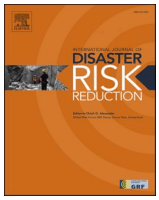
International Journal of Disaster Risk Reduction 65 (2021) 102540

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

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



DamageMap: A post-wildfire damaged buildings classifier

Marios Galanis [a], Krishna Rao [b], Xinle Yao [a], Yi-Lin Tsai [a], Jonathan Ventura [c], G. Andrew Fricker [d,*]

[a] Stanford University, Department of Civil & Environmental Engineering, USA
[b] Stanford University, Department of Earth System Science, USA
[c] California Polytechnic State University, Department of Computer Science and Software Engineering, USA
[d] California Polytechnic State University, Department of Social Sciences, USA

ARTICLE INFO

Keywords:
Deep learning
Building damage classification
Disaster management
Wildfires

ABSTRACT

The increasing frequency and severity of wildfire events in the last few decades has created an urgent need for new technologies that allow rapid surveying and assessment of post-wildfire building damage. However, existing technologies lack in accuracy and ability to scale to effectively aid disaster relief and recovery. Even today, most wildfire event inspectors need to physically visit the areas impacted by wildfires and manually classify building damage, which requires considerable time and resources. Here, we present DamageMap, an artificial intelligence-powered post-wildfire building damage classifier. DamageMap is a binary classifier (outputs are "damaged" or "undamaged"). Unlike existing solutions that require both pre- and post-wildfire imagery to classify building damage, DamageMap relies on post-wildfire images alone by separating the segmentation and classification tasks. Our model has an overall accuracy of 98% on the validation set (five wildfire events all around the world) and 92% and 98% on two independent test sets from the Camp Fire and the Carr Fire, respectively. Excellent model performance across a variety of datasets provides evidence of DamageMap's robustness to unseen data. Thus, DamageMap may help governmental and non-governmental agencies rapidly survey building damage using post-wildfire aerial or satellite imagery in wildfire-impacted areas.

1. Introduction

Post-wildfire disaster recovery depends on the ability to rapidly assess building damage. Detailed information about the locations, sizes, and values of damaged buildings is essential for multiple affected stakeholders to proceed with the disaster recovery process. First responders and disaster management agencies rely on damage information to better strategize their efforts in response and recovery. Affected residents require information about their damaged buildings to initiate insurance claims and to request financial support and disaster relief from governmental agencies. State and federal authorities depend on the building damage information and cost estimations to plan for future disaster relief resources and financial aid [1]. A successful disaster recovery for fire-affected communities thus depends on the ability to quickly assess building damage.

The need for rapid building damage classification has intensified as climate change and the growth of the wildland-urban interface (WUI), defined as the areas where human development intersects or intermingles with undeveloped wildland or vegetative fuels, have substantially increased wildfire risk, extent, and destruction globally [3–5]. Over the past five decades, the frequency of wildfire events and areas burned in the western United States have increased by more than four and six times, respectively [6,7]. Globally, wildfire seasons have lengthened significantly due to global warming in the past three decades [4]. Recent studies also indicate that wildfire intensity and areas burned will continue to increase in Australia, Canada, Chile, Portugal, and the United States [8–12]. In addition to climate change, the continuous expansion of the WUI exacerbates the ignition probability and asset exposure to wildfire events, thus leading to escalating wildfire risks [3, 13–15]. Studies have shown that human activities and ignitions inside the WUI are one of the major causes of wildfires in recent years [3,5,16, 17].

Despite the importance and urgent need for rapid building damage classification, current methods of inspecting wildfire-induced building damage are tedious and time-consuming. In the United States, damage assessment teams, usually consisting of certified specialists, are

* Corresponding author.
E-mail address: africker@calpoly.edu (G.A. Fricker).

https://doi.org/10.1016/j.ijdrr.2021.102540
Received 16 August 2021; Accepted 26 August 2021
Available online 28 August 2021

deployed by local governments to perform building damage inspections and provide safety evaluation of structures in burned areas [1,18]. Building damage assessment teams also compile reports and send them to emergency operation centers (EOC) after the completion of building damage investigation to allow the EOC staff to strategize the recovery effort. Although the manual process provides valuable damage information, the addition of an automated quick assessment and data compilation could provide a more complete picture of the damage in a more timely manner. In addition to immediate post-wildfire recovery, the automation of damage data collection and compilation could also benefit the long-term study of wildfire damage and building loss. However, as of now, there is no nation-wide consistent building loss dataset. Many past studies have had to manually compile and digitize local data to perform analyses [5].

To automate the identification and classification of damaged buildings, many studies have combined aerial or satellite imagery with a variety of methods ranging from simple spectral analyses [19–22] to complex methods involving deep learning models with more than a million parameters [23–25]. Thanks to the advances in deep learning models and availability of vast amounts of training data [26], deep learning models have surpassed traditional spectral methods in accuracy of classifying building damage.

However, most methods of automating building damage identification and classification still require images from both pre-wildfire and post-wildfire in similar settings of lighting and camera angle [see 27,28, and references within]. These building damage classification algorithms have primarily focused on change detection between pre- and post-disaster images [27,28]. Depending on the magnitude of spectral change between the two images in similar lighting conditions, either a parametric model [29] or a non-parametric machine learning model [30] is trained to identify the presence and extent of building damage. Under such constraints, maintaining a reliable pre-wildfire image database requires frequent surveys. This poses a major challenge in scaling such techniques globally since maintaining a reliable pre-wildfire image database for the entire world would be very costly. In contrast, post-wildfire imagery is easier to obtain than pre-wildfire imagery because of the availability of on-demand satellites, such as WorldView [31] and the possibility to conduct one-off aerial surveys. Therefore, identifying buildings in any pre-wildfire image (taken before a wildfire, not necessarily at similar lighting or camera angle) and then classifying building damage using post-wildfire images only has clear advantages. Although a few studies have attempted to identify damaged buildings using post-disaster images only, they have been restricted to non-deep learning solutions, thus vastly limiting their accuracy [32–34].

In this study, we combined two deep learning models to first identify building footprints in any recent pre-wildfire image and then classify building damage using post-wildfire images only. The two models work sequentially. The first model uses aerial or satellite imagery to output regions of interest (building footprints) which are then used by the second model to classify the footprints as either “damaged”, or “undamaged” buildings. We refer to the two models in combination as “DamageMap”. As the remainder of this paper details, we trained, validated, and tested DamageMap on wildfire events around the world to assess its robustness for several roof architectures and landscape characteristics.

2. Methods

Even though models exist to perform two-class segmentation for buildings, state-of-the-art models still require images sourced right before and after a disaster event to segment and classify the buildings [23,30]. Such classifiers work through change detection (e.g., if a building’s pixel values change significantly, it is likely damaged). Since predicting the timing of natural hazards like wildfires is almost impossible, obtaining aerial images just before a wildfire can be challenging.

