DOI: 10.1111/mice.12658

RESEARCH ARTICLE

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

Deep learning for post-hurricane aerial damage assessment of buildings

Chih-Shen Cheng[1] | **Amir H. Behzadan**[2] | **Arash Noshadravan**[1]

[1] Zachry Department of Civil & Environmental Engineering, Texas A&M University, College Station, TX, USA

[2] Department of Construction Science, Texas A&M University, College Station, TX, USA

Correspondence
Arash Noshadravan, Zachry Department of Civil & Environmental Engineering, Texas A&M University, College Station, TX, USA
Email: noshadravan@tamu.edu

Abstract

This study aims to improve post-disaster preliminary damage assessment (PDA) using artificial intelligence (AI) and unmanned aerial vehicle (UAV) imagery. In particular, a stacked convolutional neural network (CNN) architecture is introduced and trained on an in-house visual dataset from Hurricane Dorian. To account for the ordinality of damage level classes, the cross-entropy classification loss function is replaced with the square of earth mover's distance (EMD^2) loss. The trained model achieves 65.6% building localization precision and 61% (90% considering ±1 class deviation from ground-truth) classification accuracy. It also exhibits a positive accuracy–confidence correlation, which is valuable for model assessment in situations where ground-truth information is not readily available. Finally, the outcome of damage assessment is compared with the literature by examining the relationship between building size and number of stories, and severity of induced disaster damage.

1 | INTRODUCTION

1.1 | Background

Following a disaster event, preliminary damage assessment (PDA) is used to assign damage measurements to buildings, facilities, and infrastructure based on functionality and visible damage to the structure (FEMA, 2020). Past research has introduced several rating systems to assess disaster-induced building damage, for example, HAZUS, Federal Emergency Management Agency (FEMA) (FEMA, 2003), Kelman (Kelman, 2003), and EMS-98 (Grünthal, 1998). As listed in Table 1 and illustrated in Figure 1, FEMA guidelines, in particular, provide quantitative and straightforward definitions of five damage states (ranging from 0 through 4) which increase with the severity of the damage. To classify structural damage induced by hurricanes and storm surge, Friedland, Adams, and Levitan (2007) supplemented the standard FEMA scale with flood depth as an additional classification criterion.

Depending on the size of the affected area and the extent of damage, PDA by ground crews (e.g., windshield surveys) can be tedious and resource-intensive. Following Hurricane Irma in Florida, for example, several multi-personnel teams that were deployed to the field took almost six days to cover regions around the Gulf Coast, Atlantic Coast, and Miami/Florida Keys (Pinelli et al., 2018). Besides speed, successful ground-based PDA requires uninterrupted access to the affected site, which in the immediate aftermath of a disaster, may not be possible due to the presence of hazardous materials (e.g., chemical spills or fumes) or debris. The tradeoff between speed, accuracy, and safety has motivated new practices, including remote and aerial sensing (Gupta et al., 2019). Remote sensing using images captured by satellites or unmanned

TABLE 1 Description of FEMA damage states (FEMA, 2003)

Damage state	Quantitative damage description	Roof cover failure	Window door failures	Roof deck	Missile impacts on walls	Roof structure failure	Wall structure failure
0	No damage or very minor damage Little or no visible damage from the outside. No broken windows, or failed roof deck. Minimal loss of roof over, with no or very limited water penetration.	<2%	No	No	No	No	No
1	Minor damage Maximum of one broken window, door, or garage door. Moderate roof cover loss that can be covered to prevent additional water entering the building. Marks or dents on walls requiring painting or patching for repair.	>2% and < 15%	One window, door failure	No	<5 impacts	No	No
2	Moderate damage Major roof cover damage, moderate window breakage. Minor roof sheathing failure. Some resulting damage to interior of building from water.	>15% and < 50%	>1% and the larger of 20% and 3	1 to 3 panels	5 to 10 impacts	No	No
3	Severe damage Major window damage or roof sheathing loss. Major roof cover loss. Extensive damage to interior from water.	>50%	> the larger of 20% & 3 and < 50%	>20 impacts	>20 impacts	No	No
4	Destruction Complete roof failure and/or, failure of wall frame. Loss of more than 50% of roof sheathing.	>50%	>50%	>25%	>20 impacts	Yes	Yes

aerial vehicles (UAVs), in particular, can significantly expedite the process of damage assessment. Aerial PDA also provides valuable information to first responders, search and rescue (SAR) teams, emergency managers, and other critical stakeholders involved in disaster response and mitigation. Thomas, Kareem, and Bowyer (2014) extract features of aerial pre- and post-storm photos (e.g., edge, color, intensity) to predict the damage type of roofs such as missing tiles, collapsed roof, and holes. Gupta et al. (2019) utilized high-altitude satellite imagery to classify damage states by analyzing pre- and post-disaster photos.

While satellite data can cover large (and remote) areas and requires minimal local planning, the frequency of data collection depends on the satellite's orbital period, and object visibility in satellite photos may be limited due to cloud cover, haze, and smoke (Adams, Friedland, & Levitan, 2010; Yu, Yang, & Li, 2018). Compared to satellite imagery, UAVs can capture images with higher spatial resolution due to lower flying altitudes (Adams et al., 2010; Iizuka et al., 2018). This is ideal for building damage assessment since roofs, walls, and other building elements remain sufficiently visible. Another advantage of using UAVs is that data collection and response can be scaled up through information sharing with multiple stakeholders on site (e.g., FEMA, Red Cross, law enforcement, nonprofits, volunteers). This will ultimately lead to more heterogeneous (thus, generalizable), cost-effective, and high-frequency spatiotemporal data that can be used in all phases of disaster management (Altay & Green, 2006).

In this study, the FEMA rating system is adopted for classifying the damage scale of disaster-affected buildings in UAV imagery. In addition to the standard damage states 0 through 4, a new damage level 5 is added to capture buildings that are completely destroyed or under the early stages of construction (Figure 1). A primary challenge in utilizing this system, however, is that damage level 5 is largely

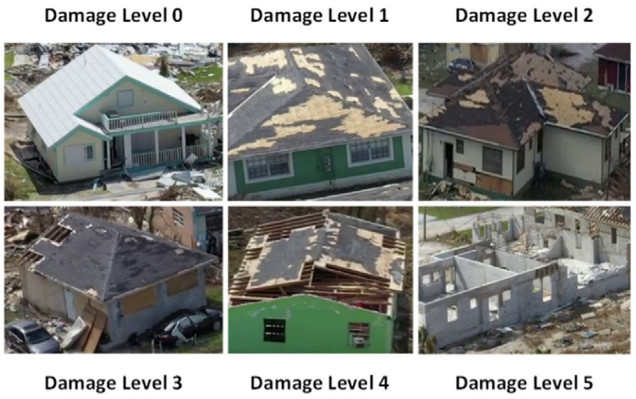


FIGURE 1 Examples of damage level 5, added to FEMA's damage levels 0 through 4

insensitive to the differences between destroyed and under-construction buildings, both sharing similar visual features (e.g., incomplete walls, no windows or roof). A potential remedy to this problem is to post-process detections labeled as damage state 5 by projecting model predictions on geocoded maps of the same region (e.g., Google Maps, OpenStreetMap) and identifying discrepancies between pre- and post-disaster objects. A destroyed building is most likely shown on the map as an intact dwelling (prior to disaster), whereas a building under construction either does not exist or is depicted as a partially completed structure on the map. A detailed discussion of this topic is not within the scope of this paper but can be found in another publication by the authors (Pi, Nath, & Behzadan, 2020b). It should be also noted that since the FEMA rating system is primarily created for ground-based PDA, a comparison of results obtained from ground assessment and aerial assessment can be conducted to further validate the results.

