Automation in Construction 122 (2021) 103474

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management

Chaofeng Wang [a,*], Qian Yu [a,b], Kincho H. Law [c], Frank McKenna [a], Stella X. Yu [a,b], Ertugrul Taciroglu [d], Adam Zsarnóczay [c], Wael Elhaddad [a], Barbaros Cetiner [d]

[a] University of California, Berkeley, United States
[b] International Computer Science Institute, United States
[c] Stanford University, United States
[d] University of California, Los Angeles, United States

A B S T R A C T

The intensity of many natural hazards, such as hurricanes, floods, tornadoes, etc., are increasing as a consequence of climate change. This increase in intensity coupled with the increase in population density, particularly along the coasts, is only magnifying the impact of such events. In order to quantify and mitigate the risk due to the hazards and to prepare for the potential impacts in a region, it is necessary to collect the information of existing buildings that are pertinent to natural hazard analysis and risk management. Gathering the building information in a region- or city-scale is a laborious and expensive undertaking. This paper presents a framework for regional scale building information generation/gathering to support regional hazard analysis. In this framework, different types of data are acquired from multiple sources (satellite and street view images, property tax assessment data, etc.) and are fused to semantically profile each building in a city. Specifically, deep learning technique is employed to extract building information from street or satellite images. A novel data mining tool is developed to overcome the data scarcity issue, quantify the uncertainty and enrich the data repository. With this framework, building inventories of cities are created to provide the data needed for disaster and risk management planning and simulations.

1. Introduction

Natural disasters pose significant destructive impacts to society, from damaging buildings, endangering lives to economic loss. For example, earthquake hazards (ground shaking, landslides, liquefaction, surface rupture) can cause damage to buildings and infrastructures because of crumbling the above ground structure and warping the underground foundation [2]. Hurricane and tornado can have detrimental effects on power lines and buildings due to the differential pressures on roofs and walls, and wind-borne debris on windows and building facades. Damages to building envelope also lead to damages to building's interior from rain and storm.

While the occurrence of natural hazards cannot be precisely predicted, their impacts to buildings and infrastructures are fairly well understood. Damages and losses can be minimized with effective management and hazard mitigation planning. Mitigation measures can be identified, prioritized, and implemented through comprehensive risk assessment studies. Buildings represent a major portion of the built environment and are vulnerable to a broad variety of natural hazards. For regional planning, buildings are of major consideration for response planning and disaster management. The first step of regional hazard risk analysis is therefore to acquire the information about the buildings.

Building Information Modeling has been adopted widely by the architecture, engineering, and construction (AEC) professionals to efficiently plan, design, construct, and manage buildings and infrastructures. Building Information Modeling is defined as a process that involves the generation and management of "shared digital representation of physical and functional characteristics of any built object [...] which forms a reliable basis for decisions" (ISO 29481-1: 2016). For the building and construction industry, the building information model (BIM) generated represents a valuable source of shareable knowledge and information about a constructed facility to support decision making over the entire life cycle (LC) of the facility, from its conception to demolition (e.g., design, construction, maintenance, risk management, deconstruction, etc.). The information contained in BIM includes building geometry, components, material, etc., which can be adapted to generate computational models suitable for a broad range of computational simulations and applications. BIM can be an ideal source of information for natural hazard risk analyses. For example, based on information captured in the BIM, structural engineers can create models

* Corresponding author.
E-mail address: c_w@berkeley.edu (C. Wang).

https://doi.org/10.1016/j.autcon.2020.103474
Received 3 May 2020; Received in revised form 13 October 2020; Accepted 5 November 2020
Available online 8 December 2020

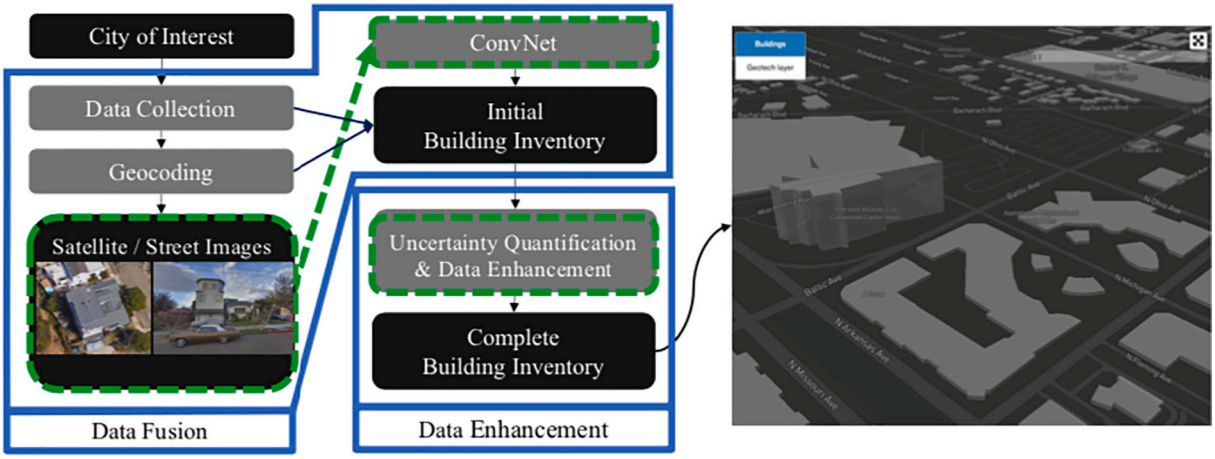


Fig. 1. Framework showing the components of the data generation workflow.

for dynamic simulations of buildings under seismic load conditions. Depending on an application, the level of details (LoD) of the information required can be quite different [13]. In addition to the basic BIM description, information about the usages of a building (e.g., commercial, residential, hospital, etc), its age (e.g., heritage, existing, new) and other conditions can have significant influence on the type of studies that can be conducted. This paper focuses on the building information that are needed for conducting natural hazard risk analysis.

To fully characterize the risk of an urban city affected by natural hazards, analyses need to be performed at multiple scales, from building level analysis to regional-based analysis. Analyses require data about the natural hazard considered as well as data on individual asset, such as buildings. However, in a regional scale, building data, even for some basic building information, such as number of stories, structural type, etc., for all the buildings in a city is scarce and difficult to obtain. Most buildings were designed and constructed before computer digitization was adopted and the concept of building information modeling was conceived. Even when the information is available, the data is often dispersed among different public and private organization. The data is heterogeneous in structure and does not follow specific standard formats. An efficient and low-cost approach that can effectively gather the existing information of buildings in a region and at the city-scale, and capture the LoD required for regional hazard analysis is needed.

Besides extracting data from existing building databases, data can also be acquired by using computer vision technologies to create BIMs for existing buildings. Image-based techniques (such as photogrammetry and videogrammetry), range-based techniques (such as laser scanning) or a combination of both have been employed for large-scale data acquisition. Range-based methods have mainly been focused on capturing the geometry of objects or scenes of interest rather than their semantic information [13]. Comparing to image-based methods, range-based methods have shown better accuracy in acquiring building geometry information, but they require high equipment cost and time consuming data processing and modeling steps [8,19].

Comparing to range-based methods, in addition to lower equipment cost and better processing efficiency, image-based approaches can take advantage of publicly available data sets. For example satellite images are now available for free or can be purchased from commercial vendors at reasonable prices. Street view images, which contain objects like buildings, trees, and cars, are also widely available from map service providers, such as Google Maps, for free or at low cost. Satellite and street view images have attracted attention of engineering researchers and social scientists alike. In particular, street view images have emerged as a popular resource for research not only because of their availability but also the rich visual information captured in the images. Researchers have found the potentials of street view images in many applications. For instance, [20] and [22] use deep learning to screen seismically vulnerable buildings from street view images. The trained classifier is able to capture the visual cues of buildings with significant geometric irregularities. [9] analyzed millions of street view images to predict the perceived safety level of a neighborhood. The visual appearance of an urban environment is used to capture the social surroundings and to derive their implication on the lives of residents. Similarly, the visual cues of cars in street view images have been used to estimate the demographic makeup of neighborhoods [3]. Most image-based studies, including applications in building detection, land usage classification, etc., utilize convolutional neural network (ConvNet or CNN) as the core to extract information from images [5,6].

The functionality of BIM depends on the level of details (LoD) and the technical specifications of the data that can be captured, processed and the model created. For natural hazard risk management at a regional scale, the LoD of a building model does not need to be high. This paper discusses the basic data needs necessary for regional hazard analysis. Creating a regional BIM database that contains sufficient information for hazard risk analyses can greatly enhance the ability to reduce the impact of natural disasters, and to better prepare, respond and recover from an aftermath. This study presents a modular framework for acquiring and integrating the building information from a multitude of sources and developing information models at a regional scale for hazard risk analysis, taking advantage of recent advances in image processing and machine learning.

2. A building information modeling framework for regional hazard analysis

This section provides an overview of a framework for developing a building information harvesting tool for regional hazard analysis and management. A standardized workflow is designed to support the creation of a building inventory database. The workflow includes a number of modules that implement the functionalities suitable for regional-scale building information modeling. The modules are generically defined according to the needed functionalities, are reusable, and can be augmented or replaced by the users with their own program codes and input data.

