundation height estimations for fusion only and fusion + TREncNet are the same for Inspection-Resnet V2, precision, recall, and F1 scores are slightly better for fusion + TREncNet. Consequently, we use Inspection-Resnet V2 with fusion + TREncNet, although one cannot conclude that fusion + TREncNet is significantly better than fusion only.

Sensitivity analysis of foundation height prediction to flood damage estimates: Fig. 10 illustrates the importance of making small improvements in the accuracy of estimates for the foundation height (i.e., first-floor elevation above grade). Using the Coastal Louisiana Risk Assessment model, we made small perturbations in the assumed first-floor elevations of all buildings at risk of storm surge-based flooding in the Louisiana coastal zone. The left pane of the figure shows the average change in the expected annual damage resulting from perturbations ranging from 10 cm below the estimated foundation height to 10 cm above it (with a floor such that the assumed foundation height cannot be negative). The right pane shows the corresponding average change in the level of damage with a 1% annual exceedance probability. Each line represents a different type of building; some asset classes in the model are excluded for clarity of the figure; however, these results are also representative of other types (e.g., educational, agriculture, religious). Negative perturbations yield higher risk, whereas assuming higher foundation heights reduces risk. We can observe that even changes in a few centimeters can produce substantial changes in the estimated risk of damage; importantly, this bias is not symmetric.

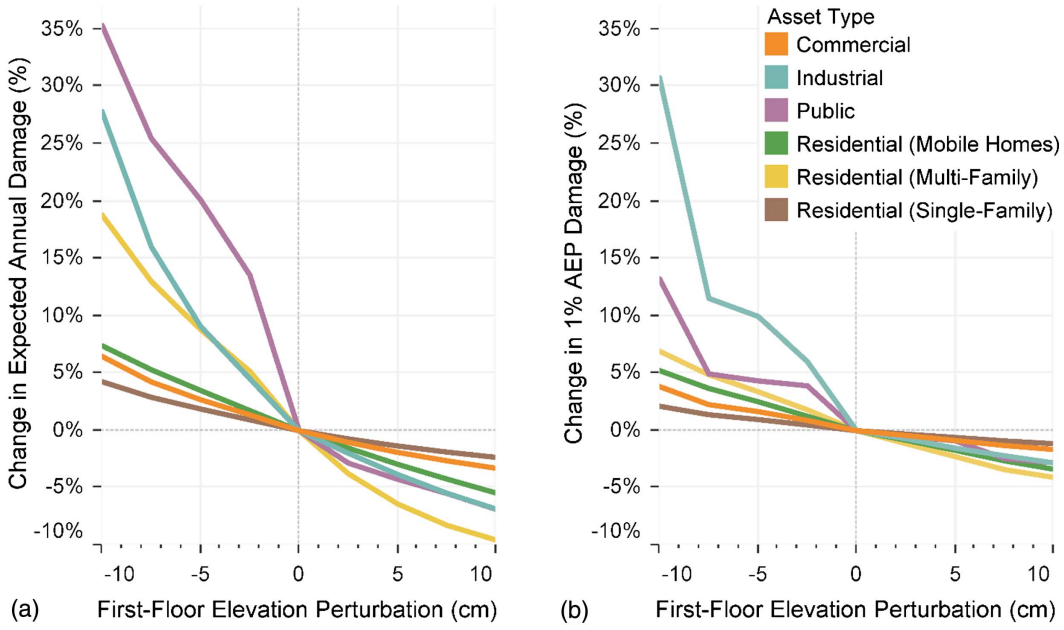


Fig. 10. (a) Average change, by asset type, in expected annual damage; and (b) 1% annual exceedance probability (AEP) damage resulting from deviations in structures' first-floor elevations away from estimates produced by AI algorithm.

Influence of imbalanced data: As shown in Fig. 4, the distributions of building attributes in the dataset are imbalanced. In this case, predicting the dominant attributes may be more accurate, whereas predictions of the tail attributes with less training data may be less accurate. To evaluate the influence of imbalanced data, Table 4 lists the heightwise MAEs and classwise f-scores of the original and weighted models. For foundation height prediction, only the MAEs of heights that had more than 2% of testing data (i.e., 172 testing GSV images) are shown. The original model applied a uniform loss weight for all of the classes and heights during training. For the weighted model, the loss weight for each class or height equaled the inverse of its percentage in the training data, which was intended to increase the influence of tail attributes during training. For instance, the loss weight for "pier foundation" equaled 100%/26% because 26% of the training data images had a "pier foundation" (Fig. 4). To prevent larger loss weights that might cause model divergence during training, all of the loss weights were clipped at a value of 10. Additionally, the initial learning rate was set to 0.00005 for the weighted model because its total loss was larger due to the weightings. Table 4 shows that all of the dominant attributes [0.153 m (0.5 foot) foundation height, slab foundation, residential building, and one-story building] have more accurate predictions when the unweighted model is used. Although the weighted model improved some of the tail attribute predictions, it also made the prediction f-scores for some dominant attributes worse because the dominant attributes had smaller loss weights than the tail attributes during training. As a result, the overall f-scores and MAE of the weighted model are worse than the predictions from the original model with uniform loss weights. Therefore, this study chose a uniform loss for each task as opposed to a height/class-weighted one.

Table 4. Heightwise MAEs and classwise F1 scores of original model using uniform losses and weighted model using weighted losses during training. Foundation heights are binned with header indicating upper end of bin range (i.e., 0.153 represents 0–0.153 m)

	Foundation height MAE (m)						Foundation type F1 score (%)				Building type F1 score (%)				Building stories F1 score (%)		
Weighted	0.153[a]	0.305	0.458	0.610	0.915	1.220	Overall	Pier	Slab[a]	Mobile	Others	Overall	Comm.	Resi[a]	Mobile	Overall	1 story[a]	2+ stories	Overall
—	0.070	0.116	0.134	0.244	0.278	0.442	0.177	70.60	88.46	74.69	77.93	77.96	77.39	96.56	74.87	83.12	99.06	90.19	94.60
Yes	0.122	0.146	0.131	0.217	0.268	0.421	0.198	68.43	87.83	73.85	78.27	77.31	77.56	96.40	74.14	82.90	99.00	89.79	94.40

[a]Dominant attributes with highest occurrence in training data.