With advancements in computing power and sensing technologies, artificial intelligence (AI), and primarily deep learning, is being increasingly utilized to solve complex problems in construction management (Rafiei & Adeli, 2018a; Rafiei, Khushefati, Demirboga, & Adeli, 2017), disaster response (Rafiei & Adeli, 2017a), and structural health monitoring (Rafiei & Adeli, 2018b). Some examples of deep learning applications for damage detection and assessment include road crack detection (K. Zhang, Cheng, & Zhang, 2018; A. Zhang et al., 2017;), rust grade classification (Xu, Gui, & Han, 2020), reinforced concrete (RC) bridge inspection (Liang, 2019), damage detection in high-rise buildings (Rafiei & Adeli, 2017b), structural damage detection (Abdeljaber, Avci, Kiranyaz, Gabbouj, & Inman, 2017; Gao & Mosalam, 2018; Kang & Cha, 2018; Y.-Z. Lin, Nie, & Ma, 2017), structural damage classification (Wang, Zhao, Li, Zhao, & Zhao, 2018), and multi-class damage detection for RC buildings (Ghosh Mondal, Jahanshahi, Wu, & Wu, 2020). Additionally, Lenjani, Yeum, Dyke, and Bilionis (2020) used a region-based convolutional neural network (CNN) to detect buildings in 2D ground-level images for post-disaster evaluation. However, there is limited body of work on aerial PDA using CNNs. Xu, Lu, Li, Khaitan, and Zaytseva (2019) presented a CNN that recognizes damaged buildings in satellite images. Nex, Duarte, Tonolo, and Kerle (2019) developed a deep learning model to detect damaged building sections in aerial imagery. They preprocessed the input image by dividing it into several equal-sized square patches. An AI model was then used to classify the content of each patch into three categories of damaged building, intact building, or no building with an 80% accuracy. Pi, Nath, and Behzadan (2020a) trained CNN models for detecting the location of damaged and undamaged roofs in hurricane footage and achieved a mean average precision (mAP) of ~45% in testing. These models, however, merely generate a binary output indicating the presence or absence of visual damage and do not differentiate between various damage states.

1.2 | Research scope and core contributions

The research presented in this paper investigates the extent to which PDA can be automated by conducting multi-state damage assessment using visual recognition (detection and segmentation of building damage) from aerial footage. For the specific application domain of this research (i.e., disaster response), it is anticipated that full annotations of new disaster footage may not be readily available due to time constraints and limited human resources. Therefore, this research also examines the potential correlation between prediction confidence and precision (i.e., quality) to verify if performance can be assessed solely based on confidence in the absence of ground-truth data. Understanding this relationship is essential to refining deep learning models, making them more predictable and explainable. Explainability, in the context of AI, is defined as the degree to which humans can interpret the output of the model (Gunning, 2017). While it is important for humans to understand and trust the output of AI tools (Samek, Wiegand, & Müller, 2017), explainability is even more critical when decisions made based on these tools can be consequential (e.g., saving victims trapped in a collapsed structure after an earthquake).

In light of this, the core contribution of this study is trifold: (1) It presents and describes an in-house visual dataset annotated in accordance with FEMA damage scales using footage from the 2019 Hurricane Dorian. This dataset can be used in future research and applications in the area of AI-enabled disaster response and mitigation;

TABLE 2 Dataset description

		Hurricane	Year	Location	Duration (s)	Resolution
Dataset 1 (Train and test)	Video 1-1	Dorian	2019	Marsh Harbor	301	1280 × 720
	Video 1-2	Dorian	2019	Marsh Harbor	301	1280 × 720
	Video 1-3	Dorian	2019	Marsh Harbor	301	1280 × 720
Dataset 2 (Unseen test)	Video 2-1	Dorian	2019	Great Guana Cay	301	1920 × 1080
	Video 2-2	Dorian	2019	Great Guana Cay	301	1920 × 1080

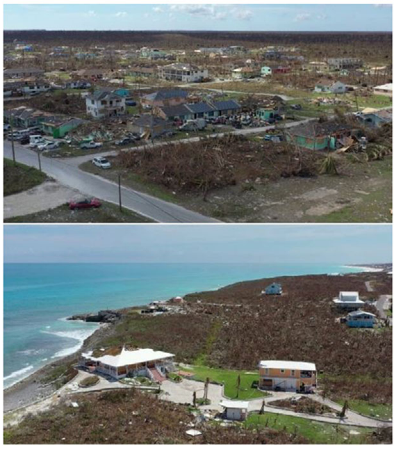


FIGURE 2 Sample images from Dataset 1 (top) and Dataset 2 (bottom)

(2) It develops a novel stacked CNN architecture for visual recognition of building damage, which increases output quality and reduces the likelihood of erroneous predictions; (3) It replaces the commonly used cross-entropy classification loss with a more robust and suitable form of loss function designed for ordinal class labels, thus accounting for the inter-class relationships.

2 | METHODOLOGY

2.1 | Dataset preparation

2.1.1 | Data collection

In this study, web mining with keyword search queries (e.g., disaster, hurricane, tornado, drone, UAV) is used to create a video dataset from the aftermath of Hurricane Dorian, which made landfall in the Bahamas in August 2019. Extracted videos are examined to ensure that they contain bird's eye views of disaster-affected residential buildings with various degrees of damage. Figure 2 shows sample frames from these videos. Overall, the dataset contains 5 UAV videos captured at 30 frames per second (FPS). Three of these videos have a 1,280 × 720 resolution, and the other two have a 1,920 × 1,020 resolution. The videos are captured from locations in Marsh Harbor and Great Guana Cay in the Bahamas.

Table 2 presents a detailed description of the dataset. To assess the generalizability of the trained models to new situations, they are first trained and tested on Dataset 1, including videos 1-1, 1-2, and 1-3 from Marsh Harbor. Next, the trained models are tested (without retraining) on Dataset 2, including videos 2-1 and 2-2 from Great Guana Cay. Dataset 2 is kept unseen from the trained models and used for the purpose of verifying if the CNN models trained on footage from one location can make acceptable predictions in a second location without retraining.

2.1.2 | Data annotation

To achieve a more visually heterogeneous dataset, considering the video speed of 30 FPS, one of every 10 consecutive frames in each video is selected for annotation (i.e., three frames from every second of the video). In doing so, frames that contain no building object are skipped. Annotation is done in Labelbox (Labelbox, 2019), with each frame taking approximately 5 minutes to annotate, although this varies based on the level of annotation difficulty and the number and diversity of building objects in each frame. In total, annotating 279 frames (1,501 buildings) in Dataset 1 and 60 frames (301 buildings) in Dataset 2 takes about 28 hours. Video frames are labeled by drawing masks along their boundaries (Figure 3), and assigning labels representing damage states (from 0 to 5). Table 3 shows the distribution of damage level annotations in both datasets. Dataset 1 is further split into training (60%), validation (20%), and testing (20%) sets.

As shown in Table 4, each instance in Dataset 1 is also assigned a single tag describing the difficulty level (easy, moderate, difficult) of labeling effort and the amount of time and personal judgment of a human annotator for labeling. The "easy" level represents cases where damage level can be assigned with high confidence within approximately 10 seconds. The "moderate" level represents cases for which the labeler can decide a damage level with

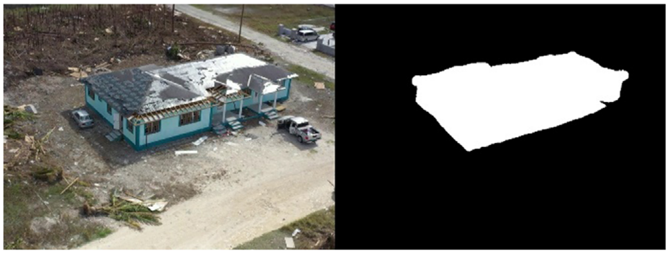


FIGURE 3 Sample image (left) and the corresponding mask (right)

TABLE 3 Distribution of annotation damage levels

		Damage level					
		Level 0	Level 1	Level 2	Level 3	Level 4	Level 5
Dataset 1 (Train and test)	Video 1-1	16	22	34	49	34	13
	Video 1-2	123	142	99	228	103	63
	Video 1-3	32	114	63	201	134	31
Dataset 2 (Unseen test)	Video 2-1	52	19	16	23	49	24
	Video 2-2	43	22	10	18	16	9

high confidence but may need more than 10 seconds of labeling time. Lastly, "difficult" denotes cases for which the annotator cannot easily decide the damage level (is in doubt between two or more levels), even when given enough time. Annotation difficulty is later used to perform sensitivity analysis and statistical testing on the performance of the developed model.

2.1.3 | Comparison of annotation and post-disaster field assessment

Door-to-door PDA in Harsh Harbor was conducted by a ground human team named FAST-1 after Hurricane Dorian using the FEMA rating system (Marshall et al., 2019). Table 5 shows a close agreement in damage level distribution between our annotation results and the findings of Marshall et al. (2019). This comparison validates the annotation quality, and further shows that UAV imagery can potentially lead to similar results to ground assessment of building damage levels.