As shown in Fig. 1, the framework for creating a regional scale building inventory database consists of two major steps: a data collection and fusion step and a data enhancement step. The purpose of the data collection and data fusion steps is to gather the relevant data, such as images, point clouds, property tax records, crowd sourcing maps, etc. and to integrate the heterogeneous data for establishing an initial building inventory. The first step of the process is to collect the basic information of individual buildings, such as address, footprint, number

of stories, year of built, structure type, occupancy, etc., from available information sources that include crowd sourcing platforms (such as OpenStreet Map), third-party commercial companies (such as Zillow), property tax assessment websites, and other public administrative offices and data sharing service providers. The buildings are then geocoded with metadata (latitude and longitude) to create unique identification tags. Based on the geocoded tagged information, further building attributes are extracted from satellite and street images with pre-trained neural networks. All the collected and preprocessed data are fused and integrated to produce a consistent, accurate, and useful initial building inventory that would otherwise be directly available from any individual data source. The building inventory data is properly organized and structured for further processing and usages.

Because the data are gathered from different sources, each has its own purpose for data collection, significant amount of building information can remain missing from the initial building inventory after the data fusion step. For instance, crowd sourcing maps may contain relatively more complete sets of data in densely urbanized areas but less in rural regions. Property tax assessment records could be missing from administrative databases. In fact, from our experience, incomplete information is commonly found in almost all the public data sources that we have examined. In the framework, using the data captured in the initial inventory database, a data enhancement process is performed to fill in the missing values by additional modeling and machine learning methods. The following sections describe in details the data collection and fusion process and the data enhancement task for constructing the building inventory data suitable for regional hazard analysis and risk management application.

3. Data collection and fusion - constructing an initial building inventory

As shown in Fig. 1, the framework for establishing an initial building inventory database includes three basic tasks, namely (1) data collection and geocoding, (2) extracting relevant information from satellite or street images using for example pretrained convolution neural networks, and (3) integrating or fusing the data.

3.1. Data collection and geocoding

For a given city or a region of interest, the first step towards establishing a building inventory is to collect the building information from available data sources. Specifically, indexed metadata (e.g. building addresses or geographic coordinates) is collected to facilitate the search for individual building in the area. Typically, obtaining the indexed metadata for buildings within a city is rather straightforward, especially for dense urban areas. For example, OpenAddress (https://openaddresses.io/) maintains a collection of addresses globally, from which the building addresses within a region of interest can be downloaded. Another example is the United State Building Footprint database (https://github.com/microsoft/USBuildingFootprints), from which the geographic coordinates of buildings within the city of interest can be computed.

The process of geocoding is to tag a building with a pair of geographic coordinates (latitude and longitude). Geocoding is particularly useful because the geographic coordinates can be used to retrieve images of individual buildings from a satellite or a street view image database. Furthermore, as will be discussed later in section 4, the data enhancement step utilizes the coordinates of the buildings to conduct spatial distribution analysis for data correlation and inference as an attempt to fill in any missing information that are deemed important for regional hazard analysis. Geocoding services, such as Google Maps API or the MapQuest Geocoding API, are publicly available and application programming interfaces (APIs) are provided to facilitate the geocoding process.

With the indexed metadata, information about a building can be collected from multiple sources and gathered to create an initial building inventory database. For example, as illustrated in Fig. 2, basic building information, such as footprints, number of stories, building type and usage, etc., can be found and downloaded from Open Street Map (OSM), a public crowd sourced database. (For a full list of building properties in the OSM database, see https://wiki.openstreetmap.org/wiki/Key:building.)

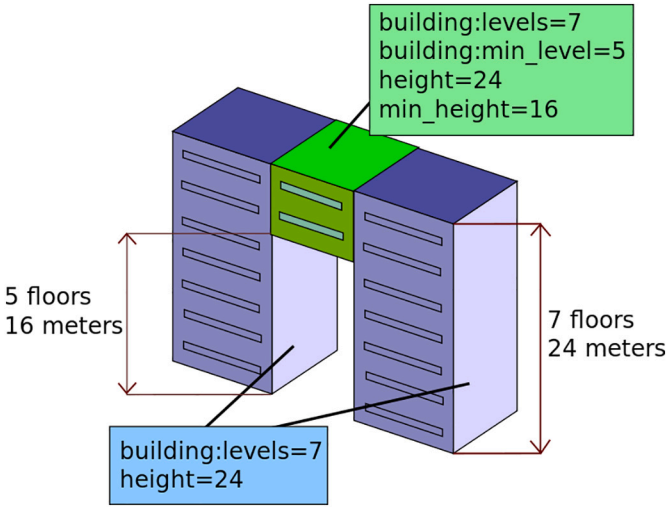


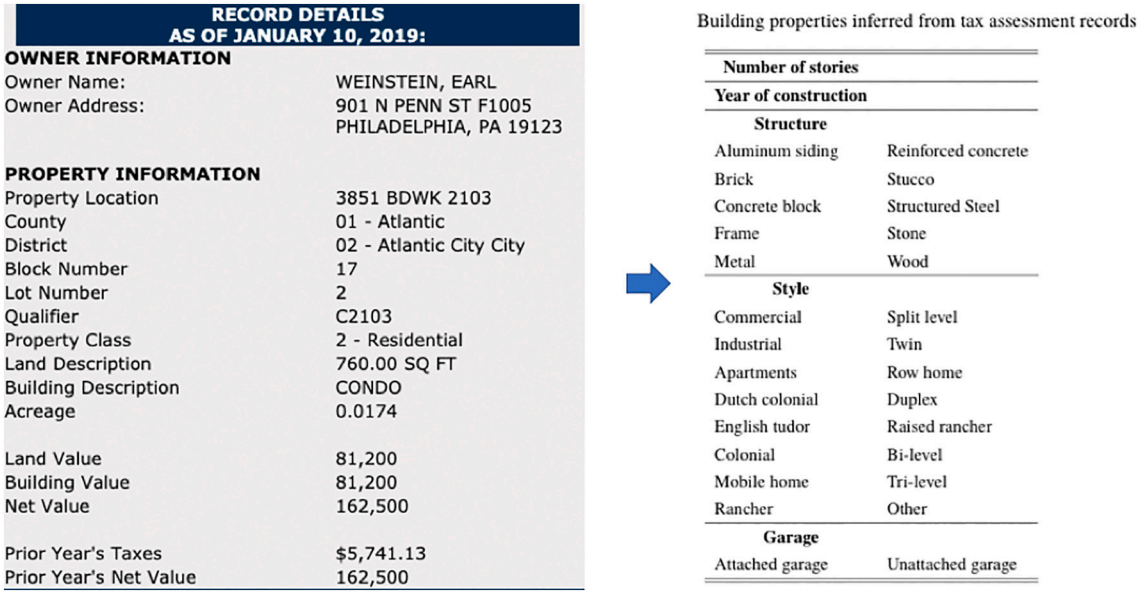
Fig. 2. OSM building (https://wiki.openstreetmap.org/wiki/File:Minlevel.svg).

Another example source of building information is the property tax assessment records, which are open to public and are available on various governmental websites. When tax assessors evaluate properties within a municipality, building information such as property value and many other descriptive data are recorded. For instance, the number of stories, exterior structural material, structural style, year of construction, type of garage, etc. can be extracted from the property tax assessment records. Fig. 3 shows an example of the tax assessment records downloaded from the website of the Department of Treasure of New Jersey (https://www.state.nj.us/treasury/taxation/lpt/TaxListSearchPublicWebpage.shtml) using the addresses of individual buildings. The property tax assessment records are a good starting source of information to establish a preliminary initial database for each building. However, it should be pointed out that the tax record on a building is often incomplete and a considerable amount of building information is missing from the tax records. For instance, about 30% of the tax records found in the aforementioned website do not contain the year of construction. As will be discussed in section 4, a machine learning-based predictive approach is developed that attempts to fill in as many as possible the missing values in the database.

3.2. Extracting building information from images

With the data collection and geocoding, the initial building inventory database contains the basic indexed information such as building addresses and geographic coordinates of individual buildings, as well as the basic descriptions of the buildings such as year built and structure type. Additional information such as material types and other building attributes can also be identified from images. Based on the indexed information, satellite or street view images of each building can be retrieved, for example, from Google Maps using the API provided. Deep learning techniques, such as convolution neural network (ConvNet) can be used to extract building attributes (that may not exist in the initial database) from the images. ConvNet has been widely used as a means to effectively analyzing images.