Impact of Accurate Attribute Values on Flood Risk Estimates

The Motivation section outlined how structure attribute data are used to estimate flood risk. Here, we empirically examine the potential impact of accurate accounting by running the structure-level foundation heights and square footage of the approximately 36,900 single-family residences from the ground-truth training datasets through the Coastal Louisiana Risk Assessment (CLARA) model. We calculate the total damage to these assets from flood depths with annual exceedance probabilities (AEP) ranging from 20% (i.e., five-year flood depths with a one-in-five chance of occurring or being exceeded in a given year) to 0.005% (i.e., 2,000-year flood depths). The flood depths are those estimated by CLARA for the coastal Louisiana landscape in 2015, which is used to represent current conditions for the state's 2017 Coastal Master Plan.

Risk estimates using the ground-truth values for foundation heights and square footage are taken as the basis for comparison. Because these attributes are typically spatially aggregated, we also run the model using the mean values of the structures in each census block. Fig. 11 shows that spatially aggregating structure attributes underestimate the risk from more frequent flood events and overestimate the risk from more rare and extreme events. The pattern is intuitive because flood depths with high AEP are lower, meaning that modeling the mean foundation heights results in an assumption that a larger percentage—if not all—of the structures have first-floor elevations above the flood depths, resulting in little damage. In other words, structures with below-average foundation heights are assigned values higher than they actually are; therefore, the estimated risk to those structures from frequent events is lower than in reality.

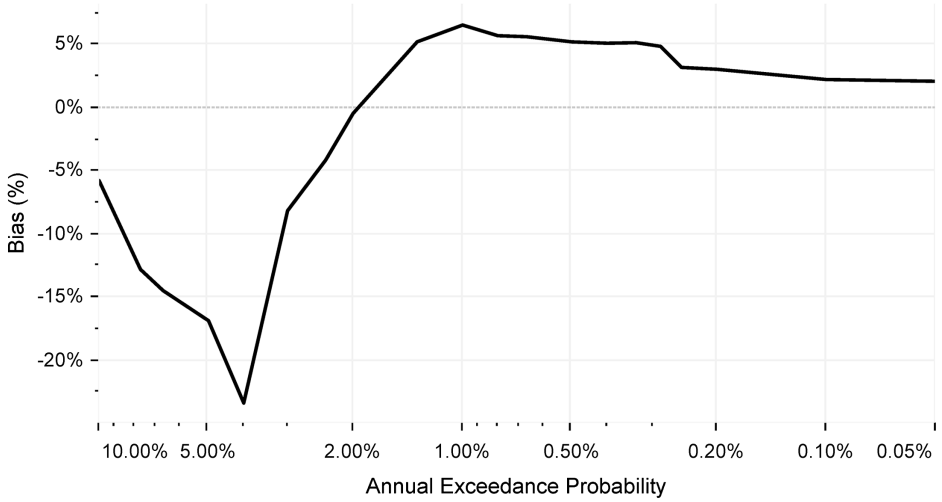


Fig. 11. Bias in estimated damage exceedances by annual exceedance probability introduced by spatially averaging foundation heights and square footage of single-family residences in training datasets at census block level.

In contrast, damage from more extreme events with lower AEP is overestimated by using mean values for structure attributes. The same intuition applies: when all structures are assigned mean values for the foundation height and are exposed to high flood depths, structures with above-average heights are assumed to incur damage when they would not in reality. The point at which, empirically, damage estimates transition from underestimating to overestimating risk is important for two policy-relevant reasons. On the one hand, the expected annual damage (EAD) from flooding produced by integrating the damage exceedances over their probability distribution is underestimated at $115.5 million compared with $117.75 million when using the ground-truth attributes (−1.9%). Because EAD is an expected value, it is driven by more frequent events that are weighted more heavily when taking a weighted average, and the risk from these events with higher AEP is biased downwards by using mean foundation heights.

On the other hand, damage from the 100-year flood event (i.e., 1% AEP) is overestimated by 6.5%. The protection from 100-year events is a common standard for flood protection in the United States, relevant to requirements for flood insurance and other policy decisions. Overestimating this quantity is at odds with underestimating EAD, which is relevant to planners engaging in benefit-cost analyses of flood protection measures or calculating actuarially fair flood insurance rates. Given that the 36,900 single-family residences used in this analysis represent only 8% of the approximately 459,000 single-family homes in coastal Louisiana vulnerable to storm surge-based flooding, our analysis indicates that using mean values for structural attributes could underestimate the expected damage by approximately $28 million per year or hundreds of millions of dollars during the coming decades.

Error Analysis

Some of the correctly predicted samples are shown in Fig. 12, and incorrectly predicted samples are shown in Fig. 13. In this section, we perform error analysis using the incorrectly predicted samples. For foundation height estimation, the system predicts significantly higher heights than the ground truth for buildings that have slab foundations but also pillars on their facade. Similarly, some mistakes also occur as a result of wrong labels. For instance, the system predicts 0.180 m as one of the buildings' foundation height, which visually appears closer to the true value but has a large numerical disparity with the ground truth (2.440 m). When looking at the wrong predictions in foundation type classification, one of the key points to note is the number of labeling mistakes across sections. In some cases, the model actually predicts the correct class, but the labels are wrong. At other times, the errors are random, such as the prediction of mobile foundations as pier, mobile foundations as others, and others as mobile, which is not surprising. Foundation type estimation is a four class task, and even our best models fail to properly characterize the minor classes.

The classifications of mobile homes as residential and residential homes as mobile seem to occur because of unclear boundaries between some of the buildings in these two classes. Some mobile homes' visual features resemble those of residential homes and vice-versa, and the model fails to capture these distinctions. The misclassification of some of the commercial buildings as residential seems to occur for the same reason; however, the misclassification of residential buildings as commercial seems random, suggesting that the model fails to characterize some of the patterns in residential homes that distinguish them from commercial homes.