TABLE 4 Distribution of annotation difficulty levels in Dataset 1

	Annotation effort		
	Easy	Moderate	Difficult
No. of samples	498 (35%)	580 (41%)	353 (24%)

TABLE 5 Distributions of annotations in Dataset 1 and field assessment

	Damage level					
	0	1	2	3	4	5
This work (%)	11	19	13	32	18	7
FAST-1 (%) (Marshall et al., 2019)	4	20	11	33	31	n/a

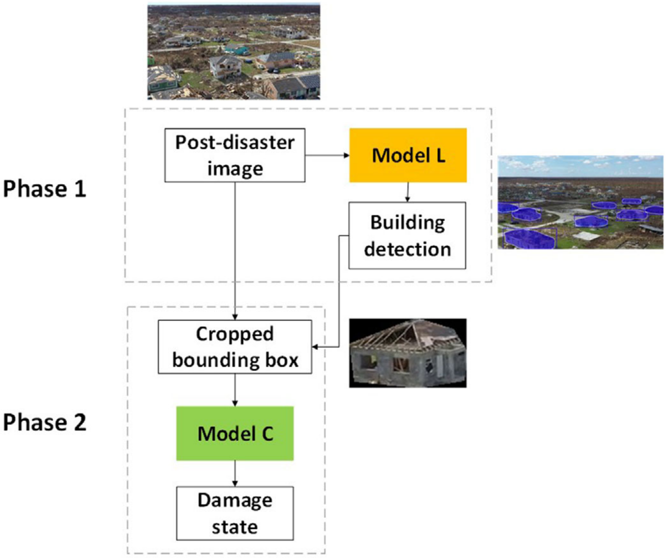
2.2 | Stacked PDA model

Unlike many visual recognition tasks that involve detecting everyday objects in natural settings, post-disaster scenes are highly complex and cluttered, and feature object classes that are difficult to differentiate (e.g., two adjacent houses with inherently different but visually similar damage levels). This could result in a relatively large number of false positive (FP) and false negative (FN) detections, ultimately deteriorating model performance. This study proposes a stacked PDA model, referred to as SPDA, to reduce the likelihood of erroneous detections. The SPDA model consists of two CNNs, namely a building localization model to distinguish between "building" and "no building" objects, and a damage level classification model to perform multi-classification task of building damage assessment. More details about these two models are presented in the following subsections. A schematic representation of the proposed SPDA model is shown in Figure 4.

2.2.1 | Building localization model: Model L

Past work by Pi et al. (2020a) has shown that damaged and undamaged building roofs can be accurately detected

FIGURE 4 Flow diagram of the developed SPDA



using YOLOv2, a deep neural network (DNN) model for real-time object detection (Redmon & Farhadi, 2017). Additionally, Pi (2020) provided detailed experiments and discussions on object detection in post-disaster settings using mask region-based CNN (Mask R-CNN). As shown in Figure 4, the first phase of the SPDA model involves a building localization model (i.e., Model L) based on the previous work by Pi (2020). This model is trained on pixel-level annotations of buildings in aerial imagery and tasked with localizing (finding the location of) building objects in each video frame regardless of damage level. Each detected building is marked with a bounding box which contains pixel-level segmentation of the building shape. The output of localization is used to automatically produce cropped images of buildings, which in turn serve as input to the damage classification model in phase two of the SPDA, as shown in Figure 4.

Model L is developed through transfer learning on a Mask R-CNN (He, Gkioxari, Dollar, & Girshick, 2017) architecture, named RetinaNet (T.-Y. Lin, Goyal, Girshick, He, & Dollár, 2017), as shown in Figure 5. RetinaNet is pre-trained on a large publicly available dataset, COCO (T.-Y. Lin et al., 2014). For better accuracy and faster processing, a feature pyramid network (FPN) (Dai, Li, He, & Sun, 2016) based on ResNet-50 (He, Zhang, Ren, & Sun, 2016) is implemented in Mask R-CNN.

2.2.2 | Damage classification model: Model C

Model C classifies cropped building images into six damage levels, FEMA's damage states 0 through 4 as well as a new damage level 5 for buildings that are completely destroyed or are under construction. To train Model C with relatively small data, the network is developed and trained through transfer learning on MobileNet (Howard et al., 2017), a fully-trained CNN for multi-classification through depth-wise separable convolutions (Figure 6). With only 28 layers, MobileNet has a light architecture, is trained on > 1,000,000 images from the ImageNet dataset (Deng et al., 2009), and can classify 1,000 object