ConvNet belongs to a class of supervised learning algorithms that



RECORD DETAILS AS OF JANUARY 10, 2019:

OWNER INFORMATION

Owner Name:	WEINSTEIN, EARL
Owner Address:	901 N PENN ST F1005 PHILADELPHIA, PA 19123

PROPERTY INFORMATION

Property Location	3851 BDWK 2103
County	01 - Atlantic
District	02 - Atlantic City City
Block Number	17
Lot Number	2
Qualifier	C2103
Property Class	2 - Residential
Land Description	760.00 SQ FT
Building Description	CONDO
Acreage	0.0174
Land Value	81,200
Building Value	81,200
Net Value	162,500
Prior Year's Taxes	$5,741.13
Prior Year's Net Value	162,500

Building properties inferred from tax assessment records

Number of stories	
Year of construction	
Structure	
Aluminum siding	Reinforced concrete
Brick	Stucco
Concrete block	Structured Steel
Frame	Stone
Metal	Wood
Style	
Commercial	Split level
Industrial	Twin
Apartments	Row home
Dutch colonial	Duplex
English tudor	Raised rancher
Colonial	Bi-level
Mobile home	Tri-level
Rancher	Other
Garage	
Attached garage	Unattached garage

Fig. 3. Building information inferred from tax assessment record.

train and learn from previously labeled (known) information. As the images need to be labeled for training, one important task is to build a labeled dataset. The OSM platform hosts a collection real world infrastructure information labeled by human. For a typical building, the information that may be found in OSM includes building height, number of stories, structure type, exterior material, footprint shape, usage, etc. However, there are only a limited number of buildings that have been labeled in OSM. This study takes advantage of the labeled information and associate them with the images to build a data set for training the neural network. The ConvNets trained on this data set are then used to predict the building properties when given images contain unseen and unlabeled buildings. As long as the satellite/street view images of a city are available, the trained ConvNets can be applied as an attempt to predict the respective building properties or attributes of interest. The basic procedure to train and to extract a specific building property from the images can be summarized as follows:

1. Identify a visually comprehensible building property or attribute type (e.g., exterior construction material) that is intended to be extracted
2. Retrieve satellite / street view images of individual buildings from Google Map using the API provided
3. Label and tag the retrieved images with the attribute types (e.g., wood, concrete, brick, etc. for exterior construction material type) found in the OSM database
4. Train a ConvNet on the labeled images for type classification

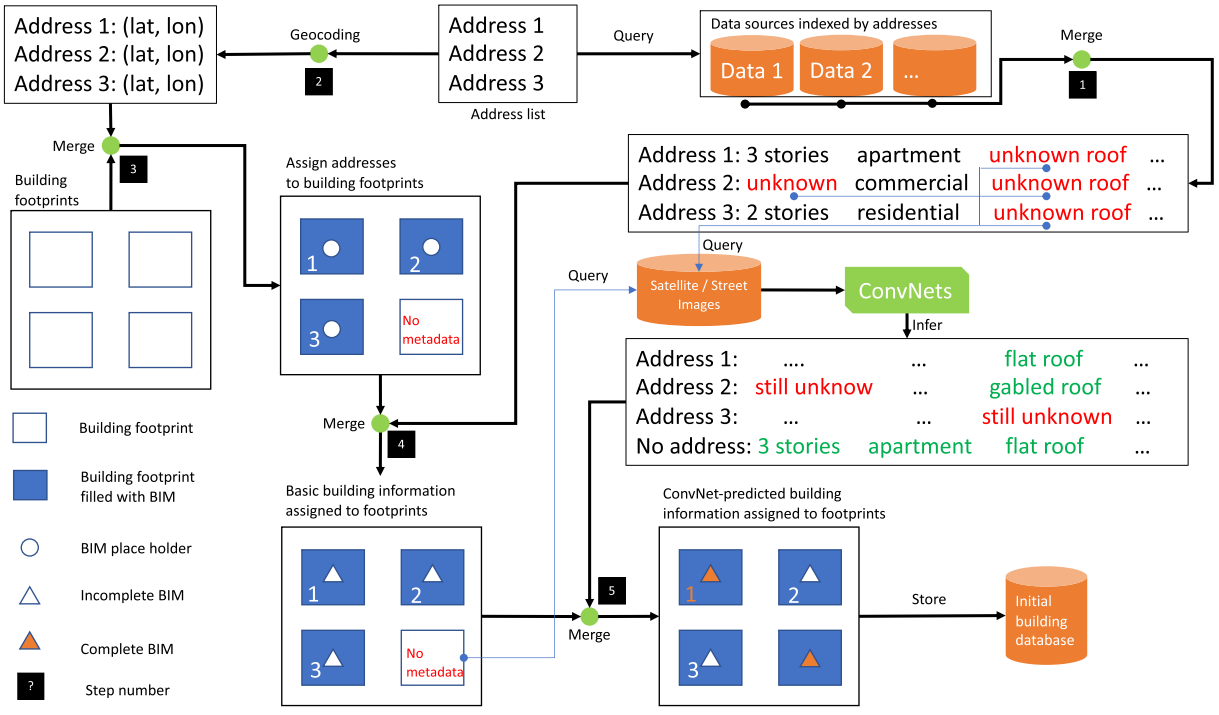


Fig. 4. Data fusion procedure to merge and enhance data from multiple sources.

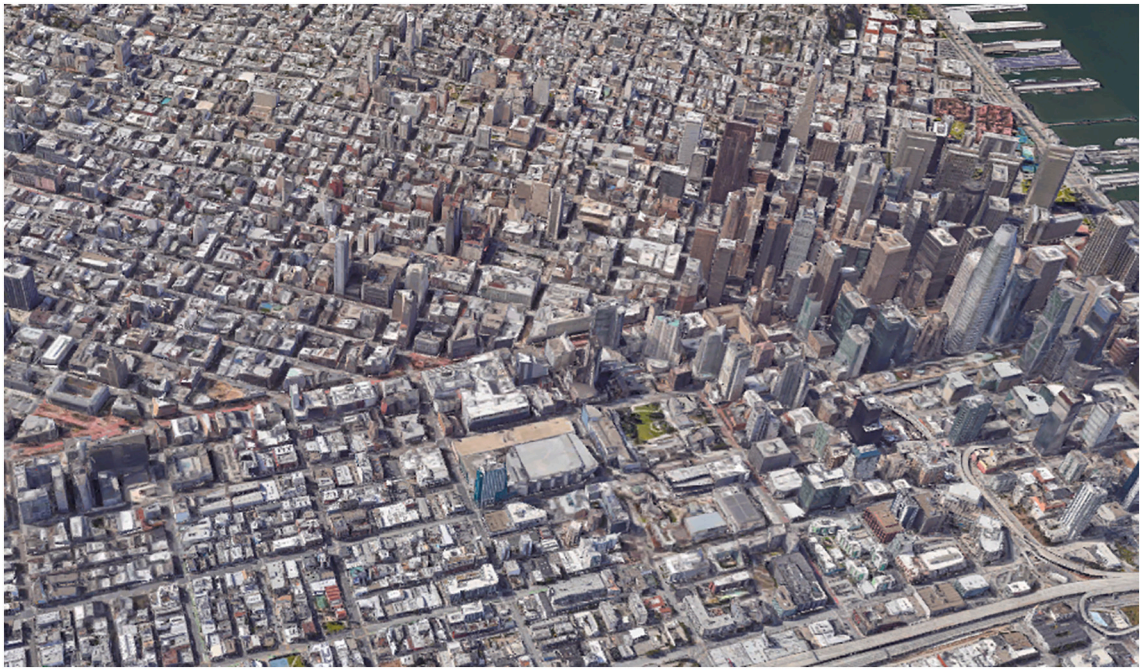


Fig. 5. A map showing the spatial distribution of buildings in San Francisco (Google Maps).

5. Apply the trained ConvNet to predict on satellite / street view images of buildings in the city of interest

The procedure can be repeated for other building attributes, such as number of stories, structural type, roof configuration, etc., as long as the attributes can be visually observable from the images. Different ConvNet models are trained and established for different building property or attribute types. While the performance of ConvNet models varies, depending on the architecture of ConvNet and the "wellness" of the trained data sets, the machine learning approach is a powerful apparatus to infer building attribute information from street and satellite images. The modular building information modeling framework can easily incorporate any ConvNet models developed for individual building attributes. The ConvNet-identified building information can then be merged into the initial database, resulting in a more detailed inventory data of buildings.

It should be noted that, due to reasons such as occlusions or bad viewpoints of the images, predictions from ConvNets do not always give acceptable confidence. In section 4, a machine learning-based approach that takes into account the spatial correlations of buildings in the region of interest will be discussed to further enhance the results for building the inventory database.

3.3. Data fusion

A data fusion procedure is designed to merge the building information retrieved from the different sources. The basic process is described as shown in Fig. 4. As shown in the figure, there are two starting points for the data merging process, namely the address list and the building footprints. As discussed earlier, the address list is used as the indexed metadata for querying the building information from the sources such as OSM, property tax assessment records and other data sets provided by the user. Once retrieved from the data source, the data is filtered and cleansed to remove duplicated attributes, to merge the data to the initial building inventory database and to put place holders on any of the missing data.