To overcome the challenge of classifying building damage without a pre-wildfire image, we used two separate models:

1. A segmenter that uses *any* pre-wildfire image (not necessarily sourced just prior to the wildfires, but at any time in recent history) to produce regions of interest for the classifier.
2. A classifier that uses *only* a post-wildfire image to classify building damage.

2.1. Dataset

For training and validating our classifier, we used a subset of the xBD dataset [26], the largest building damage assessment dataset to date which contains pre- and post-event satellite images of various natural disaster events. In this study, we only used the post-wildfire images. The locations of the five wildfires included in xBD are shown as blue dots in Fig. 1. The xBD contained 47,543 cropped images of buildings that suffered different levels of damage (“no damage,” “minor damage,” “major damage,” and “destroyed”) after wildfire events. All images were red-green-blue (RGB), and their resolution was 0.8 m. We used these images as our training set because it offered two advantages. First, the five wildfire events included in xBD had occurred in disparate locations with different building architectures. This diversity of locations could allow our model to recognize different types and styles of structures. Second, the images after each wildfire were taken separately under different lighting conditions. This variation in light conditions could help our model perform consistently on new wildfire images taken during different times of the day.

To test the performance of our model on real case studies, we used two GeoTiff RGB images of Paradise, California after the Camp Fire of 2018 (Camp Fire) and Whiskeytown–Shasta–Trinity National Recreation Area after the Carr Fire of 2018 (Carr Fire), both of which are shown in Fig. 1 with yellow dots. The two images were captured using aerial imagery and were made available by the California Department of Forestry and Fire Protection [35]. The level of damage for each building in the images was classified by damage assessment inspectors who visited each structure. The manual on-site classification means that buildings with intact rooftops, which will appear undamaged from the aerial images, can be labeled as damaged due to broken walls or windows. To overcome this challenge, we re-classified the damage to just two classes, undamaged and damaged, as explained further in Sec. 2.3.

The Camp Fire image had dimensions of 30,000 x 30,000 x 3 pixels (height x width x channels), and 0.3 m ground resolution. The Carr Fire image had dimensions of 68,000 x 80,000 x 3, and 0.15 m ground resolution. We decided to test our model on these two fires because both images were taken by different types of cameras than xBD (i.e., aerial imagery instead of satellite), and at different spatial resolution. This allowed us to test whether the model will be capable of handling new data collected from local agencies after a wildfire event. The total number of damaged and undamaged buildings per wildfire is displayed in Fig. 2.

2.2. Segmenter and classifier choice

The segmenter consists of a pre-trained ResNet34 model from Microsoft [36]. The ResNet34 (residual network) is a type of neural network that can be efficiently trained despite containing many (34) layers due to connections that allow skipping of one or more layers. The segmenter performs building extraction from satellite images in two steps. First, it recognizes building pixels in the aerial images (semantic segmentation), and then it converts the building pixels into polygons (representing individual structures). This allows us to extract the footprints of buildings. The advantage of this segmenter is that it can be used in any aerial image taken anytime recently before a wildfire event. The extracted footprints are used as post-wildfire regions of interest to classify the damage of each building. Moreover, the footprints of all the

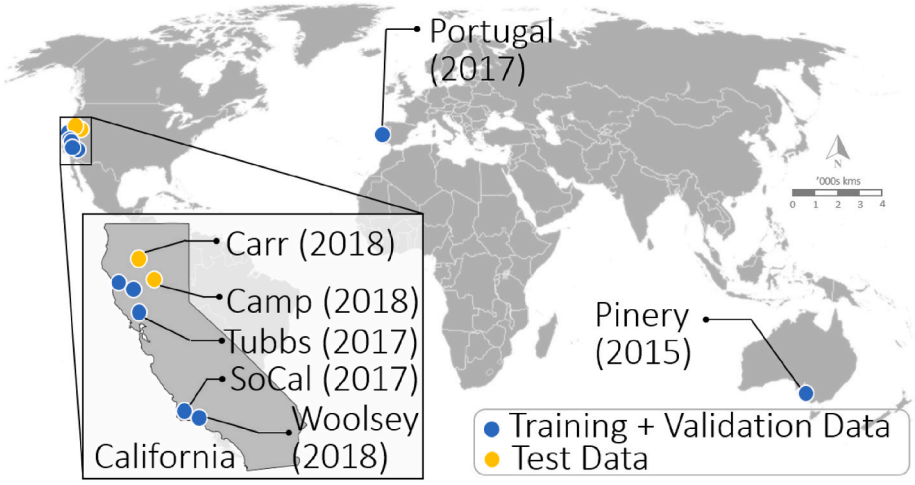


Fig. 1. Locations of all wildfire images used for training, validation, and testing. Training and validation data originated from the same source (xBD) whereas test images were sourced from the Carr Fire and the Camp Fire areas independently. Year of each wildfire is shown in parenthesis.

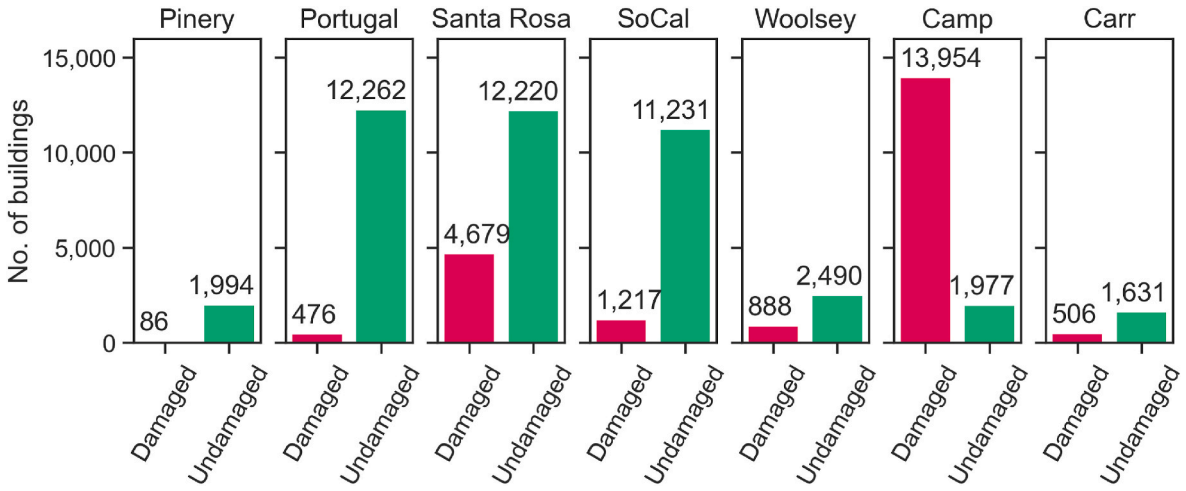


Fig. 2. Number of damaged and undamaged structures in each wildfire incident in our dataset. With the exception of the Camp Fire, all locations had a majority of undamaged structures. For the Pinery, Portugal, Santa Rosa, Socal, and Woolsey wildfires, the damage classification was performed by human labelers using satellite images, whereas in the case of the Camp and Carr fires (test set), assessment was performed by post-disaster damage assessors through ground surveys.