The errors in the one-story category occur when tall slab foundations underlie the structures. Some mistakes may also have occurred due to noise. For instance, in one of the images, the building adjacent to the building of concern seems to be multistoried. Similarly, the errors in the "two stories" category do not seem to possess an alternate explanation other than the fact that they happen because of the model's inability to properly categorize some of the multistoried buildings.

To conclude, two main kinds of errors exist in our predictions. The first one originates because of the inability of the models to properly characterize a specific class. The other is due to incorrect labeling. To prevent the former type of error from happening, more advanced versions of TREncNet and feature fusion that introduce even more relevant metadata and encode complex relationships need to be conceptualized. Similarly, to prevent the latter type of error from happening, local agencies need to be more careful when taking and keeping records of building attributes.

Key Findings and Potential Implications

The first key finding of this study is that building attribute estimation is best modeled as a multiobjective problem if more than one attribute exists. By reducing the number of models, the inference

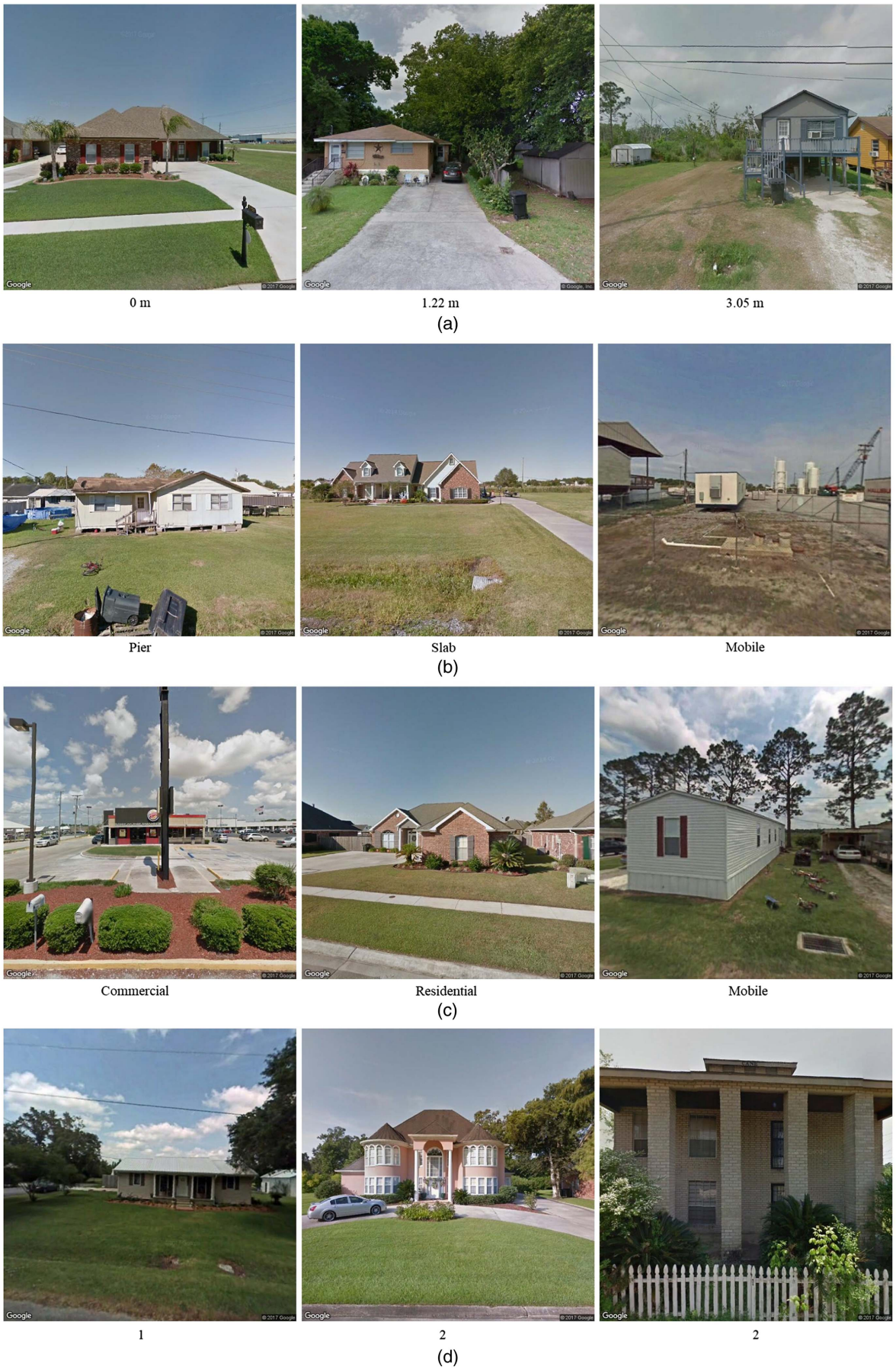


Fig. 12. Sample GSV images with correct predictions: (a) foundation height; (b) foundation type; (c) building type; and (d) number of storeys. (Images © 2017 Google; image © Google, Inc.)

Fig. 13. Sample GSV images with incorrect predictions: (a) foundation height; (b) foundation type; (c) building type; and (d) number of storeys. (Images © 2017 Google; images © Google, Inc.)

time can be decreased significantly, which allows authorities to more frequently carry out projects, such as flood risk assessment. Furthermore, the performance of the single task learning approach in estimating the most important building attribute regarding flood risk management is concerned (i.e., foundation height) is subpar. The lower foundation height MAEs in multitask models suggests that some degree of correlation exists between foundation height and the other attributes, and deep learning architectures possess enough ability to exploit them. However, the performance gains in the classification tasks were not as high, suggesting that they are relatively easy tasks and, thus, are affected by the regularization that can sometimes be induced by MTL. This disparity in the performance gain between easy and difficult tasks is consistent with the findings of MTL studies from other fields.

The second finding of this study is that additional sources of information through either feature fusion or the encoding of known relationships can improve performance. This is especially the case when the two are used together rather than separately. Similarly, the type of metadata and relationship encoding may impact the performance of individual categories. Evidence of this idea is observed in our results.