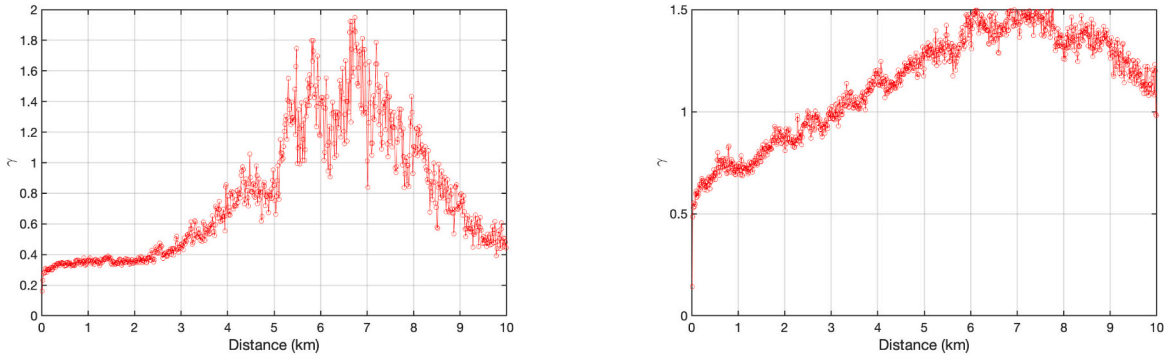
Building footprint is another important component of the city-scale building information modeling framework. In this study, the footprints employed are directly retrieved from a data set released by Microsoft. The data set contains the footprints of almost all the buildings in the United States extracted from high-resolution satellite maps. Prior works have also been reported using semantic segmentation to extract building footprints from high-resolution satellite maps [1,7,21]. In general, semantic segmentation can be applied to areas, for example many regions outside of the United States, where building footprints are not yet publicly available.

As discussed in Section 3.1, each geocoded address has a unique geographic coordinate based on longitude and latitude that can be used to merge the corresponding building footprints with the building inventory database. It should be noted that if the geocoded locations are not accurate, i.e., if a building cannot find a geocode attribute within the polygon of a footprint, the attribute will be inferred by SURF based on the neighboring buildings. The details are described in section 4. Furthermore, as discussed in Section 3.2, for buildings with missing attribute information, the pretrained ConvNets can be used to infer building features such as number of stories, roof types, etc., and to add the predicted values to the initial building inventory database.

To query a dataset means to operate on a dataset. This process uses an index (e.g., address, coordinates, ID, etc.) to search a dataset. The present framework is developed to help fuse different datasets. These datasets should be provided by the user. These datasets could contain any information about the buildings that is useful for analyzing the damage of the buildings during a hazard event. In Section 5, we showed an example of data needed for hurricane damage evaluation.

4. Building inventory database enhancement

The previous section describes the data collection and data fusion process for developing a building inventory database from publicly available data sources, such as property tax assessment records and street and satellite images. While the data is collected from multiple rich data sources, the database nonetheless is incomplete and contains considerable amount of missing data. For most public databases, building attribute values, such as year of construction, are often missing. Furthermore, some attributes of buildings may not be visually comprehensible from the images by the trained ConvNet due to, for example, occlusions in the street images. In the following, a methodology based on spatial statistics and machine learning is developed that aims to fill-in



(a) Number of stories (b) Year built

Fig. 6. Spatial patterns of building information expressed in semivariogram (The horizontal axis represents the distance between a pair of buildings, while the vertical axis represents the dis-similarity of these buildings.) These curves are calculated based on a building dataset covering five coastal cities in the Atlantic County, New Jersey.

the missing data in the database and to support city-scale regional hazard risk analysis.

Regional landscapes are seldom just random assortment of objects in space. Built environment and natural habitat often display certain orders and spatial patterns. Perceptually, they often exhibit certain spatial configurations and layouts with respect to the placements or arrangements of objects in the space. Patterns may be recognized based on distance measurements, line alignments, clustering of points or other spatial arrangements.

Spatial patterns, i.e., the arrangement of individual buildings and the geographic relationships among them, exist in the distribution of buildings in cities and neighborhoods. Buildings that are clustered or dispersed can often be identified by their attributes, such as building type, property value, construction material, etc., and are usually the manifestation of the demographic characteristics of neighborhoods, such as household income. As illustrated for the city map shown in Fig. 5, there are areas with denser number of buildings and clustering of certain types of buildings in certain regions. The capability of analyzing spatial patterns is an important step to gain a good understanding of the complicated spatial distribution of the underlying factors for a phenomenon of interest in a region.

In spatial statistics, a semivariogram function is commonly used to describe the degree of dependence of a spatially distributed random field or stochastic process, and is a useful indicator of similar or dissimilar spatial patterns. Essentially, a semivariogram measures the degree of dissimilarity between two observations as a function of the distance between them. Lower value of semivariogram means the observations are more similar to each other. The semivariogram function $\gamma(\boldsymbol{h})$ is defined as one half of the variance of two random variables separated by a vector distance $\boldsymbol{h}$ [4,12,16,17]:

$$\gamma(\boldsymbol{h}) = \frac{1}{2} Var[Z(\boldsymbol{\mu}) - Z(\boldsymbol{\mu}+\boldsymbol{h})] \tag{1}$$

where $Z(\boldsymbol{\mu})$ and $Z(\boldsymbol{\mu}+\boldsymbol{h})$ are the observations at spatial locations $\boldsymbol{\mu}$ and $\boldsymbol{\mu}+\boldsymbol{h}$, respectively. Based on the 'first rule'of geography that things closer together tend to be more similar than things that are far apart, the

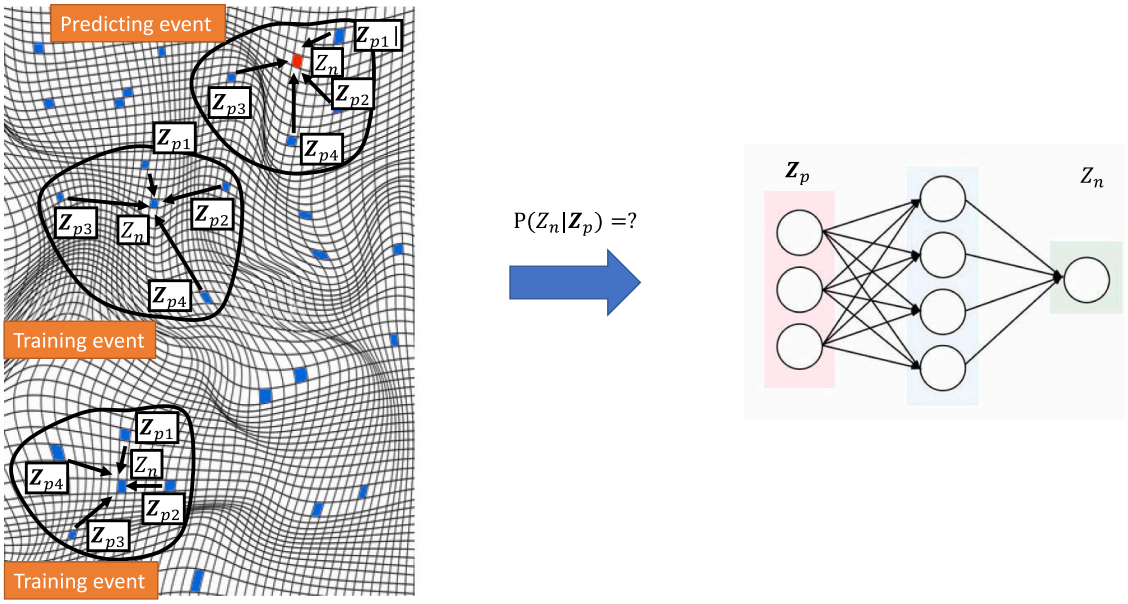


Fig. 7. Spatially distributed variables. (On the left side of the figure is a digram showing the locations of observations: The red dot Z_n represents the location of an object with unknown attributes; The blue dots represents the locations of objects with known attributes and are nearest to the unknown object Z_n. On the right side of the figure is a neural network to describe the relations between the observations and the value to be inferred.) (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. A neighborhood.

semivariogram is generally low when two locations are close to each other (i.e. observations at a point are likely to be similar to those in its neighboring points). Typically, semivariogram increases as the distance between two locations increases until at some point the locations are considered independent of each other and the semi-variance no longer increases. Under the same token, buildings farther apart from each other are expected to be more different than the buildings that are nearer to each other. Subprograms therefore provide a measure of how much two buildings are similar or dissimilar in their attributes (such as height, number of stories, etc.) based on the distance between the buildings.

The semivariogram function is applied to study the spatial patterns of the different features of buildings within a region. Indeed, the results show that, in general, buildings are indeed built following certain spatial patterns. To illustrate, the spatial subprograms of two building attributes, namely the number of stories and year of construction, are plotted as shown in Fig. 6, where the dissimilarities of a random pair of buildings are plotted against the distance between the buildings. The semivariogram figures show that with the increase of the distance between any two buildings, their semivariogram values regarding the number of stories and the year of construction, for examples, increase firstly and then fluctuate at high values in farther distances, which means the similarity decreases when increasing the distance until it exceeds a specific range, where there is no correlation and the semivariogram value starts to fluctuate. The relationships depicting the distances between buildings and their dis-similarities, however, are neither linear nor following any obvious function. It should be cautioned that the plots shown in Fig. 6 are city- or region-specific, i.e., the semivariogram curves may only apply to the region being investigated, and the same patterns may not be applicable to another region. In other words, the spatial dependence of building features are likely region-specific and the semivariogram curves would vary region by region.