buildings in the USA were already created in 2018 by Microsoft using this segmenter. For the classifier, we used a convolutional neural network (CNN) since it is the most widely used neural network architecture for analyzing visual imagery, as first proposed by LeCun et al. [37]. We selected five of the most popular CNN architectures (ResNet [38], AlexNet [39], VGG [40], SqueezeNet [41], DenseNet [42]) pre-trained on Imagenet [43] and fine-tuned them on our training set. Since the classification was only binary, we used the smallest models of each architecture. After comparing their accuracies on our validation set, we decided to keep ResNet18 as our classifier of choice due to its better performance in our application.

2.3. Training

Before training, the dataset was pre-processed. The original xBD dataset contained a total of 47,543 images of structures after wildfires. Those images were separated in four classes depending on the severity of damage the structure experienced during the wildfire incident ("no damage," "minor damage," "major damage," and "destroyed"). Since it was virtually impossible to distinguish an undamaged building from a building with "minor damage" using aerial imagery alone (see Fig. 3 in Ref. [30] and F1-scores in Ref. [26]), we reduced the number of classes to two. We labeled the first class containing the xBD classes of "no damage" and "minor damage" as "undamaged". We labeled the other class containing "major damage", and "destroyed" as "damaged". Since DamageMap is designed to serve as a rapid assessment tool for post-wildfire building damage assessment, having a more accurate model with less granular information about the extent of damage is better than having a less accurate model with more granular information.

As shown in Fig. 2 the two new classes were imbalanced with only 7,346 images depicting damaged structures, while the remaining 40,197 images contained undamaged structures. Such imbalance is common in wildfire events when most buildings in an affected area are safe. To separate the images into training, validation, and test sets, we did the following. We randomly picked 30,182 images of undamaged structures and 4,915 of damaged structures for the training set, 4,919 images of undamaged structures and 1,039 of damaged structures for the validation set, and 5,096 images of undamaged structures and 1,392 of damaged structures for the test set. To deal with the class imbalance, we augmented the damaged structure images in the training and validation set by performing three rotations (90°, 180°, and 270°), flip, and mirror transformations. These transformations were not applied to the test set– the set used to evaluate model performance.

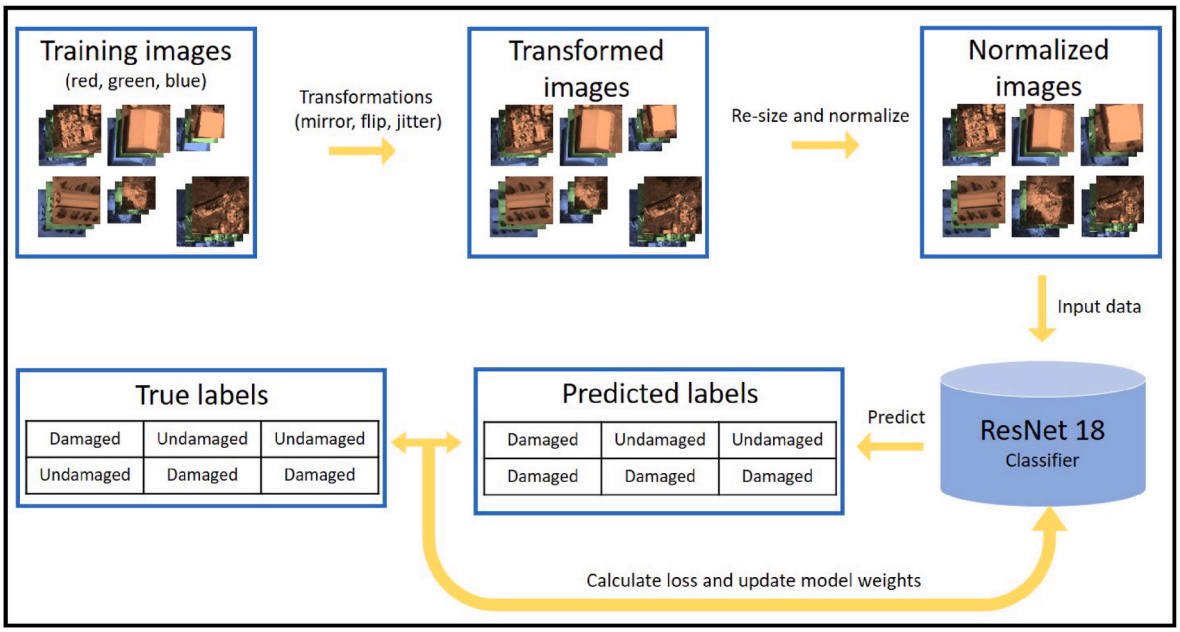


Fig. 3. Methods overview during time of model training. We input red-green-blue images into a classifier to train against known classes and tune model weights.

After preparing the three datasets, we trained the model in the following way. We started with a ResNet18 CNN pre-trained on ImageNet, a dataset that contains 1.3 million images of common objects and scenes. Then, we replaced the output layer with a new layer with only two outputs, and we fine-tuned every layer. Before feeding the training images into the model, we resized them to size 224 x 224 by using bilinear interpolation. Then, we applied random Gaussian noise to the pixels by changing the brightness, contrast, and saturation randomly. Finally, we normalized the images using the training data's mean and standard deviation for all three channels (i.e., red, green, and blue). We applied these transformations to make the model generalize better on new unseen images. For training, we used stochastic gradient descent with Nesterov momentum optimizer and cross entropy loss function. We trained for a total of fifty epochs (saving the model parameters with the best accuracy on the validation set). We chose the hyperparameters after trying many different combinations. The batch size was 128 due to the restriction of the GPU available for training. For our optimizer, we used a learning rate of 0.001, momentum factor of 0.9, and L_2 penalty of 10^{-5}. The entire training process is summarized in Fig. 3.

2.4. Prediction

We tested the trained network on the Camp Fire and the Carr Fire GeoTiff images. We obtained the true labels for both fires from post-wildfire ground surveys [35]. The ground surveyors recorded the location of the damaged buildings that we used to compare against model predictions.