For instance, TREncNet did not provide any improvement in foundation height estimation, but feature fusion did. We hypothesize that, although foundation height is correlated with other tasks (evident by the performance gain when MTL instead of STL was used), the encoded relationships apparent to us might have been equally apparent to the architectures, which made our encodings redundant. However, for feature fusion, some information such as camera-to-building distances were extraneous to the images, whereas others such as aspect ratios of bounding boxes might have been too difficult for the networks to estimate without outside help. Furthermore, the largest-to-smallest reduction in foundation height MAE was 15% for MobileNet V1, 12% for Inception V2, 9% for Inception V4, and 8% for Inception ResNet V2, which lends some credence to this theory because simpler architectures might be less prone to map harder-to-extract information by themselves.

Moreover, the results of foundation and building types show that TREncNet has a stronger impact than feature fusion. When compared with those two attributes, its impact on the number of stories and foundation height categories is negligible. To determine why, we must observe two factors: the architecture of TREncNet and the confusion matrices. Four total architectures exist; however, MobileNet V1 not only echoes the pattern in most other architectures but also accentuates it and serves as a case study for analyzing where TREncNet may or may not work or where feature fusion may or may not work. Therefore, we have illustrated its confusion matrices for the three classification tasks along with the boxplots for the foundation height task in Fig. 14, where the number of samples is 4,341, 1,252, 503, 737, 995, and 245, respectively, for each of the foundation height intervals. First, we must refer to Fig. 7.

As is observed, task encoding for the number of stories task is minimal. We only share the logits of the mobile home from building type task with one-story of the building stories task. However, from Fig. 14, the performance of the one-story class was already saturated even without TREncNet. For both foundation type and building type tasks, TREncNet improves the performance of minority classes. To understand why, let us refer back to Fig. 7. The building type category takes logits from the number of stories category. We performed a Cramer's V test using the chi-square test of independence between the two categories. Cramer's V value ranges from 0% to 100% in percentage form, and a higher value suggests a stronger relationship between nominal variables. With the residential buildings included, the result of the Cramer's V test between the number of stories and the building type was 8%. When it was excluded, Cramer's V output jumped to 20%. Although this is still a weak correlation in statistical terms, in the context of deep learning, even this small value might have been enough to significantly propel the performance of minority classes in building type estimations. The same can be said about the minority classes of foundation type and number of stories. With slab foundation excluded, the Cramer's V value jumped from 13% to 15%. Here, however, the slab foundation prediction has increased in the confusion matrix as well, which means that the initial high value and the lower increase with the slab foundation excluded are still consistent with our theory. We could not perform the same test for foundation and building types because mobile homes were a part of both tasks.

This study is the first that proposes the inclusion of encoding and metadata for building attribute estimation in flood risk assessment, and we firmly believe that more room exists for experimentation. This version of TREncNet and these metadata are only meant as proofs of concepts rather than ends in themselves. Ultimately, we hope that the main implication of this paper lies in the message that both performance and cost-efficiency can be achieved at the same time when predicting building attributes.

Limitations and Future Work

One of the limitations of this work is that the coverage of the GSV database may present a problem in certain cases. In much of the developing world (including most of Africa), GSV is almost nonexistent. Even in parts of major world economies, such as China and India, GSV images mostly cover only cities and landmarks. Be that as it may, our approach is broadly applicable in developed countries due to GSV's spatial coverage, having photographed more than 10 million miles of roads (Nieva 2019). Moreover, GSV's coverage in developing countries is increasing every day, and Google has made significant efforts toward providing the needed resources (for instance, cars and all of the necessary sensors) to motivated native people such that GSV photographs can be obtained through semicrowdsourced ventures. Such efforts have been met with marked enthusiasm, primarily on account of the young and increasingly tech-savvy populace of low- and middle-income countries. Updates are also relatively frequent, with images being refreshed approximately every one to three years in coastal Louisiana (more frequent updates are in urban areas such as New Orleans). GSV also allows fully crowdsourced contributions in some cases, enabling researchers doing postevent reconnaissance after Hurricane Laura in 2020 to publish new imagery within days to document damage to the western part of the state. Similarly, although GSV images (where available) are good sources of information on a structure's facade, they are typically information-poor for its sides and posterior (typically, only a few buildings in the GSV database have images of multiple sides). Some of the distinguishing features that are relevant to building and foundation type classifications may be apparent from these alternate vantage points as well; thus, the availability of such views may enhance performance.

One future study could be a data fusion scheme that aggregates the detection and prediction results from multiple GSV images of the same building to increase the robustness of the estimation. Moreover, we used only the camera's metadata in our feature fusion scheme; however, other types of information, such as the building's zoning and ownership, could also prove advantageous. In addition to processing only image data, combining data from audio or depth sensors may also be beneficial (Huang and Kingsbury 2013; Hazirbas et al. 2016; Wu et al. 2013; Kamel et al. 2018). To correctly train and evaluate the detection and prediction models, we manually removed the GSV images with unacceptable views

Fig. 14. Boxplots for (a) foundation height estimation, and confusion matrices for (b) foundation type; (c) building type; and (d) number of stories classification for three separate multitask learning settings of MobileNet V1.

(e.g., buildings blocked by trees). Although the quality of the views can be determined autonomously as well—for instance, a low detection score or a small bounding box during the inference phase may be used to filter them out—this is still not efficient. Future directions building on this work could be to implement an object recognition model that can estimate the view quality of details.

Last but not least, the performance of our approach in some of the key tasks and their classes needs further improvement.

Regarding foundation height, we were able to achieve an MAE as low as 0.171 m. Although a significant improvement over a single-task learning approach's MAE (0.223 m) and a plain MTL approach's MAE (0.192 m), it is still quite high considering that the average foundation height of our data was 0.45 m. The overall MAE for foundation height is small compared with uncertainty in flood depth exceedance values arising from factors such as uncertainty in hydrodynamic models, variability in levee overtopping volumes, probability of system failures, noise in lidar measurements, and randomness in the characteristics of observed historic events (for example, studies of flood hazards in Louisiana have found that 80% confidence intervals for the 100-year flood depth may be several feet in many locations (Fischbach et al. 2017) but is not as low as it could be from a computational perspective. Of particular concern is the fact that our worst relative estimates (MAE/ground truth) were usually in the lowest foundation (0.153 m; Table 4) when lower foundations are actually at greater flood risks than higher ones.