The semivariograms shown in Fig. 6 suggest that there exist distribution patterns for certain building attributes and a possibility to map or extrapolate neighboring information for an individual building. Suppose, as shown on the left side of Fig. 7, there is a field of spatially distributed variables, in which the blue dots represent a collection of objects (e.g., buildings 1, 2, 3, …) whose attributes (e.g., number of stories) denoted by $\boldsymbol{Z}_p = \{\boldsymbol{Z}_{p1}, \boldsymbol{Z}_{p2}, \boldsymbol{Z}_{p3}, \ldots\}$ are known, while the red dot represents an object with a known location but its attribute value Z_n is unknown. The task is to train a neural network that takes $\boldsymbol{Z}_p$ as the input

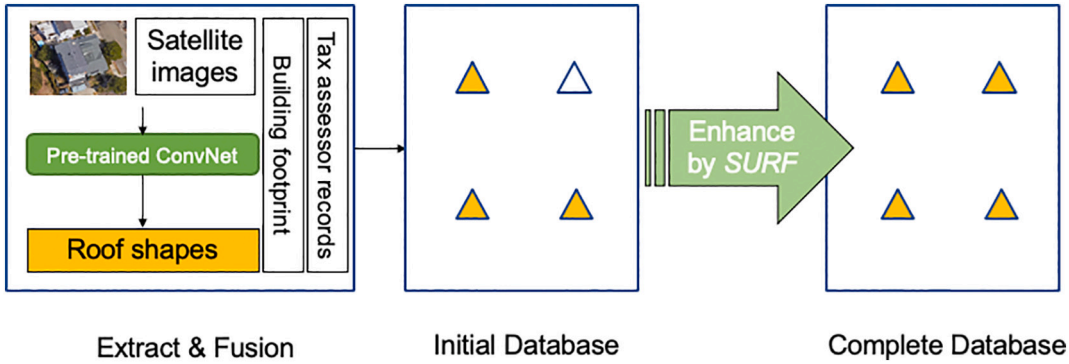


Fig. 9. An example of the implementation of the framework (white triangles represent buildings with incomplete information; yellow triangles stand for buildings with complete information). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

data and predicts Z_n as the target output. The data used for training is called training events, as depicted in Fig. 7. For each training event, there is a target object located at the center of the event and several nearby neighbors surrounding the target object. A number of events are first created and fed into the neural network for training. Once trained, the neural network can be used to predict the unknown attribute values of a target object using the known attribute values from a set of nearby neighbors. In other words, the trained neural network takes the neighboring objects with known attributes as the input and predict the "missing" attribute value for an unknown target object. The prediction events can be created for any number of unknown targets. The above process has been automated in the modular "Uncertainty Quantification & Data Enhancement" framework as shown in Fig. 1 based on an opensource code library "Spatial Uncertainty Research Framework", or "SURF", which is developed at the NHERI SimCenter to facilitate the study of spatial pattern with large data sets [14]. This library is originally designed to facilitate the pattern recognition of datasets in natural and social science, capturing the variations and nonlinearities in the spatial dependence structure. The first part of this paragraph showed the theory behind SURF. In the following, we use an example to further illustrate its applicability to our problems. More details about the usage of SURF can be found in section 5 and in the supplementary to this paper.

As discussed in the previous sections, significant amount of building information remain missing from the initial building inventory database after the data collection and data fusion process. The spatial statistical based neural network methodology can be applied to predict the missing values based on the known values of neighboring buildings. To illustrate, Fig. 8 shows an example of four neighboring buildings. The attributes, such as number of stories, occupancy, structure type, etc., of buildings A, C, D can be obtained from the collected data or extracted from the images of individual buildings using the pretrained ConvNets as discussed in section 3. However, for building B, which is heavily occluded by trees if viewed from the street, the information can not be extracted from the image of the building itself with sufficient confidence. Because the attributes of buildings within a community are likely well correlated with each other as illustrated in Fig. 6, the spatial machine learning approach can be used to infer the features of the building in the middle based on the information of its neighbors. In short, SURF can be used to effectively infer the missing attribute values of a building based on the known information of the neighboring buildings, thereby filling the gaps and enhancing the regional building inventory information database. More details about SURF can be found in the supplementary.

Though it might be true that the'First Law of Geography' dominates in many regions, the neural network is not trying to find just the first law. The neural network prediction is made based on the current location + the weighted neighbor information. (More details can be found in the supplementary.) Different from the traditional geostatistical method, the neural network here also considers the current location, which makes the algorithm aware of geo-locations. If a sub-region is dominated by the first law, then the neural network should be able to predict in this sub-region based on the first law it learned here. If a sub-region has strong patterns that disobey the first law, the neural network is then supposed to learn the specific patterns in this region. But if the disobeying pattern is not strong (i.e., only a small portion of the data disobeys), a neural network will treat such cases as noises during the training period. In a nutshell, the traditional geostatistical method is model-driven, so the data has to obey the first law. The neural network is data-driven, and it is capable of learning more complex patterns.

5. Application example

Risk analyses of different natural hazards (e.g., earthquake, wind, flood, wildfire) require different input data to describe the buildings. For example, it is important to know the foundation type of a building in order to better evaluate its seismic resistance during earthquakes, while it is not necessarily a required parameter for fire-related vulnerability analyses. Roof type is a crucial information for wind hazard analyses, yet it is not needed for flood risk assessments. Even for the same hazard (for example, earthquake), risk analyses could be conducted at different levels of fidelity, requiring different LoD of building information as the input data. Therefore, the specific building information to be collected should be decided by the investigator based on the needs. Rather than showing what kinds of information are needed for risk analysis of a specific hazard, the objective of this study is to present a framework that can be used to obtain specific building information when needed.

This section describes an example of regional wind risk analysis, in which we use the developed framework to create a building inventory database of several coastal cities in New Jersey that are threatened by hurricane hazards. The created building inventory is then fed into fragility functions for wind-caused damage and risk analyses. As illustrated in Fig. 9, three sources of data are retrieved and fused to build the initial building inventory; these include:

1. Tax assessor records from the administrative website of the Department of Treasury of New Jersey (https://www.state.nj.us/treasury/taxation/lpt/TaxListSearchPublicWebpage.shtml). As shown earlier in Fig. 3, the information available includes the number of stories, the year of construction, structure type, occupancy type, and other building attributes.
2. The publicly available United State Building Footprint dataset from Microsoft.
3. High-resolution satellite images obtained using the Google Maps API.

As discussed below, the base data obtained from the information sources alone are not sufficient for hurricane hazard assessment. The workflow described in the previous sections is employed to establish a

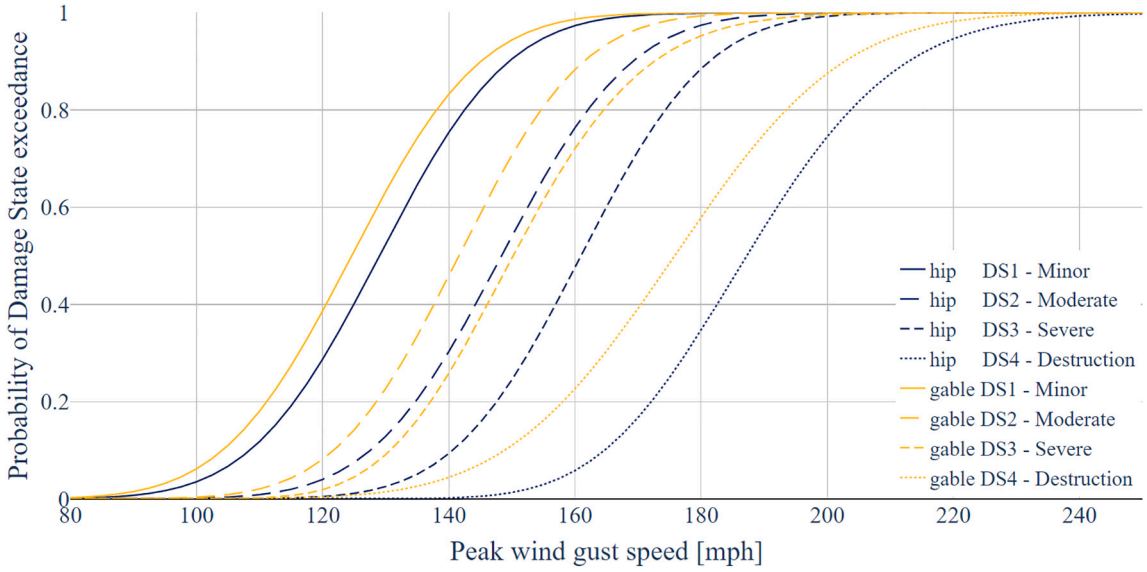


Fig. 10. Fragility functions corresponding to two structural configurations that are identical in all attributes, but the roof type.

Table 1
Attributes of wooden structures as per HAZUS MH 2.1.

Attribute	Valid entries
Occupancy type	single-family home | multi-unit housing, hotel, or motel
Number of stories	1 | 2 | 3+
Roof shape	gable | hip | flat
Secondary water resistance	yes | no | N/A
Roof cover	BUR | SPM | N/A
Roof quality	good | poor | N/A
Roof deck attachment	6d @ 6/12 | 8d @ 6/12 | 6d/8d mix | 8d @ 6/6
Roof-wall connection	strap | toe-nail
Shutters	yes | no
Garage	SFBC 1994 | standard | weak | no
Terrain roughness	open | light suburban | suburban | light trees | trees

regional building inventory database usable by commonly used hazard models, such as FEMA's HAZUS MH 2.1.