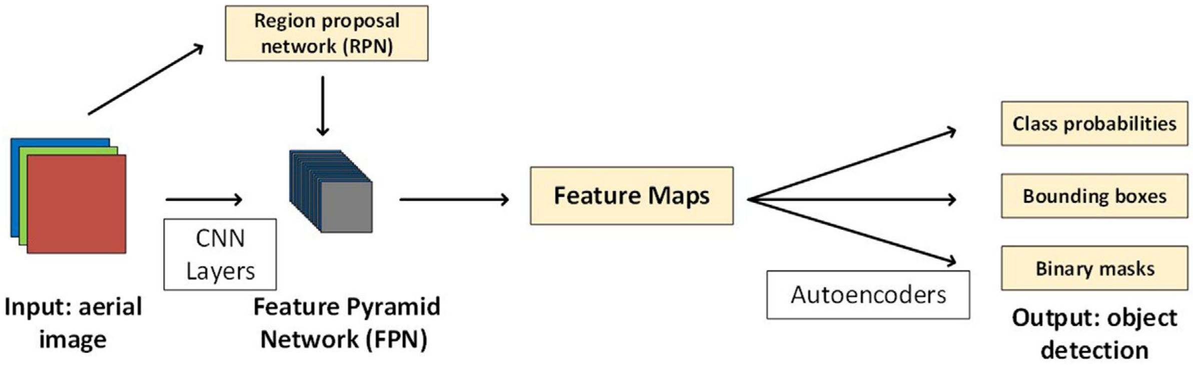


FIGURE 5 Architecture of Model L (localization) based on RetinaNet

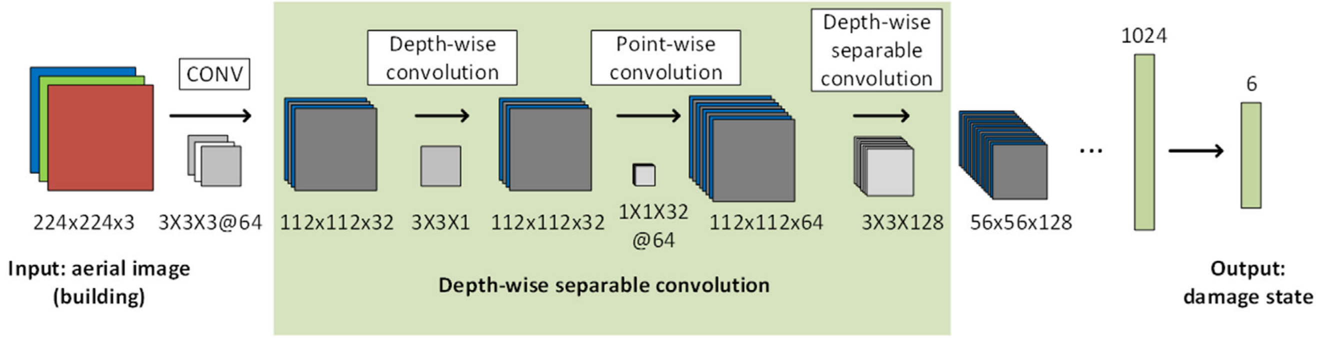


FIGURE 6 Architecture of Model C (classification) based on MobileNet

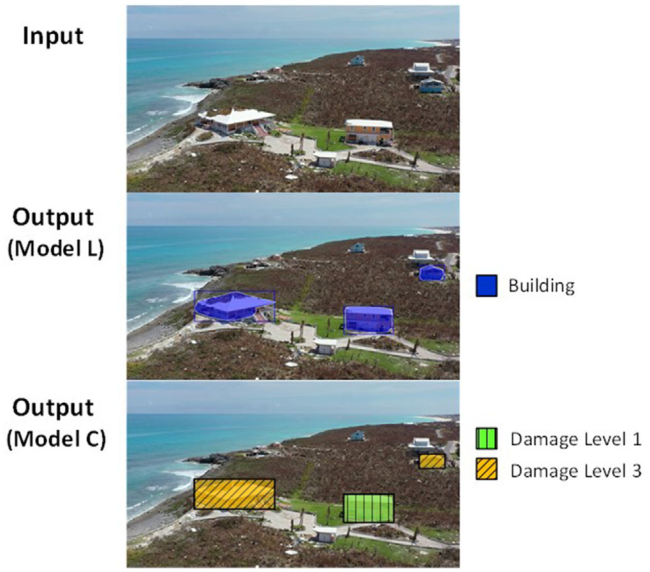


FIGURE 7 Example of building damage assessment using SPDA

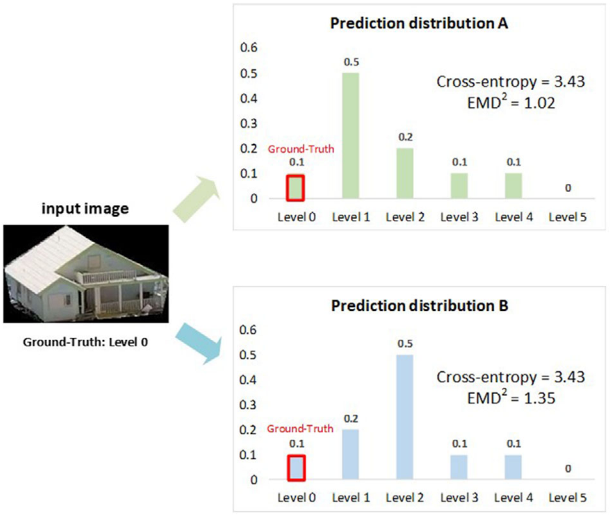


FIGURE 8 Comparison of prediction distributions (same cross-entropy, different EMD^2)

classes including person, car, boat, and animals. Since Model C is pre-trained on a very large number of feature representations from diverse images, it can be efficiently fine-tuned using the in-house dataset in this research. Figure 7 shows a sample operation performed on a single video frame by Model L and Model C successively.

2.3 | Loss function for multi-classification training

Loss function is an objective function for optimizing and tuning the weights in neural network layers (Chollet, 2017). For the multi-classification task, cross-entropy is by far the most popular loss function designed to minimize the difference between ground-truth and prediction values by tuning the weights in the network (Krizhevsky, Sutskever, & Hinton, 2017). The cross-entropy loss is defined as,

$$\text{Cross} - \text{Entropy Loss} = -\sum_{i=1}^{C} t_i \log(p_i) + (1 - t_i) \log(1 - p_i) \tag{1}$$

in which C is the total number of classes, t_i is the ith value in the ground-truth vector, and p_i is the ith value in the prediction vector.

One of the primary limitations of cross-entropy is that when the similarity of prediction classes is very high, the training effectiveness significantly decreases (Hou, Yu, & Samaras, 2016). Considering the six building damage states illustrated in Figure 1, it is seen that the target classes are highly and orderly related to each other. For example, the difference between damage levels 0 and 5 is much more significant than that of damage levels 0 and 1. To explain this, as shown in Figure 8, consider the ground-truth of damage level 0 which can be expressed as a vector [1, 0, 0, 0, 0, 0]. Let us also assume two prediction distributions, $\mathbf{A}$ = [0.1, 0.5, 0.2, 0.1, 0.1, 0] and $\mathbf{B}$ = [0.1, 0.2, 0.5, 0.1, 0.1, 0]. These two predictions differ only in the second and third elements, corresponding to damage levels 1 and 2. Using Equation (1), the cross-entropy losses of both predictions **A** and **B** are identical, despite the fact that prediction **A** is closer to the ground-truth vector. In other words, the cross-entropy loss function does not consider inter-class relationships.

To address this limitation, for training Model C, a different yet more representative loss function is adopted based on the notion of earth mover's distance (EMD) (Rubner, Tomasi, & Guibas, 2000). EMD is defined as the minimum cost needed to transform one distribution to another and can be used as a metric to evaluate the dissimilarity between two probability density functions (PDFs). Calculating the EMD between two distributions is a demanding optimization problem. However, EMD has been shown to be equivalent to Mallows distance, which is also a metric of difference between distributions and has a closed-form solution when distributions satisfy certain conditions (Levina & Bickel, 2001). According to Hou et al. (2016), these conditions are satisfied in ordered-class classification problems using softmax as the prediction layer in DNN. Furthermore, to achieve faster convergence, Hou et al. (2016) proposed using the square of EMD (also known as EMD^2) as the loss function when optimizing with gradient descent. EMD^2 can be mathematically expressed as,

$$\mathrm{EMD}^2(\mathbf{p},\mathbf{t}) = \sum_{i=1}^{C} \left(CDF_i(\mathbf{p}) - CDF_i(\mathbf{t})\right)^2 \qquad (2)$$

where index i refers to any of the classes, C is the total number of classes, **p** and **t** are the prediction and ground-truth vectors, respectively, and $CDF_i(\mathbf{x})$ denotes the ith element of cumulative distribution function (CDF) of a vector **x**. For DNN training, it has been shown that EMD^2 can be used as a loss function for multi-classification problems and is particularly suitable for classifying highly-related sequential classes, such as facial images of human age-groups (Hou et al., 2016).

In this research, EMD^2 is utilized as the loss function for training Model C with the gradient descent optimizer. The EMD^2 loss function converges smoothly (Hou et al., 2016; Shalev-Shwartz & Tewari, 2011), a behavior that is observed in this research as well. In particular, the EMD^2 loss function exhibited convergence after 13 training epochs, compared to 18 epochs when the cross-entropy loss was used. Using EMD^2 instead of the original cross-entropy loss function allows for the incorporation of class order in the multi-classification task of building damage assessment. Recalling the above example, prediction **A** should be better than prediction **B**, a conclusion that cannot be reached considering the identical values of cross-entropy loss. Using Equation (2), however, prediction **A** is found to have a smaller EMD^2 (1.02) than prediction **B** (1.35) because values in prediction **A** are closer to the ground-truth vector (damage level 0). In other words, the distance between prediction **A** and ground-truth is smaller than the distance between prediction **B** and ground-truth. This comparison is illustrated in Figure 8.

3 | EXPERIMENTS AND RESULTS

3.1 | Model development

3.1.1 | Training the localization model (Model L)

The RetinaNet neural network (Figure 5), pre-trained on COCO dataset (T.-Y. Lin et al., 2014), is retrained on 60% of Dataset 1 and validated on 20% of this dataset. The focal loss (T.-Y. Lin et al., 2017), which is based on classic cross-entropy, is utilized to handle the imbalance between object classes and image background during training, and Adam optimizer (Kingma & Ba, 2014) with a 10^{-5} learning rate and a batch size of 1 is used. After each epoch, the validation loss is monitored to prevent the model from overfitting, and training is terminated once the validation loss starts to increase. The total training time for Model L is 6.70 hours based on Intel Xeon E5-2670 v2 2.5 GHz 10-core CPU, 128GB RAM, and NVIDIA k20 GPU (Ada, 2020).

3.1.2 | Training the classification model (Model C) and data augmentation

In MobileNet, all deep layers consist of pre-trained weights that can extract useful features to efficiently classify images. However, in the newly added output layers, weights are initialized with random values and updated through retraining the model on the domain dataset for differentiating new target classes (i.e., damage levels). In this study, Model C is trained following a two-step process suggested by Nath, Behzadan, and Paal (2020). In the first step, the weights in the original layers are kept unchanged and the model is trained for 25 epochs with a learning rate of 10^{-3}. This allows the model to fit with the new classes and learn how to use pre-learned features to distinguish between damage states. In the second step, the entire model is fine-tuned by updating the weights in all layers with a lower learning rate of 10^{-4}. Using a lower learning rate allows the model to smoothly modify the pre-learned weights to find more effective features.

Moreover, real-time data augmentation is used during training by randomly transforming training images in each epoch to prevent overfitting (Kingma & Ba, 2014).

TABLE 6 Comparison of training performance using cross-entropy and EMD^2 loss

	Accuracy after first step training	Accuracy after second step training
Model trained with cross-entropy loss (%)	63	71
Model trained with EMD^2 loss (%)	77	80

Data augmentation helps expose the network to more diverse data, leading to a more generalizable output. For training model C, each image is randomly reshaped by scaling up or down by ±20%, rotating by ±15 degrees, and horizontally flipping randomly selected 50% of images. Additionally, a 50% dropout is implemented to prevent overfitting (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). The total training time for Model C is 16 minutes based on Intel Xeon E5-2670 v2 2.5 GHz 10-core CPU, 128GB RAM, and NVIDIA k20 GPU (Ada, 2020).