Similarly, in some buildings with a foundation height of 0.153 m, the prediction is as high as 3.0 m (Fig. 14). Still, this study is the first of its kind for inexpensive and efficient regression of foundation heights from GSV images using deep learning techniques and in a MTL environment; therefore, we are positive that improvements will be made in the future by building on it. The performances in the number of stories task were promising for all the mentioned approaches. However, they were still heavily skewed toward the majority class. From a financial planning perspective, this is inefficient because the classification of a significant proportion of two-storied buildings as one-storied can drastically overestimate flood damage. Ideally, we want a balanced performance if improving the performances of both classes is not possible, such that we are neither more conservative nor more generous than optimal. The same can be said about foundation type and building type classification. By erroneously predicting a quarter of commercial buildings as residential, we may be underestimating flood damage (commercial buildings may have more to lose if flooded). Similarly, the detection of pier foundations might be more important than the detection of slab foundations in some areas because the former is typically located in riskier places (such as immediately next to coasts within a larger coastal region). The fact that our approach's F1 score in pier foundation detection is only slightly higher than the baseline is a limitation. Overall, we improved the performances of some classes to an extent, but they still need significant improvements if we are to accurately assess flood risks (i.e., neither overestimate nor underestimate damages). The best way forward from this research is as follows: for classification tasks, to find ways to keep the performances in the majority classes at least at the current level and raise the performances of the minority classes to the level observed in the majority classes; and for foundation height, significantly improve the performance for smaller heights. In summary, a selective approach might be a more suitable next step to improve the overall performance.

Conclusion

To collect comprehensive and up-to-date structure attribute data to assess flood risks without labor-intensive street surveys, this study details a deep learning-based framework that can simultaneously estimate multiple building attributes from GSV imagery. An extensive evaluation is done to select the optimal building detection model. Furthermore, a feature fusion scheme is proposed that combines image features with meta information that improves the prediction of foundation heights. Additionally, TREncNet is introduced to encode task relations as network connections for MTL that enhance the predictions of classification tasks.

The proposed framework achieves a 0.177-m foundation height prediction MAE and classification f-scores of 77.96% for foundation type, 83.12% for building type, and 94.60% for building stories and requires less than five days to predict the attributes of 0.8 million buildings in the coastal Louisiana study region.

Given its capabilities and efficiency, the proposed framework saves time and money for flood risk studies. The method also has the potential for predicting other types of building attributes relevant to estimating hurricane, tornado, or seismic hazards. Consequently, the data set of inferred foundation heights (Chen et al. 2020) is planned for direct use by the risk model used in Louisiana's 2023 Coastal Master Plan (Brown et al. 2020). However, the classification of nonresidential assets (e.g., distinguishing movie theaters, banks, and hospitals from each other) would be challenging without much larger ground-truth training sets, and GSV images often do not capture the full structure for large, multistory assets such as office towers or hotels.

Deep learning-based methods have been used in the geospatial context in the past; however, such methods have primarily focused on performance at the expense of cost. This study proves that both inexpensive and accurate building attribute prediction schemes are possible by combining three methods: MTL, feature fusion, and TREncNet.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author on reasonable request. The coastwide data set of estimated structural features is publicly available (Chen et al. 2020). The ground-truth data collected by government sources are available on request to Dr. Johnson. The availability of the images used in this work is restricted by the terms of use of the GSV API.

Acknowledgments

This study was funded in part by the Andrew W. Mellon Foundation. We also extend our sincere thanks to Mr. Ahmad Bassel Abdallah, a freshman at Purdue University, for helping us with some of the figures used in this paper.

References

Alipour, M., and D. K. Harris. 2020. "A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training." *J. Civ. Struct. Health Monit.* 10 (2): 313–332. https://doi.org/10.1007/s13349-020-00386-4.

Arrighi, C., B. Mazzanti, F. Pistone, and F. Castelli. 2020. "Empirical flash flood vulnerability functions for residential buildings." *SN Appl. Sci.* 2 (5): 1–12. https://doi.org/10.1007/s42452-020-2696-1.

Bao, Y., Z. Tang, H. Li, and Y. Zhang. 2019. "Computer vision and deep learning–based data anomaly detection method for structural health monitoring." *Struct. Health Monit.* 18 (2): 401–421. https://doi.org/10.1177/1475921718757405.

Brown, S., E. White, Z. Cobell, and D. R. Johnson. 2020. *2023 coastal master plan—Technical modeling workshop*. Baton Rouge, LA: Louisiana Coastal Protection and Restoration Authority.

Cai, J., L. Yang, Y. Zhang, S. Li, and H. Cai. 2021. "Multitask learning method for detecting the visual focus of attention of construction workers." *J. Constr. Eng. Manage.* 147 (7): 04021063. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002071.

Caruana, R. 1997. "Multitask learning." *Mach. Learn.* 28 (1): 41–75. https://doi.org/10.1023/A:1007379606734.

Chen, F.-C., M. R. Jahanshahi, D. R. Johnson, and E. J. Delp. 2020. *Structural attributes derived from Google Street View imagery.* West Lafayette, IN: Purdue Univ. Research Repository.

Cheng, C.-S., A. H. Behzadan, and A. Noshadravan. 2021. "Deep learning for post-hurricane aerial damage assessment of buildings." *Comput. Aided Civ. Infrastruct. Eng.* 36 (6): 695–710. https://doi.org/10.1111/mice.12658.

Dai, J., Y. Li, K. He, and J. Sun. 2016. "R-FCN: Object detection via region-based fully convolutional networks." In Vol. 29 of *Proc., 30th Int. Conf. on Neural Information Processing Systems*, 379–387. New York: Association for Computing Machinery.

Dall'Osso, F., M. Gonella, G. Gabbianelli, G. Withycombe, and D. Dominey-Howes. 2009. "A revised (PTVA) model for assessing the vulnerability of buildings to tsunami damage." *Nat. Hazards Earth Syst. Sci.* 9 (5): 1557–1565. https://doi.org/10.5194/nhess-9-1557-2009.