In this example, the damage models proposed by FEMA in HAZUS MH 2.1 are adopted to assess the effects of high winds and hurricanes on the building structures. The potential damage types are grouped into a number of damage states, each of which shows damages of similar type that requires similar amount of repair work and effort. As shown in Fig. 10, fragility functions are defined to relate wind speed with the probability of being in or exceeding a particular damage state.

HAZUS MH 2.1 provides fragility functions for a wide variety of structural types and building configurations, each is identified by a set of attributes. Table 1 shows an example of the attributes required by HAZUS MH 2.1 model to describe the 3520 wooden structures in the area. Although many of the attribute values are directly available from the existing data sources, many of the attributes and their values, however, are missing. For instance, the initial building inventory database from the tax assessment records include the data on the year of construction for only 70% of the buildings in the region of interest. Furthermore, roof type, a key attribute needed for consideration of wind effects on structures, is not available in the tax assessor records. Although we can ignore missing data and treat them explicitly as uncertainties, the regional damage estimates obtained from the hazard risk assessment models will also inherit the uncertainties and produce misleading results with further more uncertainties. For instance, Fig. 10 shows the significant variations in the probability of exceedance for the two set of fragility curves with two different roof types. The lack of information about the buildings will significantly hamper the use of the hazard models for regional damage estimates.

In an attempt to determine roof type for every building in a region, a ConvNet classifier, as illustrated in Fig. 11, is trained to take a satellite image of a building and predicts its roof type. A training data set of 6000 satellite images (2000 for each roof type: flat, gabled, hipped) is collected. Specifically, InceptionV3 [11], which is a widely-used ConvNet architecture for image feature recognition that has been shown to attain good results with an accuracy greater than 78.1% on the ImageNet dataset, is employed. The InceptionV3 model pretrained on the ImageNet dataset is taken as the initial model and transferred to train

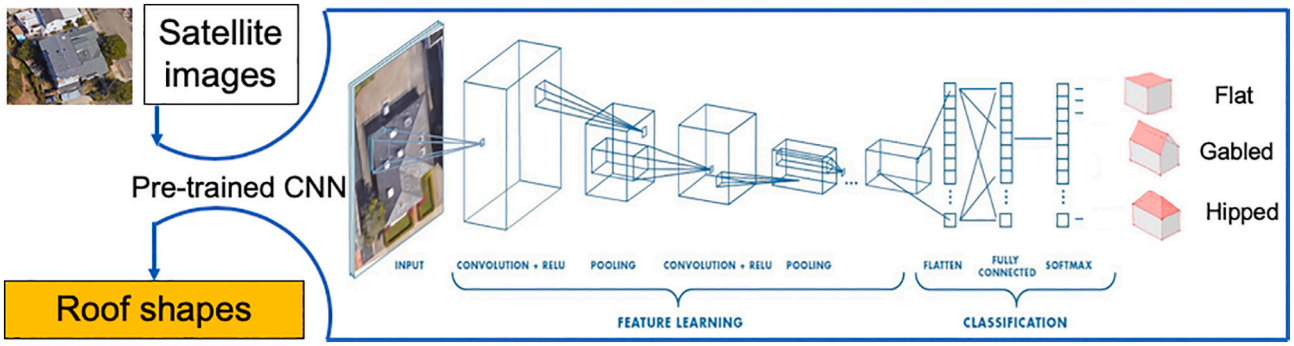


Fig. 11. Convolutional neural network for roof type classification.

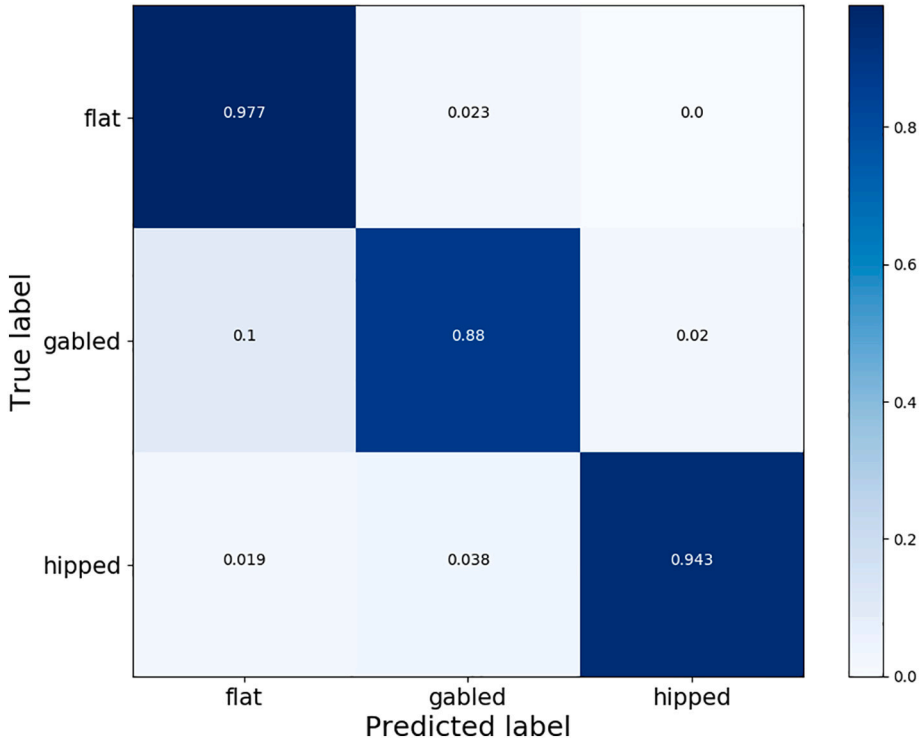


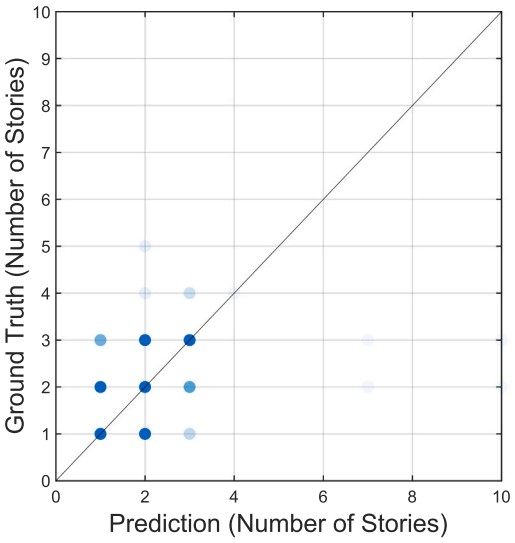
Fig. 12. Confusion Matrix.

Table 2
Performance of the roof classifier.

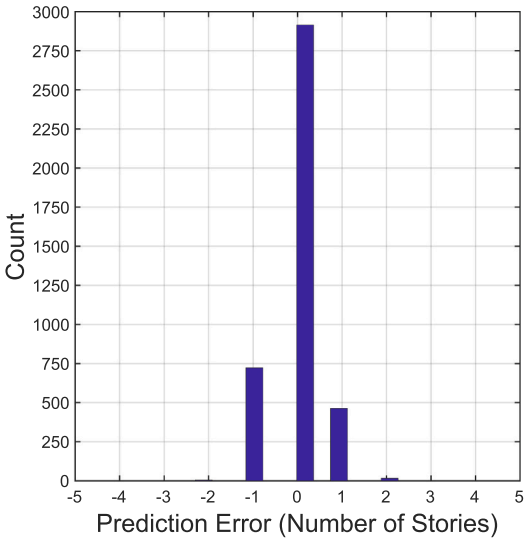
Class	Precision	Recall	F1-score	Overall accuracy
Flat	75.0%	97.7%	84.8	
Gabled	88.0%	88.0%	88.0	93.2%
Hipped	96.2%	94.3%	95.2	

the roof classifier using the 6000 satellite images of the region. When training InceptionV3, the initial learning rate and momentum are set to be 0.01 and 0.001, RMSProp optimizer is used with weight decay 0.00004. Each batch has 64 images. During training, we first fine-tune the final FC layer for 3000 iterations and then fine-tune all layers for another 20,000 iterations. All the experiments are implemented with Tensorflow and conducted on Maverick2 at Texas Advanced Computing Center (TACC). The computing node has 4 NVidia GTX 1080 Ti GPUs. This computing resource is made available to the author through NHERI

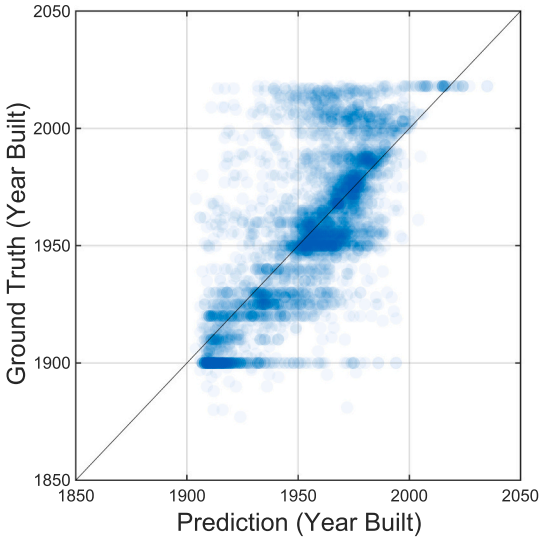
Fig. 13. Regional scale BIM database.