In order to assess the influence of loss function on classification performance, using the same training process, batch size, and learning rate, model C is trained once using the cross-entropy loss and another time with the EMD^2 loss. As Table 6 shows, during training with the cross-entropy loss, the model reaches 63% accuracy after 25 epochs, and a final accuracy of 71%. In comparison, when trained with the EMD^2 loss, the model reaches 77% accuracy after 25 epochs in step 1, and a final accuracy of 80%.

3.2 | Model testing and performance analysis

The most common metrics used to measure the performance of CNN models in object detection are intersection over union (IoU), precision, recall, and average precision (AP) (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; T.-Y. Lin et al., 2014; Rezatofighi et al., 2019). IoU measures the percentage of the overlapping region between the ground-truth and prediction (Rezatofighi et al., 2019). IoU values between 50% and 90% are commonly used for various object detection problems (Everingham et al., 2010; Pi et al., 2020a). In this study, due to the high complexity of the post-disaster scene, the detection is considered successful if IoU is greater than 50%. For any object class, the number of true positive (TP), FP, and FN predictions are needed to calculate precision, recall, and AP. Precision is defined as the number of TP cases over TP + FP, and recall is the number of TP cases over TP + FN (Powers, 2011). When plotted together, the area under the resulting precision-recall curve is termed AP (Everingham et al., 2010). A threshold value of 0.5 is used in this research to filter out low-confidence predictions by Model L, similar to a previous study by the authors (Pi et al., 2020a). Testing Model L on the remaining 20% of Dataset 1 yields a 65.6% AP for detecting building objects. Consequently, testing Model C yields an overall 61% classification accuracy. As results in Table 7 show, damage level 0 is most confused with damage level 1. This can be explained considering the high object similarity between these two classes. However, Model C shows a satisfactory performance in identifying the correct damage state ±1 class. For instance, ground-truth damage state 1 is classified within ±1 class (i.e., 0, 1, or 2) in 94% of cases. Overall, the classification accuracy of Model C considering ±1 class deviation from ground-truth is 90% which is sufficiently high for the intended PDA application. The average speeds of processing a single video frame (1,280 × 720) are 2.87 FPS for Model L and 20.12 FPS for Model C, respectively, based on NVIDIA Quadro P2000 GPU. Next, the relationship between prediction confidence and accuracy is investigated. For each prediction made by Model C, the maximum of softmax layer outputs is used as a confidence metric. Table 8 summarizes the findings based on the test portion of Dataset 1. In Figure 9, the horizontal axis represents the average confidence values in 10% increments, ranging from 20% to 100%. In each interval, the average accuracy of samples belonging to that interval is shown on the vertical axis. For example, the group containing samples with prediction confidence values between 80% and 90% has an average prediction accuracy of 66.7%. Figure 9 shows that confidence and accuracy are positively correlated with an R^2 value of 0.896 on the testing samples in Dataset 1. In other words, Model C is expected to yield a reliable prediction when the prediction confidence is high, and the confusion diminishes with distance from diagonal values in each row.

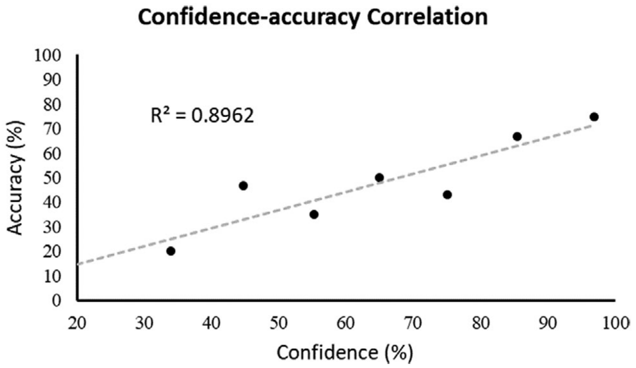


FIGURE 9 Positive correlation between confidence and accuracy of Model C (classification) tested on Dataset 1

TABLE 7 Confusion matrix for Model C (damage state classification for the test portion of Dataset 1)

		Prediction					
Damage level		**0**	**1**	**2**	**3**	**4**	**5**
Ground-truth	0	23 (68%)	10 (29%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)
	1	11 (21%)	36 (68%)	3 (6%)	3 (6%)	0 (0%)	0 (0%)
	2	1 (3%)	8 (21%)	13 (33%)	16 (41%)	0 (0%)	1 (3%)
	3	2 (2%)	12 (13%)	11 (12%)	66 (71%)	2 (2%)	0 (0%)
	4	1 (2%)	0 (0%)	4 (8%)	13 (25%)	33 (63%)	1 (2%)
	5	1 (5%)	0 (0%)	0 (0%)	2 (9%)	10 (45%)	9 (41%)

TABLE 8 Description of confidence-accuracy correlation for Model C (classification) tested on Dataset 1

	Confidence interval (%)								
	20–30	**30–40**	**40–50**	**50–60**	**60–70**	**70–80**	**80–90**	**90–100**	**Total**
No. of samples	0	5	15	20	36	37	45	135	293
Confidence mean (%)	0	34.0	44.7	55.2	65.0	75.1	85.4	96.9	81.8
Accuracy (%)	n/a	20.0	46.7	35.0	50.0	43.2	66.7	74.8	61.4

In addition, the performance of Model C considering annotation difficulty is examined using tags representing the level of annotation effort (i.e., easy, moderate, difficult) for each building object. As the results in Table 9 show, the model achieves 83%, 77%, and 75% accuracy on easy, moderate, and difficult samples. Applying one-tailed z-test with 95% confidence reveals that the prediction accuracy of easy samples (83%) is statistically higher than prediction accuracies of moderate and difficult samples (77% and 75%). It is also found that there is no statistically significant difference between the prediction accuracies of moderate and difficult samples (77% and 75%).

3.3 | Model generalizability to unseen videos

Testing Model L (localization) on unseen videos (Dataset 2) yields an AP of 63.3%, which is slightly lower than (yet sufficiently close to) the AP = 65.6% obtained from testing the same model on the test portion of Dataset 1. As shown in Figure 2, visual analysis of the videos shows that the surrounding conditions in Dataset 2 are to a large extent different than in Dataset 1, although all videos were taken from the same disaster event. For example, in Dataset 1, the imagery captures residential buildings surrounded by vegetation, whereas in the unseen videos of Dataset 2, buildings are located near the ocean backdrop.

TABLE 9 Evaluating Model C (classification) performance on different annotation difficulty levels

	Annotation effort		
	Easy	**Moderate**	**Difficult**
Confidence mean (%)	92.3	84.8	82.0
Accuracy (%)	82.9	76.7	74.8

Testing Model C (classification) on unseen videos (Dataset 2) yields an accuracy of 30%, compared to 61% obtained from testing the model on the test portion of Dataset 1. The confusion matrix is shown in Table 10, which highlights that the model is not confused merely with the ground-truth level ±1 class comparing to Table 7. For example, damage level 0 is confused with not only level 1 (15%) but also with level 3 (18%). However, the correct damage level ±1 class still accounts for a large proportion of predictions. Using the same approach as in Section 3.2, considering the ground-truth damage level ±1 class, Model C achieves an overall 64% classification accuracy.

Furthermore, for this classification task, the correlation between prediction confidence and accuracy is illustrated in Figure 10. It can be seen that even though the model is tested on completely unseen data, confidence and accuracy values have maintained their previously positive correlation, albeit with a smaller R^2 value of 0.761.