D'Ayala, D., K. Wang, Y. Yan, H. Smith, A. Massam, V. Filipova, and J. J. Pereira. 2020. "Flood vulnerability and risk assessment of urban traditional buildings in a heritage district of Kuala Lumpur, Malaysia." *Nat. Hazards Earth Syst. Sci.* 20 (8): 2221–2241. https://doi.org/10.5194/nhess-20-2221-2020.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. "ImageNet: A large-scale hierarchical image database." In *Proc., 2009 IEEE Conf. Comput. Vision Pattern Recognition (CVPR'09)*, 248–255. New York: IEEE.

Emanuel, K. 2005. "Increasing destructiveness of tropical cyclones over the past 30 years." *Nature* 436 (7051): 686. https://doi.org/10.1038/nature03906.

Fischbach, J. R., D. R. Johnson, K. Kuhn, M. Pollard, C. Stelzner, R. Costello, E. Molina-Perez, R. Sanchez, H. J. Roberts, and Z. Cobell. 2017. *2017 coastal master plan attachment C3-25: Storm surge and risk assessment*. Baton Rouge, LA: Louisiana Coastal Protection and Restoration Authority.

Gao, Y., and K. M. Mosalam. 2018. "Deep transfer learning for image-based structural damage recognition." *Comput.-Aided Civ. Infrastruct. Eng.* 33 (9): 748–768. https://doi.org/10.1111/mice.12363.

Gonzalez, D., D. Rueda-Plata, A. B. Acevedo, J. C. Duque, R. Ramos-Pollan, A. Betancourt, and S. Garcia. 2020. "Automatic detection of building typology using deep learning methods on street level images." *Build. Environ.* 177 (3): 106805. https://doi.org/10.1016/j.buildenv.2020.106805.

Gopalakrishnan, K., S. K. Khaitan, A. Choudhary, and A. Agrawal. 2017. "Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection." *Constr. Build. Mater.* 157 (Sep): 322–330. https://doi.org/10.1016/j.conbuildmat.2017.09.110.

Hailegeorgis, T. T., and K. Alfredsen. 2017. "Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for Mid-Norway." *J. Hydrol.: Reg. Stud.* 9 (Feb): 104–126. https://doi.org/10.1016/j.ejrh.2016.11.004.

Hall, W. A., and D. T. Howell. 1963. "Estimating flood probabilities within specific time intervals." *J. Hydrol.* 1 (3): 265–271. https://doi.org/10.1016/0022-1694(63)90006-4.

Hallegatte, S., C. Green, R. J. Nicholls, and J. Corfee-Morlot. 2013. "Future flood losses in major coastal cities." *Nat. Clim. Change* 3 (9): 802. https://doi.org/10.1038/nclimate1979.

Haralick, R. M., K. Shanmugam, and I. H. Dinstein. 1973. "Textural features for image classification." *IEEE Trans. Syst. Man Cybern.* 6: 610–621. https://doi.org/10.1109/TSMC.1973.4309314.

Hazirbas, C., L. Ma, C. Domokos, and D. Cremers. 2016. "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture." In *Proc., Asian Conf. on Computer Vision*, 213–228. New York: Springer.

He, K., X. Zhang, S. Ren, and J. Sun. 2016a. "Deep residual learning for image recognition." In *Proc., IEEE Conf. on computer vision and pattern recognition*, 770–778. New York: IEEE.

He, K., X. Zhang, S. Ren, and J. Sun. 2016b. "Identity mappings in deep residual networks." In *Proc., European Conf. on Computer Vision*, 630–645. New York: Springer.

Hoffmann, E. J., Y. Wang, M. Werner, J. Kang, and X. X. Zhu. 2019. "Model fusion for building type classification from aerial and street view images." *Remote Sens.* 11 (11): 1259. https://doi.org/10.3390/rs11111259.

Hoskere, V., Y. Narazaki, T. A. Hoang, and B. Spencer. 2020. "Madnet: Multi-task semantic segmentation of multiple types of structural materials and damage in images of civil infrastructure." *J. Civ. Struct. Health Monit.* 10 (12): 757–773. https://doi.org/10.1007/s13349-020-00409-0.

Howard, A. G., M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. 2017. *MobileNets: Efficient convolutional neural networks for mobile vision applications*. Ithaca, NY: Cornell Univ.

Huang, J., et al. 2017a. "Speed/accuracy trade-offs for modern convolutional object detectors." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 7310–7311. New York: IEEE.

Huang, J., and B. Kingsbury. 2013. "Audio-visual deep learning for noise robust speech recognition." In *Proc., IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 7596–7599. New York: IEEE.

Huang, Y., L. Zhuo, H. Tao, Q. Shi, and K. Liu. 2017b. "A novel building type classification scheme based on integrated lidar and high-resolution images." *Remote Sens.* 9 (7): 679. https://doi.org/10.3390/rs9070679.

Huber, P. J. 1992. "Robust estimation of a location parameter." In *Breakthroughs in statistics*, 492–518. Berlin: Springer.

Iannelli, G. C., and F. Dell'Acqua. 2017. "Extensive exposure mapping in urban areas through deep analysis of street-level pictures for floor count determination." *Urban Sci.* 1 (2): 16. https://doi.org/10.3390/urbansci1020016.

Ioffe, S., and C. Szegedy. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." In *Proc., 32nd Int. Conf. Mach. Learning (ICML'15)*, 448–456. San Diego: JMLR.

Johnson, D. R., J. R. Fischbach, and D. S. Ortiz. 2013. "Estimating surge-based flood risk with the coastal Louisiana risk assessment model." *J. Coastal Res.* 67 (1): 109–126. https://doi.org/10.2112/SI_67_8.

Kamel, A., B. Sheng, P. Yang, P. Li, R. Shen, and D. D. Feng. 2018. "Deep convolutional neural networks for human action recognition using depth maps and postures." *IEEE Trans. Syst. Man Cyber. Syst.* 49 (9): 1806–1819. https://doi.org/10.1109/TSMC.2018.2850149.

Kang, J., M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu. 2018. "Building instance classification using street view images." *ISPRS J. Photogramm. Remote Sens.* 145 (12): 44–59. https://doi.org/10.1016/j.isprsjprs.2018.02.006.