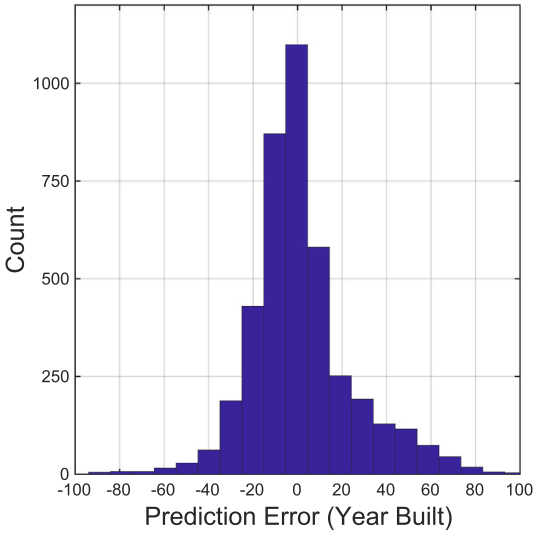
(a) Prediction error of story numbers, $MAE = 0.32$



(b) Histogram of prediction error of stories



(c) Prediction error of year built, $MAE = 16.39$



(d) Histogram of prediction error of year built

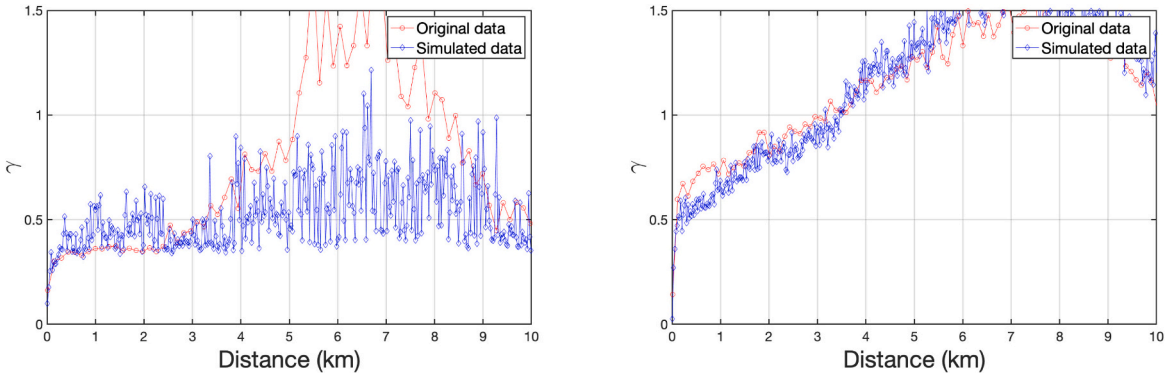
Fig. 14. Prediction errors.

DesignSafe [10]. An overall accuracy of 93.15% is obtained when the trained classifier is tested on a ground truth dataset [15]. The confusion matrix is plotted in Fig. 12. The precision, recall and F1 for each category are listed in Table 2.

Although it is not the focus of this work, it should be noted that, in addition to the InceptionV3 model pretrained on the ImageNet data set, other ConvNet models pretrained on different datasets can also be explored to potentially further improve the predictive performance. The new ConvNet models can be easily incorporated in the modular building information modeling framework.

As missing attribute information is determined, the framework is applied to fuse and enhance the initial building inventory to establish the regional BIM database. Note that many important attributes, such as structure type, roof type, occupancy, number of stories, year of construction, and footprint geometry, are collected, identified if necessary, fused and imported to the database. Fig. 13 shows the region with approximately 20,000 buildings that are included in the BIM database. The BIM data of a typical building is highlighted as shown on the right side of the figure.

To examine the current modules implemented in the framework, Fig. 14 shows the prediction errors for two example building attributes, namely the number of stories and the year of construction. It can be seen from the figure that the prediction errors for the number of stories are mostly zero. For the year of construction, the prediction errors are mostly within a 10-year period. Furthermore, Fig. 15 shows the comparison of the semivariograms of the predicted building attributes against those of the collected and observed data. As can be seen in the figure, the semivariograms agree well with each other, indicating that the spatial variations in the initial datasets from the original data sources are preserved in the enhanced datasets obtained using the data fusion and enhancement process.



(a) Number of stories (b) Year built

Fig. 15. Comparison of semivariograms between original data and simulated data.

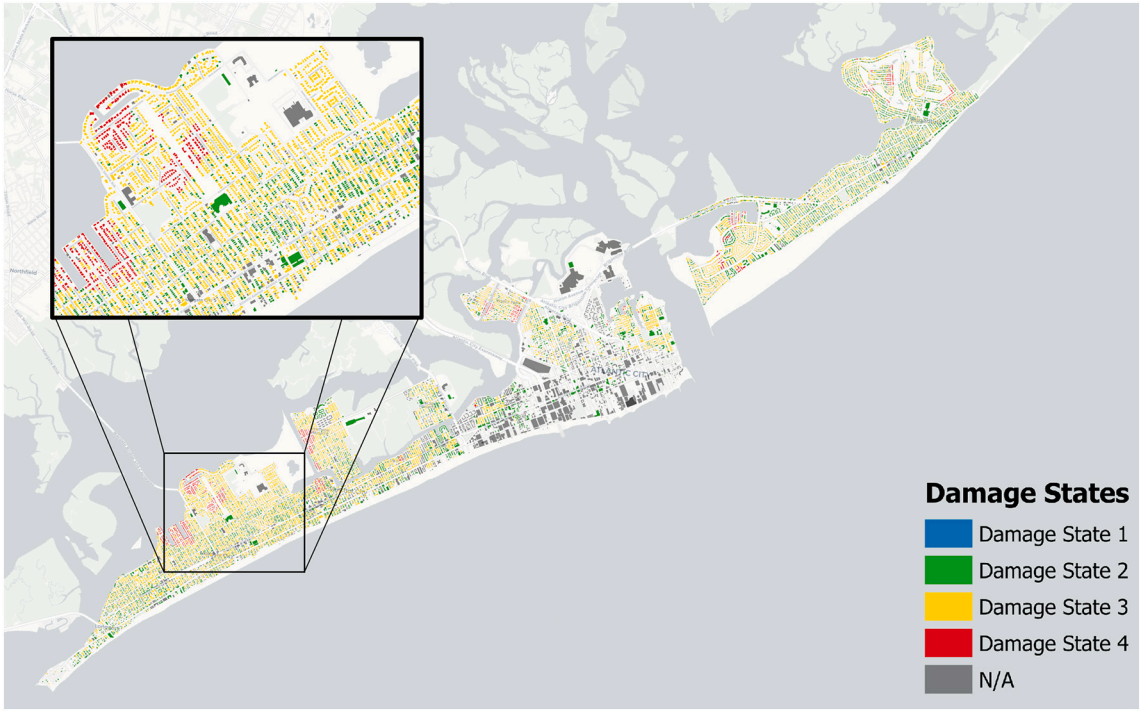


Fig. 16. Estimated damage state of wooden buildings after a Category 5 hurricane makes landfall in the Atlantic City area in New Jersey.

6. Discussion

The source code, the data, the pre-trained ConvNets, and the BIM database created in this study has been made available at [18]. The framework developed in this study can be applied to a broad range of applications in regional assessment analyses. First, additional building attributes of interest (e.g., window area, facade material, first floor elevation) can be extracted and added to the database. The data attributes can be acquired from data sources if available or obtained by training a ConvNet for each attribute, as long as the attribute type is visually comprehensible by the ConvNet from the images. For instance, it has been reported that geometric features (such as building outlines) can be extracted from images using ConvNets-based segmentation [1].

The NHERI SimCenter has developed rulesets that link the BIM information shown in Fig. 13 to design details, such as secondary water resistance or the type of roof-deck attachment, that are deemed important attributes of certain building types. The rulesets are developed based on the building codes that have been used in the State of New Jersey for quite some time. Given these rulesets, together with the enhanced BIM information produced by the presented framework as discussed in this study, we are able to build a more accurate description of the building stock in Atlantic City and four of its neighbors along the coast of New Jersey. In other words, more realistic damage and loss models can be applied to provide risk assessment estimates with less uncertainty. For instance, Fig. 16 shows the high-resolution damage maps that are enabled by the enhanced data obtained through the framework presented. Such damage maps can then be fed into loss models to predict and estimate repair costs. Using the loss models built at NHERI SimCenter, Fig. 17 shows the resulting loss estimates for the five cities as a percentage of building replacement cost using the fragility information in the HAZUS models.