4 | DISCUSSION

Results show that SPDA performs well on post-disaster aerial UAV imagery. A key limitation of current CNN model training and testing practices, however, is that new data samples must be first annotated and used to retrain the model for better performance on unseen footage. In the disaster management domain, this may pose a challenge since disasters have unknowable visual footprints,

TABLE 10 Confusion matrix for Model C (damage state classification for unseen data in Dataset 2)

		Prediction					
Damage level		**0**	**1**	**2**	**3**	**4**	**5**
Ground-truth	0	48 (50%)	14 (15%)	9 (9%)	17 (18%)	8 (8%)	0 (0%)
	1	26 (63%)	5 (12%)	1 (2%)	6 (15%)	3 (7%)	0 (0%)
	2	11 (42%)	1 (4%)	2 (8%)	5 (19%)	7 (27%)	0 (0%)
	3	13 (32%)	12 (29%)	0 (0%)	7 (17%)	9 (22%)	0 (0%)
	4	11 (17%)	1 (2%)	3 (5%)	24 (37%)	26 (40%)	0 (0%)
	5	5 (15%)	0 (0%)	0 (0%)	3 (9%)	22 (67%)	3 (9%)

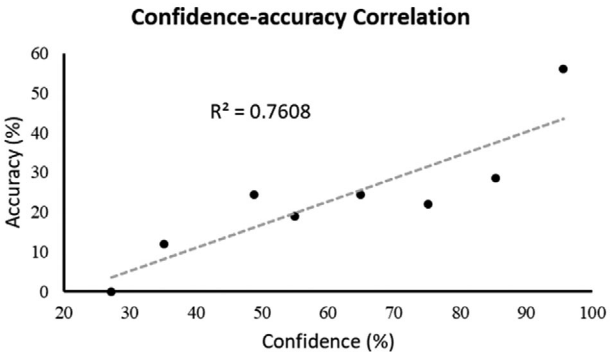


FIGURE 10 Positive correlation between confidence and accuracy of Model C (classification) tested on Dataset 2 (unseen)

causing the space of visual phenomena to be ill-defined in conventional visual recognition terms. As such, obtaining new data that adequately describes key visual patterns of future damage is not feasible. Even if such data is available at mass scales, annotating video footage is an extremely resource-intensive task. For example, in this study, it took an average of approximately 5 minutes per video frame to annotate buildings with one of the six damage states (Figure 1). In a disaster aftermath, time is of the essence, and retraining CNN models on newly collected images may not be practical for immediate response and mitigation. As an alternative, this paper demonstrated that a positive correlation exists between prediction confidence and prediction accuracy, leading us to believe that in the absence of ground-truth data, the performance of the designed SPDA model can still be sufficiently quantified.

Using the developed SPDA model, FP cases generated by Model L will be still processed by Model C for damage state classification. To avoid a large number of FP cases, one can either increase the prediction threshold of Model L to parse out only detections with high confidence, or rely on pre-disaster imagery (e.g., from Google Maps) to decide if a detected building is, in fact, a building that used to exist in that location. With regard to the former approach (adjusting the threshold value), a common value used in the literature is 0.5, and therefore, this value is used in Model L. Using this threshold, the model only outputs a few FP cases (approximately 1.5%), which do not influence the overall performance of Model C.

It is also worth noting that this study has used deep learning models in a deterministic setting. More rigorous quantification of confidence in model predictions requires a nondeterministic approach that deploys uncertainty quantification and propagation. While outside the scope of this work, such outcome can be achieved, for instance, using a Bayesian deep learning framework, which enables modeling epistemic and aleatoric uncertainty (Gal & Ghahramani, 2016; Kendall & Gal, 2017; Sajedi & Liang, 2020).

Given that class labels used in this study (i.e., building damage states) are ordinal, in addition to reporting accuracy in its traditional sense, this study also calculates the performance of Model C in terms of predicting the correct damage level ±1 class. Using this notion, for a building with a ground-truth damage state of 3, for example, short of accurately predicting this damage state, the user may still be interested in knowing if Model C can at least predict a sufficiently close label (i.e., 2 or 4, but not 1 or 5). This approach is consistent with past studies that have shown that even human experts may have different opinions when describing damage levels. Thus, an AI system that is intended not to surpass human capacity but to complement it and to gain human trust, should demonstrate the same behavior. For instance, Lue, Wilson, and Curtis (2014) asked experts to assess the tornado damage (on a 5-point scale) of 32 buildings by looking at site photos, and found that in 11 cases, the experts could not agree on one single damage label, but rather picked the top two as the most likely labels.

Moreover, the fidelity of the designed SPDA model is subject to factors such as quality and heterogeneity of input imagery as well as building characteristics and location, as described below.

4.1 | Effect of image quality on model performance

The quality of aerial imagery can influence the performance of the models. For example, buildings visible in the upper part of aerial photos are far from the camera, and thus appear smaller, have lower resolution, and are sometimes blurry. Similarly, splitting the input image for more accurate damage assessment outcomes (i.e., classifying damage levels for each floor of a multi-story building) may reduce the quality of the image, thus affecting the accuracy of damage classification. In this work, most of these low-resolution samples were marked during annotation as instances with high annotation difficulty. The experiment results in Section 3 indicate that the classification performance (Model C) is low on these samples. One potential remedy to this problem could be adjusting camera zoom and angle (occasionally during UAV flight) or post-processing photos to enhance image quality. A detailed description of this problem and its impact on the performance of visual recognition models can be found in Nath and Behzadan (2020).

Aerial damage assessment could also be difficult if the object of interest (in this case, building) is not fully visible in the camera view. For instance, if a building (or its components such as roof or walls) is partially blocked by vegetation or other artifacts, the model can make predictions based on what is visible, which may or may not represent the entire extent of the damage. When the same building is viewed from a different angle or height, the predicted damage state may be revised by the model. This behavior is expected and, in fact, consistent with how human experts label buildings for damage; they walk around the building, and step closer or farther away to see all building elements in detail. Ultimately, for each building, the final estimated damage label can be determined by selecting the maximum (most conservative) or average value of all labels generated by the model for that building when viewed in different images from the scene. However, it must be noted that by selecting the maximum value, the final damage scale outcome may be skewed by false predictions. Similarly, selecting the average value has a risk of misprediction if too many unreliable predictions occur. As a potential remedy, one may choose to only rely on high-confidence predictions. Using this approach, predictions with confidence values less than a minimum threshold (e.g., 95%) are filtered out and not considered when damage labels corresponding to any single building are aggregated. The choice of the aggregation method, to some extent, reflects the level of risk one is willing to take to arrive at the final damage score.

4.2 | Effect of building attributes and location on model performance

Previous research has investigated the pattern and intensity of damage for different building types, wind intensity, and hurricane path and duration (Chock, 2005; Crandell, 1998; De Silva, Kruse, & Wang, 2008; Egnew, Roueche, & Prevatt, 2018; Fronstin & Holtmann, 1994; Jain, Guin, & He, 2009; Lindell & Prater, 2003; Xian, Lin, & Hatzikyriakou, 2015). In this study, the correlation between two key building attributes, size and number of stories, and the severity of induced damage is further investigated. Egnew et al. (2018) reviewed the literature and summarized the relation between building attributes and the wind-induced damage or loss to the building. They considered building age, number of stories, area, value, value per area, and location, as listed in Table 11. De Silva et al. (2008) pointed out that larger buildings are at risk of more severe damage or loss. Likewise, the severity of damage to residential buildings in the aftermath of the 2011 Joplin, MO tornado was found to be much higher for buildings with a lower cost per unit of living area (Egnew et al., 2018). Although detailed information on damaged houses in the Bahamas after the 2019 Hurricane Dorian is not available, the authors manually divided a small portion (200 samples) of randomly selected building images in Dataset 1 into two groups of large and small based on their visual characteristics. A one-tailed z-test with 95% confidence reveals that the average damage level in large buildings (2.82) is higher than the average damage level in small buildings (2.44), which is in agreement with the findings in De Silva et al. (2008). The same 200 samples were also used to study the correlation between the number of stories and the damage level. Using a two-tailed z-test with 95% confidence, results are in general agreement with Egnew et al. (2018) indicating that there is no significant difference between the average damage level of one-story buildings (2.61) and buildings with more than two stories (2.67). Detailed results are included in Table 11.