To use the same model to classify damage in buildings of all shapes and sizes, we cropped square images around the building footprints. The dimension of the square was chosen to be the square root of 75th percentile of footprint area of all buildings in the test region. We chose the 75th percentile for the dimension of the cropped image after we qualitatively found that the 25th and 50th percentile areas led to images being too zoomed-in and the 100th percentile led to images being too zoomed-out for majority of the buildings. After obtaining the square crops of the buildings' segments, we normalized each image using the training data's mean and standard deviation for all three channels. Then, we used our model to predict if every structure in the image was undamaged or damaged. This whole process took approximately 60 ms per image on an Nvidia Tesla K80 graphics processing unit. The prediction process is summarized in Fig. 4.

To evaluate our model's performance, we used four key metrics: accuracy, precision, recall, and F_1 score. For our application it is important to correctly identify the damaged buildings even if they are considerably less than the undamaged ones. Therefore, we focussed on the F_1 score, which is considered the most important metric when the false negatives and false positives are crucial. The four metrics are defined below:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$\text{Recall} = \frac{\text{Number of correctly predicted damaged buildings}}{\text{Total number of damaged buildings}}$$

$$\text{Precision} = \frac{\text{Number of correctly predicted damaged buildings}}{\text{Total number of buildings predicted as damaged}}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

3. Results

3.1. xBD validation set

Most of the aerial images collected after wildfire events showed an area with either a very high percentage of damaged buildings or a high percentage of undamaged buildings. This means that in an area with high concentration of damaged buildings, a model could achieve high accuracy by simply predicting every building as damaged. Therefore, it is necessary to investigate the performance of our model for more metrics than just accuracy. Since the main goal of the proposed model is to classify correctly the damaged structures after a wildfire event, recall is a crucial metric to evaluate.

As shown in Table 1 and Fig. 5, when evaluated on the validation set of the xBD, our model had a recall of 0.98, while accuracy, precision, and F_1 score were 0.98, 0.99, and 0.98, respectively. This is a substantial improvement in comparison to a baseline model, a linear classifier trained on xBD that achieved an accuracy of 0.71 on the validation set.

None of the existing models is aimed at solving an identical problem

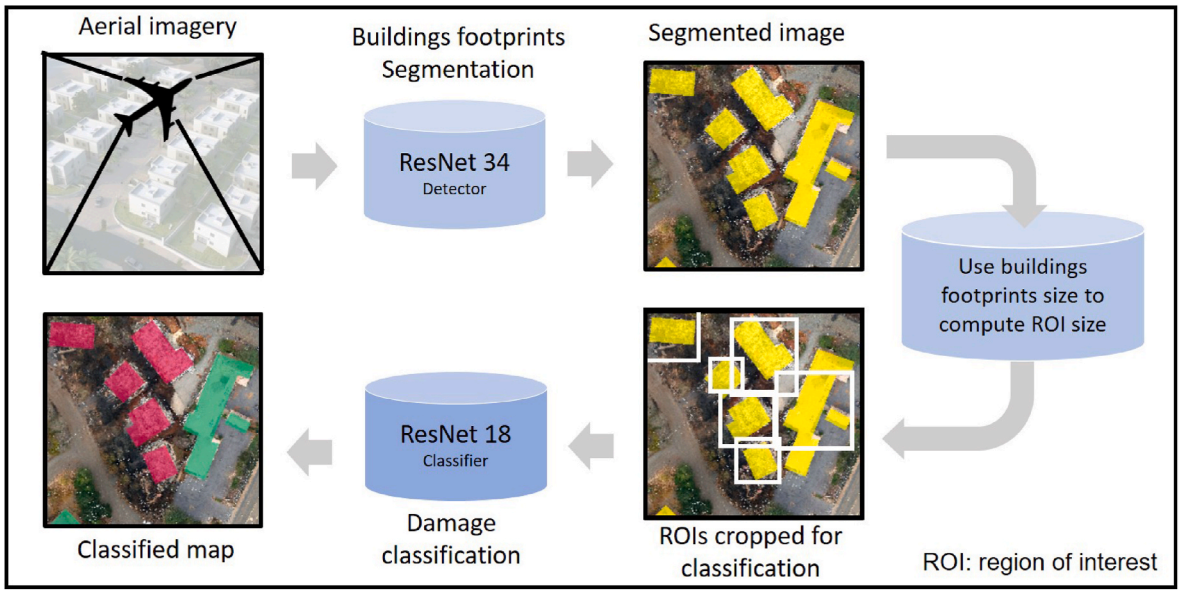


Fig. 4. Methods overview during time of prediction. The DamageMap system ingests an aerial post-wildfire image and outputs polygons for undamaged and damaged buildings. The input image is fed through two convolutional neural networks designed to first segment buildings and then classify whether the buildings were undamaged (green) or damaged (red) due to the wildfire.

Table 1
Evaluation metrics of the model on each dataset. Refer to sec. 3.1 for definitions of metrics.

Dataset	Accuracy	Precision	Recall	F1 Score
xBD wildfires (validation set)	0.98	0.99	0.98	0.98
Camp Fire (Test set 1)	0.92	0.92	0.99	0.96
Carr Fire (Test set 2)	0.98	0.97	0.95	0.96

as our model, but some perform similar tasks. Three of the best attempts for assessing building damage from satellite imagery using the xBD dataset can be viewed in Refs. [24,26,30]. The difference in these models is that they were applied to many natural disasters, including floods, wildfires, and earthquakes. In addition to the variety of natural disasters, these models were also tasked to do both segmentation and classification of buildings into four different classes ("no damage," "minor damage," "major damage," and "destroyed"). In contrast, our model only classified the already segmented buildings into two classes ("undamaged" and "damaged"), as explained in Sec. 2.3. That difference makes a fair comparison between the performance of their models and ours challenging. To address this problem and make the comparison as fair as possible, we compared our model's performance for our two classes ("undamaged" and "damaged") with their models' performance for two of their classes ("undamaged" and "destroyed"), disregarding the other two classes ("minor damaged" and "major damaged"), which none of the existing models can identify accurately. In Table 2 we show the F_1 scores for classes undamaged and damaged of the models from Gupta

Table 2
F_1 score comparison of four different models on the two classes, undamaged and damaged.