7. Conclusions

This paper presents a novel city-scale building information modeling framework designed for regional natural hazard risk management. The

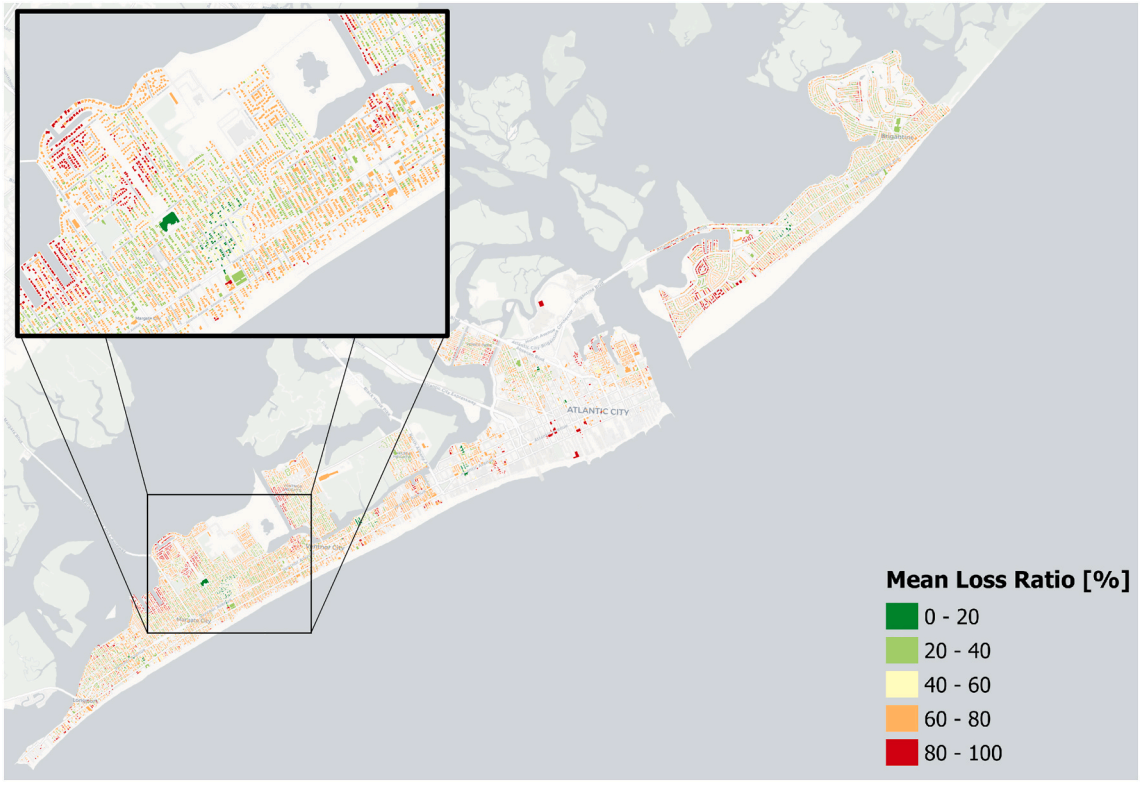


Fig. 17. Estimated losses in wooden buildings from a Category 5 hurricane in the Atlantic City area in New Jersey. Losses are expressed as a percentage of building replacement cost.

framework takes multiple sources of data as input and employs machine learning technologies to extract building information from the publicly available data sources and enhance the data to form an regional scale building inventory database. Firstly, the framework collects and fuses multiple sources of data (for examples, building metadata from the internet, satellite/street view images from Google Maps, building tags from OpenStreetMap, etc) to form an initial database, in which the BIM schema is defined with both semantic representations (such as structural type, number of stories, year of construction) and detailed geometric representations (such as building footprints). Furthermore, the framework includes an uncertainty quantification module (SURF) that employs machine learning to quantify missing information, to fill in the gaps and to enhance the building inventory database. Since the main input for the framework is image data which is readily available, and is relatively inexpensive to collect but contains ample valuable building information, both the level of details and cost-efficiency are achieved. Last but not least, the framework is designed for large scale BIM applications and the construction of region-based inventory database.

To illustrate the utility of the framework, a regional building inventory database with sufficient BIM details is created for hurricane hurricane damage and loss assessment covering five coastal cities in the State of New Jersey. The ability to detect features of infrastructures from images at a regional scale can find many important applications for disaster management planning and simulations. In addition, the BIM databases created using the framework have the potential to provide the data and to support the application tools for solving a wide range of social and economic problems of cities.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study is based upon work supported by the National Science Foundation under Grant No. 1612843. NHERI DesignSafe [10] and Texas Advanced Computing Center (TACC) are acknowledged for the generous allotment of compute resources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.autcon.2020.103474.

References

[1] B. Bischke, P. Helber, J. Folz, D. Borth, A. Dengel, Multi-task learning for segmentation of building footprints with deep neural networks, in: 2019 IEEE International Conference on Image Processing, 2019, pp. 1480–1484, https://doi.org/10.1109/ICIP.2019.8803050.
[2] G. Deierlein, A. Zsarnóczay, et al., State-of-Art in Computational Simulation for Natural Hazards Engineering, 2019, https://doi.org/10.5281/zenodo.2579582 (February).
[3] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E.L. Aiden, L. Fei-Fei, Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states, Proc. Natl. Acad. Sci. 114 (50) (2017) 13108–13113, https://doi.org/10.1073/pnas.1700035114.
[4] P. Goovaerts, Geostatistics for Natural Resources Evaluation, Oxford University Press, 1997 (ISBN:9780195115383).
[5] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Proces. Syst. (2012) 1097–1105, https://doi.org/10.1145/3065386.
[6] S. Lawrence, C.L. Giles, A.C. Tsoi, A.D. Back, Face recognition: A convolutional neural-network approach, IEEE Trans. Neural Netw. 8 (1) (1997) 98–113, https://doi.org/10.1109/72.554195.
[7] W. Li, C. He, J. Fang, J. Zheng, H. Fu, L. Yu, Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data, Remote Sens. 11 (4) (2019) 403, https://doi.org/10.3390/rs11040403.
[8] T. Mill, A. Alt, R. Liias, Combined 3D building surveying techniques–terrestrial laser scanning (TLS) and total station surveying for BIM data management purposes, J. Civ. Eng. Manag. 19 (sup1) (2013) S23–S32, https://doi.org/10.3846/13923730.2013.795187.

[9] N. Naik, J. Philipoom, R. Raskar, C. Hidalgo, Streetscore-predicting the perceived safety of one million streetscapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, https://doi.org/10.1109/CVPRW.2014.121.
[10] E.M. Rathje, C. Dawson, J.E. Padgett, J.-P. Pinelli, D. Stanzione, A. Adair, P. Arduino, S.J. Brandenberg, T. Cockerill, C. Dey, et al., Designsafe: new cyberinfrastructure for natural hazards engineering, Nat. Hazards Rev. 18 (3) (2017), https://doi.org/10.1061/(ASCE)NH.1527–6996.0000246, 06017001.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826, https://doi.org/10.1109/CVPR.2016.308.
[12] E. Vanmarcke, Random fields: analysis and synthesis, World Sci. (2010), https://doi.org/10.1142/5807.
[13] R. Volk, J. Stengel, F. Schultmann, Building Information Modeling (BIM) for existing buildings—Literature review and future needs, Autom. Constr. 38 (2014) 109–127, https://doi.org/10.1016/j.autcon.2013.10.023.
[14] C. Wang, NHERI-SimCenter/SURF: v0.2.0, 2019, https://doi.org/10.5281/zenodo.3463676.
[15] C. Wang, Random Satellite Images of Buildings, 2019, https://doi.org/10.5281/zenodo.3521067.
[16] C. Wang, Q. Chen, A hybrid geotechnical and geological data-based framework for multiscale regional liquefaction hazard mapping, Géotechnique 68 (7) (2018) 614–625, https://doi.org/10.1680/jgeot.17.P.074.
[17] C. Wang, Q. Chen, M. Shen, C.H. Juang, On the spatial variability of cpt-based geotechnical parameters for regional liquefaction evaluation, Soil Dyn. Earthq. Eng. 95 (2017) 153–166, https://doi.org/10.1016/j.soildyn.2017.02.001.
[18] C. Wang, Q. Yu, F. McKenna, B. Cetiner, S.X. Yu, E. Taciroglu, K.H. Law, NHERI-SimCenter/BRAILS: v1.0.1, 2019, https://doi.org/10.5281/zenodo.3483208 (October).
[19] A. Watson, Digital buildings–challenges and opportunities, Adv. Eng. Inform. 25 (4) (2011) 573–581, https://doi.org/10.1016/j.aei.2011.07.003.
[20] Q. Yu, C. Wang, B. Cetiner, S.X. Yu, F. Mckenna, E. Taciroglu, K.H. Law, Building information modeling and classification by visual learning at a city scale, in: 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, 2019, https://doi.org/10.5281/zenodo.3996808.
[21] K. Zhao, J. Kang, J. Jung, G. Sohn, Building extraction from satellite images using mask r-cnn with building boundary regularization, in: Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 247–251, https://doi.org/10.1109/CVPRW.2018.00045.
[22] Q. Yu, C. Wang, F. McKenna, S.X. Yu, E. Taciroglu, B. Cetiner, K.H. Law, Rapid visual screening of soft-story buildings from street view images using deep learning classification, Earthquake Engineering and Engineering Vibration 19 (4) (2020) 827–838, https://doi.org/10.1007/s11803-020-0598-2.