Kellens, W., T. Terpstra, and P. De Maeyer. 2013. "Perception and communication of flood risks: A systematic review of empirical research." *Risk Anal.: Int. J.* 33 (1): 24–49. https://doi.org/10.1111/j.1539-6924.2012.01844.x.

Kendall, A., Y. Gal, and R. Cipolla. 2018. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 7482–7491. New York: IEEE.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet classification with deep convolutional neural networks." In *Proc., Advances Neural Information Process System 25 (NIPS'12)*, 1097–1105. Cambridge, MA: MIT Press.

Landsea, C. W., B. A. Harper, K. Hoarau, and J. A. Knaff. 2006. "Can we detect trends in extreme tropical cyclones?" *Science* 313 (5786): 452–454. https://doi.org/10.1126/science.1128448.

Laudan, J., V. Rözer, T. Sieg, K. Vogel, and A. H. Thieken. 2017. "Damage assessment in Braunsbach 2016: Data collection and analysis for an improved understanding of damaging processes during flash floods." *Nat. Hazards Earth Syst. Sci.* 17 (12): 2163–2179. https://doi.org/10.5194/nhess-17-2163-2017.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. "Deep learning." *Nature* 521 (7553): 436–444. https://doi.org/10.1038/nature14539.

Lehmann, J., D. Coumou, and K. Frieler. 2015. "Increased record-breaking precipitation events under global warming." *Clim. Change* 132 (4): 501–515. https://doi.org/10.1007/s10584-015-1434-y.

Lenjani, A., S. J. Dyke, I. Bilionis, C. M. Yeum, K. Kamiya, J. Choi, X. Liu, and A. G. Chowdhury. 2020. "Towards fully automated post-event data collection and analysis: Pre-event and post-event information fusion." *Eng. Struct.* 208 (12): 109884. https://doi.org/10.1016/j.engstruct.2019.109884.

Li, W., B. Xu, and J. Wen. 2016. "Scenario-based community flood risk assessment: A case study of Taining county town, Fujian Province, China." *Nat. Hazards* 82 (1): 193–208. https://doi.org/10.1007/s11069-016-2187-2.

Li, X., C. Zhang, and W. Li. 2017. "Building block level urban land-use information retrieval based on Google Street View images." *GISci. Remote Sens.* 54 (6): 819–835. https://doi.org/10.1080/15481603.2017.1338389.

Li, Y., Y. Chen, A. Rajabifard, K. Khoshelham, and M. Aleksandrov. 2018. "Estimating building age from Google Street View images using deep learning (short paper)." In *Proc., 10th Int. Conf. on Geographic Information Science (GIScience 2018)*. Wadern, Germany: Schloss Dagstuhl.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. "Microsoft COCO: Common objects in context." In *Proc., European Conf. on Computer Vision*, 740–755. New York: Springer.

Liu, C.-J., V. A. Krylov, P. Kane, G. Kavanagh, and R. Dahyot. 2020. "IM2ELEVATION: Building height estimation from single-view aerial imagery." *Remote Sens.* 12 (17): 2719. https://doi.org/10.3390/rs12172719.

Liu, W., D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. 2016. "SSD: Single shot multibox detector." In *Proc., European Conf. on Computer Vision*, 21–37. New York: Springer.

Louisiana Coastal Protection and Restoration Authority. 2012a. *Louisiana's comprehensive master plan for a sustainable coast.* Baton Rouge, LA: State of Louisiana.

Louisiana Coastal Protection and Restoration Authority. 2012b. *Louisiana's comprehensive master plan for a sustainable coast.* Baton Rouge, LA: State of Louisiana.

Ma, J., Z. Zhao, X. Yi, J. Chen, L. Hong, and E. H. Chi. 2018. "Modeling task relationships in multi-task learning with multi-gate mixture-of-experts." In *Proc., 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*. New York: Association for Computing Machinery.

Maniat, M., C. V. Camp, and A. R. Kashani. 2021. "Deep learning-based visual crack detection using Google Street View images." In *Neural computing and applications*, 1–18. New York: Springer.

McGrath, H., E. Stefanakis, M. McCarthy, and M. Nastev. 2014. *Data preparation for validation study of Hazus Canada flood model*. West Lafayette, IN: Purdue Univ.

Meehl, G. A., J. M. Arblaster, and C. Tebaldi. 2005. "Understanding future patterns of increased precipitation intensity in climate model simulations." *Geophys. Res. Lett.* 32 (18): 120–125. https://doi.org/10.1029/2005GL023680.

Meyer, G. P. 2021. "An alternative probabilistic interpretation of the Huber loss." In *Proc., IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 5261–5269. New York: IEEE.

Meyerson, E., and R. Miikkulainen. 2017. *Beyond shared hierarchies: Deep multitask learning through soft layer ordering*. Ithaca, NY: Cornell Univ.

Michael, J. A. 2007. "Episodic flooding and the cost of sea-level rise." *Ecol. Econ.* 63 (1): 149–159. https://doi.org/10.1016/j.ecolecon.2006.10.009.

Mou, L., and X. X. Zhu. 2018. *IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network*. Ithaca, NY: Cornell Univ.

Neumann, J. E., K. Emanuel, S. Ravela, L. Ludwig, P. Kirshen, K. Bosma, and J. Martinich. 2015. "Joint effects of storm surge and sea-level rise on US coasts: New economic estimates of impacts, adaptation, and benefits of mitigation policy." *Clim. Change* 129 (1–2): 337–349. https://doi.org/10.1007/s10584-014-1304-z.

Nieva, R. 2019. *Google Maps has now photographed 10 million miles in street view.* San Francisco: CNET.

Pi, Y., N. D. Nath, and A. H. Behzadan. 2020. "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery." *Adv. Eng. Inf.* 43 (Sep): 101009. https://doi.org/10.1016/j.aei.2019.101009.

Pinelli, J., D. Rodriguez, D. Roueche, K. Gurley, M. Baradaranshoraka, S. Cocke, S. Dong-Wook, L. Lapaiche, and R. Gay. 2018. "Data management for the development of a flood vulnerability model." In *Proc., European Safety and Reliability Conf., Trondheim, Norway*, 17–21. Boca Raton, FL: CRC Press.