Model	Undamaged	Damaged
Weber and Kané	0.906	0.837
Gupta and Shah	0.883	0.808
xView2 challenge winners	0.923	0.864
Our approach	**0.979**	**0.983**

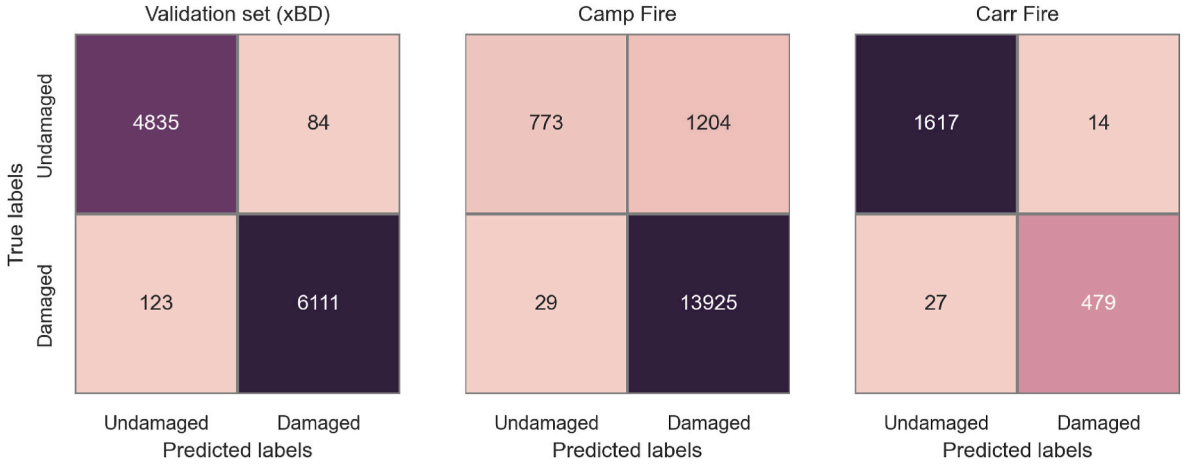


Fig. 5. Confusion matrices showing the performance of our model on each dataset. True labels refers to damage classification obtained from human labelers or ground surveyors. Predicted labels refer to damage classification from our model. Validation set from xBD consists of 12.5% of xBD images. Darker colors correspond to bigger numbers in each matrix.

and Shah [30], Weber and Kané [24], xView2 challenge winners, and our own. Despite the fact that our model displays a better performance, we point out again that a comparison between our model and the other three can never be totally fair due to the reasons mentioned before.

To understand better what makes our model give a false prediction, we analyzed the images that our model failed to classify correctly. A few characteristic images, from the xBD validation set, leading to inaccurate predictions are shown in Fig. 6. To investigate the inaccuracies, saliency maps are presented. These maps are created such that the pixels with higher contribution to the decision making of the classifier display a brighter red color. As seen from Fig. 6, the classifier mostly focused on building pixels when making a decision because most of the pixels with dense red color on the saliency maps correspond to building pixels. In addition, most of the incorrect predictions of the classifier are from buildings that are either not visible in the image (e.g., the building is located under a tree), or damaged but with intact roofs. Finally, sometimes the building of interest is undamaged, but there exists another damaged building in the same image due to errors in segmenting or cropping.

3.2. Camp Fire (Paradise, CA) and Carr Fire (Redding, CA)

A successful model should perform well even on new unseen datasets. To evaluate our model on such datasets, we used the images collected from the Camp Fire and the Carr Fire. These two aerial images were retrieved from aeroplanes rather than satellites. From Fig. 7, one can observe the large differences among the color channel distributions of each dataset caused by the different sources of the imaging. Despite the variations, our model generalized well on new data collected with different methods. In the case of the Camp Fire, it achieved a recall of 0.99, while accuracy, F_1 score, and precision were 0.92, 0.96, and 0.92, respectively, as seen in Fig. 5 and Table 1. A sample of predictions from the Camp Fire are presented in Fig. 9. To investigate the relatively low precision (0.92), a sample of undamaged buildings from the Camp Fire, incorrectly classified as "damaged", are presented in Fig. 8. When four humans were asked to classify 100 building images from the Camp Fire, their precision ranged from 83% to 93%. Thus model performance, despite being low, was on par with human performance in the case of the Camp Fire.

When evaluated on the Carr Fire, our model had a recall of 0.95 and the accuracy, F_1 score, and precision were 0.98, 0.96, and 0.97, respectively, as shown in Fig. 5 and Table 1. Eight representative images from the Carr Fire are shown in Fig. 10. Once again, saliency maps suggest that our model correctly focused on the buildings' pixels to classify the images. In addition, most of the incorrect predictions were difficult to classify properly given only a top view because of the presence of trees or intact roofs.

4. Discussion

4.1. Adequacy of post-wildfire images for damage classification

Obtaining pre- and post-wildfire images under similar lighting, view angle, and camera can be a monumental task for three main reasons. First, since wildfires can occur in any year, maintaining a reliable pre-fire image database would demand mapping the urban and rural world at very high resolution (<1 m) every year. Second, any change in the sensor capturing the pictures would demand a whole new set of pre-fire images captured with the new sensor camera. This is because even minute changes in camera angle and sensor can degrade model performance if higher-level features are not first extracted [45]. Lastly, and perhaps most importantly, building damage assessments immediately after a wildfire can be performed with airborne cameras only because lingering smoke and clouds may obstruct spaceborne imagery. This leaves disaster assessment using post-wildfire images as the only option, since airborne surveys are hardly ever performed at regular intervals. Thus, by exclusively using post-wildfire images, we reduced data sourcing challenges.

Although earlier studies showed that using only post-disaster images may result in significant under-counting of damaged buildings [32], more recent studies using deep learning have found that the loss in accuracy is negligible. In the case of Xu et al. [45], accuracy dropped from 0.83 to 0.80. The results of our study show through the five validation fires (Fig. 6) and two explicit test cases (Figs. 9 and 10) that our deep learning model maintained high accuracy with solely post-wildfire images. The saliency maps in the validation as well as the test set show that when the model's predictions were correct, it focused correctly on the buildings (indicated by brighter red colors over the footprint area). Validation as well as test accuracy for damage classification exceeded 90% in all cases (Table 1), and exceeded previous approaches (Table 2), suggesting that post-disaster images alone might be sufficient for a binary building damage classification task.