To evaluate the performance of models on different disaster events and landscapes, a smaller tertiary dataset is created with 10 frames (35 building instances) extracted from a video taken in Texas after the 2017 Hurricane Harvey (Pi et al., 2020a). Testing Model L (localization) without retraining on these new frames yields an AP of 44% for detecting buildings. Subsequently, testing Model C (classification) without retraining yields an accuracy of 29% for damage state identification, and 55% considering ±1 class deviation from ground-truth. While results show that testing previously trained models on disaster footage from new locations may lead to a drop in performance (an expected

TABLE 11 Correlation of typical building attributes with wind-induced damage or loss

	Building attribute					
	Age (years)	Number of stories	Area (m^2)	Value ($)	Value per area ($/ m^2)	Dist. to shore (m)
Fronstin and Holtmann (1994)	(–)	n/a	n/a	(–)	n/a	n/a
Crandell (1998)	n/a	(+)	n/a	n/a	n/a	n/a
Chock (2005)	(+)	(–)	n/a	n/a	n/a	n/a
De Silva et al. (2008)	(–)	(–)	(+)	n/a	n/a	n/a
Jain et al. (2009)	(+)	n/a	n/a	n/a	n/a	n/a
Xian et al. (2015)	(+)	n/a	n/a	n/a	n/a	(–)
Egnew et al. (2018)	(–)	(+) (weak)	(+)	(–)	(–)	(+)
This study	n/a	(0)	(+)	n/a	n/a	n/a

Note: (+): positive correlation, (-): negative correlation, and (0): no correlation.

outcome), including more heterogeneous training data is expected to increase the accuracy of building localization and damage classification tasks on new (unseen) samples.

4.3 | Potential directions for future research

In this study, only one trained human annotator was recruited to assign ground-truth damage states to buildings. The output of the annotation process, therefore, was a binary damage vector with a single value of 1 (for the vector element corresponding to the perceived damage state) and five 0 values. For example, $\mathbf{D} = [0, 0, 0, 1, 0, 0]$ implies a damage state of 3. A possible direction for future research is to scale up the data annotation effort through crowdsourcing to obtain a larger pool of labels. Crowdsourcing is becoming a popular data collection technique for post-disaster response and mitigation (Akter & Wamba, 2019). Special attention, however, must be paid to the reliability of annotators (Hsueh, Melville, & Sindhwani, 2009) which is also referred to as task approval rating (Hauser & Schwarz, 2016). Even when working with reliable annotators, multiple annotators may label the same building differently, especially when the building condition does not warrant a clear damage state (i.e., perceived damage appears to be in between two damage states). Therefore, ground-truth labels obtained via crowdsourcing may be non-deterministic; that is the ground-truth vectors are nonbinary as they may contain partially distributed values over several candidate classes. As an example, $\mathbf{D}' = [0.1, 0.3, 0.5, 0.1, 0, 0]$ implies a 50% likelihood that the ground-truth damage state is 2, and a 30% likelihood that the ground-truth damage state is 1. Previous research shows that a high disagreement among nonprofessional annotators deteriorates model performance, but obtaining extra (i.e., redundant) annotations can improve data quality (Novotney & Callison-Burch, 2010). Therefore, if crowdsourced annotators are well-trained, relying on probabilistic ground-truth data can potentially lead to better-performing CNN models for predicting building damage states with the same EMD^2 loss function.

Designing targeted data balancing strategies is another possible direction for future research. As results indicate, trained CNNs perform with higher prediction confidence on images with less annotation difficulty. As Table 4 shows, the number of samples with easy, moderate, and difficult tags is fairly equally distributed in the datasets used in this research. However, it is not hard to imagine that new data from a future disaster event may not be balanced with respect to annotation difficulty. Past research (Dupret & Koda, 2001; López, Fernández, García, Palade, & Herrera, 2013; Pi et al., 2020a) shows that training models on unbalanced data can lead to low performance and prediction bias. Therefore, further investigation is needed to understand the impact of data imbalance on model performance, and whether a targeted strategy (e.g., up-sampling, down-sampling, or a combination of both) could resolve the data imbalance problem with minimum computational cost.

5 | CONCLUSION

The objective of the research was to create and benchmark AI-assisted visual recognition models for post-disaster PDA using UAV imagery. To train AI models, an in-house dataset was created using web mining, and buildings visible in video frames were manually labeled for damage states, as well as assigned an additional tag to describe annotation effort. The dataset contained videos of post-disaster scenes in two locations in the Bahamas, Marsh Harbor and Great Guana Cay. Annotated video frames from the first location were used to train and test the mode, while videos from the second location were kept unseen from the trained model, and utilized only for verifying whether a model trained on one geographical location can produce satisfactory predictions when applied to footage from a new location without retraining.

A stacked PDA (SPDA) model based on two CNN architectures was developed to first localize building objects (Model L in phase 1) and then, use the localization output to classify damage state (Model C in phase 2). Model L was developed through transfer learning on a Mask R-CNN backbone, yielding a 65.6% AP, and Model C was constructed using transfer learning on MobileNet. To train Model C, the cross-entropy loss function was replaced with EMD^2 to enable handling the ordinality of classes corresponding to damage states. This improvement led to an overall 61% classification accuracy by Model C. To assess the generalizability of the trained models to new locations, they were also tested on unseen video frames from the second location (i.e., Great Guana Cay). Model L yielded an AP of 63.3%, and Model C produced a 30% accuracy. Moreover, this study explored the relationship between prediction confidence and accuracy, with results indicating a positive correlation between the two. Finally, in a randomized subset of the dataset, a generally positive correlation was found between building size and damage level, and no significant correlation was observed between the number of stories and damage severity.

Although the designed SPDA model can successfully detect buildings and classify induced damage with good accuracy, its performance in detecting building damage as a result of other types of hazards (e.g., earthquake, landslide) still needs to be evaluated. Additionally, while this study utilized only one trained human annotator, the authors plan to recruit crowdworkers in the future in order to scale up data annotation and obtain large-sized and more heterogeneous labels. Moreover, devising data balancing techniques that target specific characteristics of input data (e.g., high imbalance with respect to annotation difficulty) is another possible direction of future research.

ACKNOWLEDGMENTS

The authors would like to thank Texas A&M University's High Performance Research Computing (HPRC) for providing necessary computing infrastructure for model training. Any opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily represent the views of the HPRC.

REFERENCES

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M., & Inman, D. J. (2017). Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. *Journal of Sound and Vibration*, *388*, 154–170.

Ada (2020). Texas A&M University high performance research computing. Ada: An Intel x86-64 Linux cluster. Retrieved from https://hprc.tamu.edu/wiki/Ada:INtro

Adams, S., Friedland, C., & Levitan, M. (2010). *Unmanned aerial vehicle data acquisition for damage assessment in hurricane events*. Paper presented at the Proceedings of the 8th International Workshop on Remote Sensing for Disaster Management, Tokyo, Japan.

Akter, S., & Wamba, S. F. (2019). Big Data and disaster management: A systematic review and agenda for future research. *Annals of Operations Research*, *283*(1), 939–959.

Altay, N., & Green, W. G. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research*, *175*(1), 475–493.

Chock, G. Y. K. (2005). Modeling of hurricane damage for Hawaii residential construction. *Journal of Wind Engineering and Industrial Aerodynamics*, *93*(8), 603–622.

Chollet, F. (2017). *Deep learning with Python*. Shelter Island, NY: Manning Publications Company.

Crandell, J. H. (1998). Statistical assessment of construction characteristics and performance of homes in hurricanes Andrew and Opal. *Journal of Wind Engineering and Industrial Aerodynamics*, *77-78*, 695–701.

Dai, J., Li, Y., He, K., & Sun, J. (2016). *R-FCN: Object detection via region-based fully convolutional networks*. Paper presented at 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., & Li, F.-F. (2009). *ImageNet: A large-scale hierarchical image database*. Paper presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL.

De Silva, D. G., Kruse, J. B., & Wang, Y. (2008). Spatial dependencies in wind-related housing damage. *Natural Hazards*, *47*(3), 317–330.

Dupret, G., & Koda, M. (2001). Bootstrap re-sampling for unbalanced data in supervised learning. *European Journal of Operational Research*, *134*(1), 141–156.

Egnew, A. C., Roueche, D. B., & Prevatt, D. O. (2018). Linking building attributes and tornado vulnerability using a logistic regression model. *Natural Hazards Review*, *19*(4), 04018017.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

FEMA. (2003). Multi-hazard loss estimation methodology: Hurricane model HAZUS-MH MR3, Technical manual. Retrieved from https://www.hsdl.org/

FEMA. (2020). FEMA preliminary damage assessment guide. Retrieved from https://www.fema.gov/

Friedland, C. J., Adams, B. J., & Levitan, M. L. (2007). *Remote sensing classification of hurricane storm surge structural damage*. Paper presented at the Forensic Engineering Conference at Structures Congress, May 16–19, June 15–20, Long Beach, CA.

Fronstin, P., & Holtmann, A. G. (1994). The determinants of residential property damage caused by Hurricane Andrew. *Southern Economic Journal*, *61*(2), 387–397.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning, NY*.

Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering*, *33*(9), 748–768.

Ghosh Mondal, T., Jahanshahi, M. R., Wu, R.-T., & Wu, Z. Y. (2020). Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance. *Structural Control and Health Monitoring*, *27*(4), e2507.

Grünthal, G. (1998). European macroseismic scale 1998. Retrieved from http://lib.riskreductionafrica.org/

Gunning, D. (2017). Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency, DARPA/I20.

Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., …, Gaston, M. (2019). xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the 16th International Conference on Computer Vision*, Venice, Italy (pp. 2961–2969).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 27—30, Las Vegas, NV.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hou, L., Yu, C.-P., & Samaras, D. (2016). Squared earth mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T. ,…, Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: A study of annotation selection criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, Boulder, CO.

Iizuka, K., Itoh, M., Shiodera, S., Matsubara, T., Dohar, M., & Watanabe, K. (2018). Advantages of unmanned aerial vehicle (UAV) photogrammetry for landscape analysis compared with satellite data: A case study of postmining sites in Indonesia. *Cogent Geoscience*, *4*(1), 1498180.

Jain, V. K., Guin, J., & He, H. (2009). Statistical analysis of 2004 and 2005 hurricane claims data. *Proceedings of the 11th Americas Conference on Wind Engineering*, San Juan, Puerto Rico, June 22–26.

Kang, D., & Cha, Y.-J. (2018). Autonomous UAVs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging. *Computer-Aided Civil and Infrastructure Engineering*, *33*(10), 885–902.

Kelman, I. (2003). *Physical flood vulnerability of residential properties in coastal, Eastern England* (Ph.D. dissertation). University of Cambridge, Cambridge, United Kingdom.

Kendall, A., & Gal, Y. (2017). *What uncertainties do we need in Bayesian deep learning for computer Vision?* Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Labelbox. (2019). Labelbox: The leading training data solution. Retrieved from https://labelbox.com

Lenjani, A., Yeum, C. M., Dyke, S., & Bilionis, I. (2020). Automated building image extraction from 360° panoramas for postdisaster evaluation. *Computer-Aided Civil and Infrastructure Engineering*, *35*(3), 241–257.

Levina, E., & Bickel, P. (2001). The earth mover's distance is the mallows distance: some insights from statistics. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, July 7—14, Vancouver, BC, Canada.

Liang, X. (2019). Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization. *Computer-Aided Civil Infrastructure Engineering*, *34*(5), 415–430.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. *Proceedings of the 16th International Conference on Computer Vision*, Venice, Italy.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele & T. Tuytelaars (Eds.) *Computer Vision–ECCV 2014. Lecture Notes in Computer Science*. Cham, Switzerland: Springer.

Lin, Y.-Z., Nie, Z.-H., & Ma, H.-W. (2017). Structural damage detection with automatic feature-extraction through deep learning. *Computer-Aided Civil and Infrastructure Engineering*, *32*(12), 1025–1046.

Lindell, M. K., & Prater, C. S. (2003). Assessing community impacts of natural disasters. *Natural Hazards Review*, *4*(4), 176–185.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Lue, E., Wilson, J. P., & Curtis, A. (2014). Conducting disaster damage assessments with spatial video, experts, and citizens. *Applied Geography*, *52*, 46–54.

Marshall, J., Smith, D., Lyda, A., Roueche, D., Davis, B., Djima, W., …, Mosalam, K. (2019). StEER - Hurricane Dorian: Field Assessment Structural Team (FAST-1) Early Access Reconnaissance Report (EARR). *DesignSafe-CI*. https://doi.org/10.17603/ds2-4616-1e25

Nath, N. D., & Behzadan, A. H. (2020). Deep convolutional networks for construction object detection under different visual conditions. *Frontiers in Built Environment*, *6*, 97.

Nath, N. D., Behzadan, A. H., & Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, *112*, 103085.

Nex, F., Duarte, D., Tonolo, F. G., & Kerle, N. (2019). Structural building damage detection with deep learning: Assessment of a state-of-the-art CNN in operational conditions. *Remote Sensing*, *11*(23), 2765.

Novotney, S., & Callison-Burch, C. (2010). *Cheap, fast and good enough: Automatic speech recognition with non-expert transcription*. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA.

Pi, Y. (2020). *Aerial disaster information retrieval framework using artificial intelligence* (Ph.D. dissertation). Texas A&M University, College Station, TX.

Pi, Y., Nath, N. D., & Behzadan, A. H. (2020a). Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics*, *43*, 101009.

Pi, Y., Nath, N. D., & Behzadan, A. H. (2020b). Disaster impact information retrieval using deep learning object detection in crowdsourced drone footage. *Proceedings of the 27th International Workshop on Intelligent Computing in Engineering (EG-ICE)*, Berlin, Germany.

Pinelli, J.-P., Roueche, D., Kijewski-Correa, T., Plaz, F., Prevatt, D., Zisis, I., . . . Moravej, M. (2018). *Overview of damage observed in regional construction during the passage of Hurricane Irma over the state of Florida*. Paper presented at Eighth Congress on Forensic Engineering November 29–December 2, 2018, Austin, TX.

Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63

Rafiei, M. H., & Adeli, H. (2017a). NEEWS: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. *Soil Dynamics and Earthquake Engineering*, *100*, 417–427.

Rafiei, M. H., & Adeli, H. (2017b). A novel machine learning-based algorithm to detect damage in high-rise building structures. *Structural Design of Tall and Special Buildings*, *26*(18), e1400.

Rafiei, M. H., & Adeli, H. (2018a). Novel machine-learning model for estimating construction costs considering economic variables and indexes. *Journal of Construction Engineering and Management*, *144*(12), 04018106.

Rafiei, M. H., & Adeli, H. (2018b). A novel unsupervised deep learning model for global and local health condition assessment of structures. *Engineering Structures*, *156*, 598–607.

Rafiei, M. H., Khushefati, W. H., Demirboga, R., & Adeli, H. (2017). Supervised deep restricted Boltzmann machine for estimation of concrete. *ACI Materials Journal*, *114*(2), 237.

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 21–26, Honolulu, HI.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., & Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 15–20, Long Beach, CA.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Sajedi, S. O., & Liang, X. (2020). Uncertainty-assisted deep vision structural health monitoring. *Computer-Aided Civil Infrastructure Engineering*. https://doi.org/10.1111/mice.12580

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Shalev-Shwartz, S., & Tewari, A. (2011). Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, *12*(Jun), 1865–1892.

Thomas, J., Kareem, A., & Bowyer, K. W. (2014). Automated post-storm damage classification of low-rise building roofing systems using high-resolution aerial imagery. *IEEE Transactions on Geoscience and Remote Sensing*, *52*(7), 3851–3861.

Wang, N., Zhao, Q., Li, S., Zhao, X., & Zhao, P. (2018). Damage classification for masonry historic structures using convolutional neural networks based on still images. *Computer-Aided Civil and Infrastructure Engineering*, *33*(12), 1073–1089.

Xian, S., Lin, N., & Hatzikyriakou, A. (2015). Storm surge damage to residential areas: A quantitative analysis for Hurricane Sandy in comparison with FEMA flood map. *Natural Hazards*, *79*(3), 1867–1888.

Xu, J., Gui, C., & Han, Q. (2020). Recognition of rust grade and rust ratio of steel structures based on ensembled convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, *35*(10), 1160–1174.

Xu, J. Z., Lu, W., Li, Z., Khaitan, P., & Zaytseva, V. (2019). Building damage detection in satellite imagery using convolutional neural networks. *arXiv preprint arXiv:.06444*.

Yu, M., Yang, C., & Li, Y. (2018). Big data in natural disaster management: A review. *Geosciences*, *8*(5), 165.

Zhang, A., Wang, K. C. P., Li, B., Yang, E., Dai, X., Peng, Y., … Chen, C. (2017). Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. *Computer-Aided Civil and Infrastructure Engineering*, *32*(10), 805–819.

Zhang, K., Cheng, H. D., & Zhang, B. (2018). Unified approach to pavement crack and sealed crack detection using preclassification based on transfer learning. *Journal of Computing in Civil Engineering*, *32*(2), 04018001.

How to cite this article: Cheng C-S, Behzadan AH, Noshadravan A. Deep learning for post-hurricane aerial damage assessment of buildings. *Comput Aided Civ Inf*. 2021;*36*:695–710. https://doi.org/10.1111/mice.12658