Qi, F., J. Z. Zhai, and G. Dang. 2016. "Building height estimation using Google Earth." *Energy Build.* 118 (Mar): 123–132. https://doi.org/10.1016/j.enbuild.2016.02.044.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. "You only look once: Unified, real-time object detection." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 779–788. New York: IEEE.

Ren, S., K. He, R. Girshick, and J. Sun. 2015. "Faster R-CNN: Towards real-time object detection with region proposal networks." In *Proc., Advances Neural Information Process System 28 (NIPS'15)*, 91–99. Cambridge, MA: MIT Press.

Ruder, S. 2017. *An overview of multi-task learning in deep neural networks*. Ithaca, NY: Cornell Univ.

Rundle, A. G., M. D. Bader, C. A. Richards, K. M. Neckerman, and J. O. Teitler. 2011. "Using Google Street View to audit neighborhood environments." *Am. J. Preven. Med.* 40 (1): 94–100. https://doi.org/10.1016/j.amepre.2010.09.034.

Russakovsky, O., et al. 2015. "ImageNet large scale visual recognition challenge." *Int. J. Comput. Vision* 115 (3): 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Scawthorn, C., et al. 2006. "HAZUS-MH flood loss estimation methodology. II. Damage and loss assessment." *Nat. Hazard. Rev.* 7 (2): 72–81. https://doi.org/10.1061/(ASCE)1527-6988(2006)7:2(72).

Sen, M. K., S. Dutta, and J. I. Laskar. 2021. "A hierarchical Bayesian network model for flood resilience quantification of housing infrastructure systems." *J. Risk Uncertainty Eng. Syst. Part A: Civ. Eng.* 7 (1): 04020060. https://doi.org/10.1061/AJRUA6.0001108.

Silberman, N., and S. Guadarrama. 2016. *TensorFlow-Slim image classification model library.* San Francisco: Github.

Simonyan, K., and A. Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition." In *Proc., 3rd Int. Conf. on Learning Representations*, edited by Y. Bengio and Y. LeCun. La Jolla, CA: International Conference on Learning Representations.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. "Dropout: A simple way to prevent neural networks from overfitting." *J. Mach. Learn. Res.* 15 (1): 1929–1958.

Stamataki, I., and T. R. Kjeldsen. 2021. "Reconstructing the peak flow of historical flood events using a hydraulic model: The city of Bath, United Kingdom." *J. Flood Risk Manage.* 14 (3): e12719. https://doi.org/10.1111/jfr3.12719.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi. 2017. "Inception-v4, Inception-ResNet and the impact of residual connections on learning." In *Proc., 31st AAAI Conf. on Artificial Intelligence*, 4278–4284. Palo Alto, CA: Association for the Advancement of Artificial Intelligence.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. "Going deeper with convolutions." In *Proc., IEEE Conf. Computer Vision Pattern Recognition (CVPR'15)*, 1–9. New York: IEEE.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. "Rethinking the inception architecture for computer vision." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 2818–2826. New York: IEEE.

Thrun, S. 1996. "Is learning the n-th thing any easier than learning the first?" In *Advances in neural information processing systems*, 640–646. Cambridge, MA: MIT Press.

USACE. 2009. *Louisiana coastal protection and restoration final technical report.* Washington, DC: USACE.

Valenzuela, J. T., R. S. Carredo, C. Z. Coca, C. L. Patiño, and J. R. Sinogaya. 2016. "Web-and mobile-based data collection using VGI for building feature mapping/attribution in the flood-prone zones of western Visayas, Philippines." In *Proc., GSDI 15 World Conf.*, 129. Gilbertville, IA: GSDI Association Press.

Viola, P., et al. 2001. "Rapid object detection using a boosted cascade of simple features." In Vol. 511–518 of *Proc., 2001 IEEE Computer*

Society Conf. on Computer Vision and Pattern Recognition. New York: IEEE. https://doi.org/10.1109/CVPR.2001.990517.

Wan, H.-P., and Y.-Q. Ni. 2019. "Bayesian multi-task learning methodology for reconstruction of structural health monitoring data." *Struct. Health Monit.* 18 (4): 1282–1309. https://doi.org/10.1177/1475921718794953.

Wright, D. B. 2015. *Methods in flood hazard and risk assessment*. Washington, DC: The World Bank.

Wu, Q., Z. Wang, F. Deng, Z. Chi, and D. D. Feng. 2013. "Realistic human action recognition with multimodal feature selection and fusion." *IEEE Trans. Syst. Man Cyber. Syst.* 43 (4): 875–885. https://doi.org/10.1109/TSMCA.2012.2226575.

Xie, J., and J. Zhou. 2017. "Classification of urban building type from high spatial resolution remote sensing imagery using extended MRS and soft BP network." *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 10 (8): 3515–3528. https://doi.org/10.1109/JSTARS.2017.2686422.

Yang, S., W. Wang, C. Liu, and W. Deng. 2018. "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles." *IEEE Trans. Syst. Man Cyber. Syst.* 49 (1): 53–63. https://doi.org/10.1109/TSMC.2018.2868372.

Yu, Q., C. Wang, F. McKenna, X. Y. Stella, E. Taciroglu, B. Cetiner, and K. H. Law. 2020. "Rapid visual screening of soft-story buildings from street view images using deep learning classification." *Earthquake Eng. Eng. Vib.* 19 (4): 827–838. https://doi.org/10.1007/s11803-020-0598-2.

Zhang, K., L. Zheng, Z. Liu, and N. Jia. 2020. "A deep learning based multi-task model for network-wide traffic speed prediction." *Neurocomputing* 396 (Apr): 438–450. https://doi.org/10.1016/j.neucom.2018.10.097.

Zoph, B., V. Vasudevan, J. Shlens, and Q. V. Le. 2018. "Learning transferable architectures for scalable image recognition." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 8697–8710. New York: IEEE.

Zou, S., and L. Wang. 2021. "Detecting individual abandoned houses from Google Street View: A hierarchical deep learning approach." *J. Photogramm. Remote Sens.* 175 (Mar): 298–310. https://doi.org/10.1016/j.isprsjprs.2021.03.020.