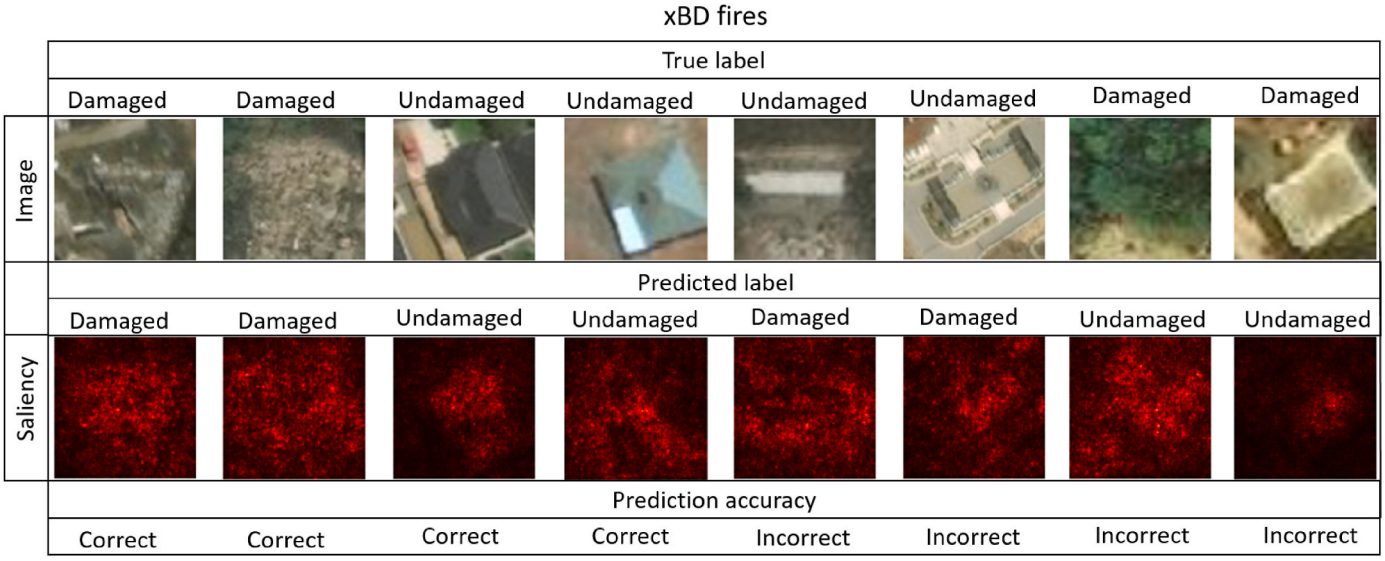


Fig. 6. xBD wildfire images and prediction visualization. A sample of input images after cropping around building segments is provided in the top row. Corresponding saliency maps from our model are in the bottom row. The saliency maps are created such that the pixels with higher contribution to the decision making of the classifier display a brighter red color.

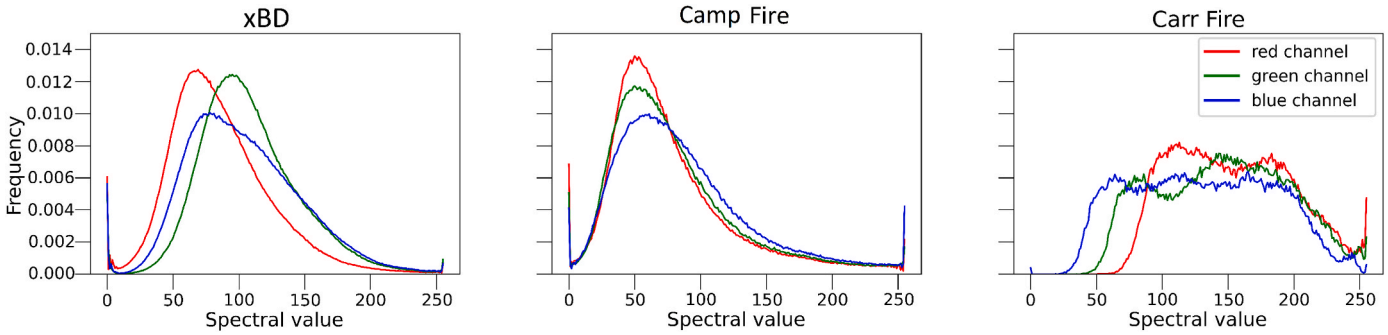


Fig. 7. Color channel histograms for the three datasets in our study. Both the Camp Fire and xBD wildfires had similar histograms whereas the Carr Fire images had a bi-modal distribution due to its highly urbanised area.

Fig. 8. A sample of images of undamaged buildings from the Camp Fire, incorrectly classified as "damaged". Such false positives mostly occur when tree canopies cover the building roofs either partially or fully.

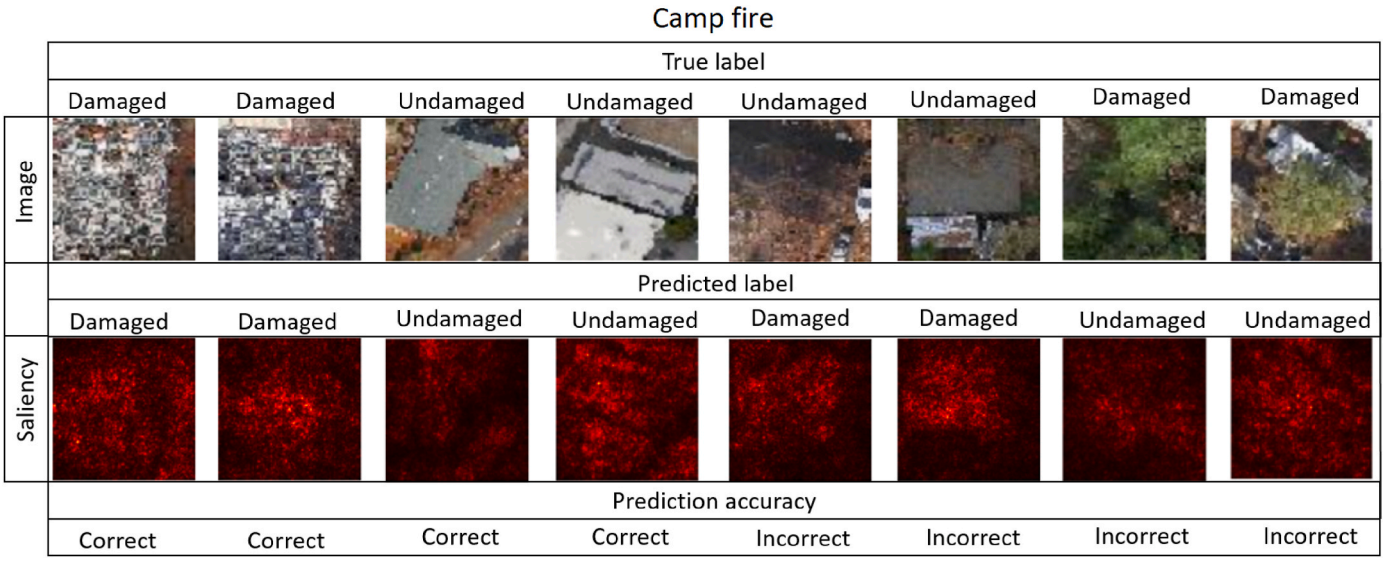


Fig. 9. Camp Fire images and prediction visualization. A sample of input images after cropping around building segments are provided in the top row, and corresponding saliency maps are in the bottom row.

The metrics of both test cases combined with the saliency maps suggest that our model can be used in new datasets with promising results. Despite the fundamental differences between the three different datasets (xBD, Camp Fire, Carr Fire) the classifier was able to correctly predict building damage for most buildings. With the trained and validated model, we created a web application [46] to allow users to interactively explore damage predictions on the test set.

4.2. Causes for prediction error

Although our model exhibited an overall accuracy exceeding 98% on the validation set and in the Carr Fire, its performance was considerably lower in the Camp Fire, with accuracy = 92% (Table 1). The overwhelming majority of the incorrect examples in the Camp Fire were of buildings that were actually undamaged but predicted as damaged (precision = 92%, but recall = 99%, Table 1). An investigation of the incorrectly classified images from the Camp Fire revealed that model performance suffered in two scenarios. First, when trees covered undamaged buildings, the model predicted them as "damaged" (Figs. 8 and 9). This was perhaps caused by the similarity in texture between tree canopies and damaged buildings' debris. Secondly, the model incorrectly classified images which had a smaller building and a relatively higher portion of the background landscape visible in the cropped region of interest. Brown ground appears visually similar to damaged roofs, thus degrading the accuracy of our model in the Camp Fire.

In contrast, the images from the Carr Fire were sourced from a

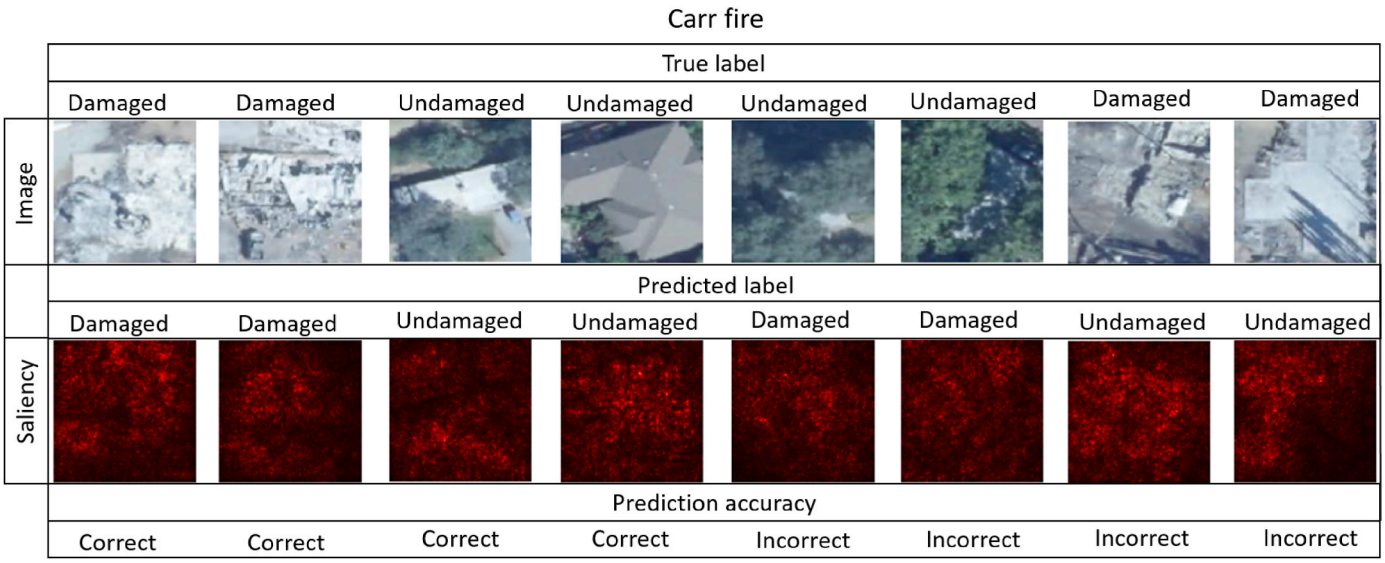


Fig. 10. Carr Fire images and prediction visualization. Eight building images and their true labels are provided in the top row. Corresponding predicted labels and saliency maps are in the bottom row.

developed environment (Fig. 10) that had brighter background colors (Fig. 7) and thus lesser likelihood of encountering brown ground. Furthermore, owing to the built-up environment of the Carr Fire's location, trees were less likely to be present adjacent to the buildings, thus making classification easier and more robust. Nevertheless, given the fact that in the Camp Fire, the errors are of commission and not omission, the lower accuracy is less concerning for use in an operational setting since the resulting damage assessment is more conservative. Moreover, since model performance matched human-level prediction (92% for model, 93% for humans), it is likely that model accuracy may not be improved with the current image quality of 0.3 m resolution.

4.3. Model generalization

The ability of empirical models to generalize to new unseen data can be beneficial in an operational environment. New images sourced from future fires could be from new places where the current model was neither trained nor validated. Thus, to quantify generalized model accuracy, it is insufficient to evaluate it on the validation set alone (which is typically used to tune model hyperparameters). An independently-sourced test set is needed to understand true ability of models to generalize. This is because the validation data contain buildings from the same place, with similar roof architectures, roof colors, or background colors unlike in the case of test data where images are sourced from completely different places using different instruments (e.g., aerial cameras as opposed to satellite-based cameras).

On comparing the model performance on different datasets, we observed that the accuracy remained almost unchanged between the training set (99%), validation set (98%), and test sets (98% in the Carr Fire and 92% in the Camp Fire, Table 1). This indicates that the tuning of model hyperparameters and data normalization likely helped prevent overfitting to the training or the validation datasets, despite significant differences in their color spectrum (Fig. 7).

For model tuning, we imposed a stringent regularization mechanism by constraining the L_2 norm of the weight matrix. We also introduced random changes to brightness, contrast, and saturation of the input images between each epoch to help the model generalize better. Finally, we subtracted the mean reflectances from all bands before dividing it by the standard deviation to normalize the color spectrum of each image (Fig. 3). This helped maintain high model performance on the Carr Fire data (test set) despite its vastly different color histogram (Fig. 7). The saliency maps show that the model's attention was correctly focused on the building footprints, instead of the background landscape in the Carr Fire (Fig. 10). This finding is consistent with other image recognition studies where spectrum normalization and random shifting of image brightness and contrast have helped improve overall model performance on unseen test data [47].

On the other hand, relying on the validation set alone may not always reveal a model's ability to generalize. This is because the validation set is usually comprised of images similar to the images in the training set without changes in the type of cameras used to source the images, camera angle, or camera resolution. For instance, when Xu et al. [45], trained their CNN model on a dataset of damaged buildings in Mexico, they found that the model had an accuracy of 71% on unseen data in Mexico. However, the accuracy dropped to 60% when the model was trained on a dataset from Haiti and tested in Mexico. In spite of the need to test model generalization ability, most studies of building damage detection in the past were validated on a portion of the same data source (for e.g., 30, 48, 26, 49 with the notable exception of Weber et al., [24]). Thus, there is an urgent need to test model generalizability on not just unseen data from the same data source, but also on images from new places, captured under different settings.

4.4. Limitations and future work

There are some limitations to our study. Incorrect classification of building damage mostly occurred at buildings located under trees or partially damaged buildings. Future efforts must address classification challenges with such partial information. Since buildings adjacent to trees are likely to be present in intermix zones of the WUI, it is of prime importance that any damaged building classifier perform well in such regions due to the increased risk of wildfire damage in such areas [5]. To overcome the challenges posed by tree-covered buildings for building damage classification, three solutions may help. First, a more advanced cropping technique for selecting regions of interest could be used. Instead of just cropping an image around building footprints, pixels not belonging to the building footprints may be masked out before classifying the images [50,51]. Second, using imagery with more spectral bands may help distinguish between built-up area and vegetation. In particular, near-infrared band reflectance, or an index constructed from it such as the normalized difference vegetation index can assist the model to distinguish between rooftops and vegetation canopy due to direct sensitivity to chlorophyll content in vegetation [52]. Third, microwave radars may be used to penetrate through thin canopies [53]. An

added advantage of microwaves is that they can penetrate smoke and clouds, thus allowing for consistent data capturing. Microwave synthetic aperture radar (SAR) has shown promise in damage identification, although the speckle (spatial noise) in SAR images has posed significant challenges in accurately identifying damage at the building-scale [54].

In addition, our model might under-perform with new data collected from unique and rare environments that our model has never seen in the training and validation process. Another limitation is that our model is a binary classifier, whereas for more precise damage assessment we might need more granular damage classifications. Supplementing aerial imagery with other remotely-sensed information like microwave backscatter may help in this regard. Finally, even though our model does not need images just before the wildfire for classifying building damage, it still needs an image taken at some time before the fire to identify building perimeters (which are then fed to the classifier). However, if the time period between the image collection and fire is considerably long, our approach will miss buildings that were constructed in that time period. This is of particular concern in places experiencing rapid expansion of the WUI [55]. In such cases, locally-sourced information from authorities can help in obtaining the latest information on location of buildings to supplement the segmentation task. This can be done by taking the union of known location of buildings from local authorities and from the output of Microsoft's building footprints dataset.

The issue of building damage identification is not unique to wildfire disasters alone. Other natural disasters such as hurricanes and earthquakes also demand rapid damage assessments [26], since current methods are manual and laborious [56–59]. Extending DamageMap to other natural disasters can thus be beneficial. Dataset compiled by Ref. [26] may be used for this purpose, although independent test sets to estimate model generalizability are lacking. Furthermore, an analysis of the type of damage caused by each natural disaster would be required because damages caused by floods, for instance, may not be apparent using aerial imagery alone.

5. Conclusion

The artificial intelligence-powered DamageMap has the potential to assist disaster managers and first responders in building damage assessment and post-disaster recovery. Unlike other studies that require imagery both before and after wildfires occur, our model achieved high accuracy in classifying building damages by only using post-wildfire images. Our results suggest that our model can potentially save time and reduce the costs of collecting pre-disaster aerial imagery. Moreover, most of the existing models of building damage classification for wildfires used the same data sources for training and validation, so their accuracy may drop significantly when applied to other unseen data at different locations. In contrast, our model used different sources of training and test datasets and showed promising results in both the validation set of the xBD dataset, and the test set of wildfire events, such as the Camp Fire and the Carr Fire. These promising results demonstrate that our model has the potential to generalize well to unseen datasets of wildfires and other natural disasters.

Author contributions

GAF conceived of the study. All authors designed the analysis. MG and KR performed the analysis. All authors interpreted the results and contributed to writing of the manuscript.

Data accessibility statement

The damage predictions on the test set can be interactively explored at https://kkraoj.users.earthengine.app/view/damagemap. The scripts supporting the analysis can be found at https://github.com/MariosGalanis/DamageMap.

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.

Acknowledgements

Our research was supported by the California Polytechnic State University, San Luis Obispo 2019 Research, Scholarly, and Creative Activities Grant Program. The research project was started by Gustave Rousselet as part of his undergraduate thesis at Cal Poly and the Swedish KTH, School of Electrical Engineering and Computer Science. The authors would like to thank Google for hosting the 2019 Geo For Good conference where KR and GF built an initial prototype of DamageMap. The authors would also like to thank William Brewer and Steven Hawks from Cal Fire for allowing access to the imagery and field collected damage points, Jonas Junnior and Brendan Palmieri for helping with exploratory work for this study, and Derek Fong and Rebecca Miller for providing guidance during the early stages of the project. KR was funded by the NASA Earth and Space Science Fellowship and the Stanford Data Science Scholarship.

References

[1] FEMA, Preliminary damage assessment guide. FEMA Preliminary Damage Assessment Guide, FEMA, May 2020. https://www.fema.gov/disasters/preliminary-damage-assessment-reports/guide.
[3] Volker C. Radeloff, David P. Helmers, H. Anu Kramer, Miranda H. Mockrin, Patricia M. Alexandre, Avi Bar-Massada, Van Butsic, Todd J. Hawbaker, Sebastián Martinuzzi, Alexandra D. Syphard, Susan I. Stewart, Rapid growth of the US wildland-urban interface raises wildfire risk, in: Proceedings of the National Academy of Sciences of the United States of America, vol. 115, 2018, pp. 3314–3319, 13.
[4] W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, David M.J.S. Bowman, Climate-induced variations in global wildfire danger from 1979 to 2013, Nat. Commun. 6 (May) (2015) 1–11.
[5] Heather Anu Kramer, Miranda H. Mockrin, Patricia M. Alexandre, Volker C. Radeloff, High wildfire damage in interface communities in California, Int. J. Wildland Fire 28 (9) (2019) 641–650.
[6] Philip E. Dennison, James D. Brewer, C. Simon, Arnold, Max A. Moritz, Large wildfire trends in the western United States, Geophys. Res. Lett. 41 (April) (2014) 6413–6419.
[7] Anthony Le Roy Westerling, Increasing western US forest wildfire activity: sensitivity to changes in the timing of spring, Phil. Trans. Biol. Sci. 371 (1696) (2016).
[8] B.M. Wotton, M.D. Flannigan, G.A. Marshall, Potential climate change impacts on fire intensity and key wildfire suppression thresholds in Canada, Environ. Res. Lett. 12 (9) (2017).
[9] David M.J.S. Bowman, Andrés Moreira-Muñoz, Crystal A. Kolden, Roberto O. Chávez, Ariel A. Muñoz, Fernanda Salinas, Álvaro González-Reyes, Ronald Rocco, Francisco de la Barrera, Grant J. Williamson, Nicolás Borchers, Luis A. Cifuentes, John T. Abatzoglou, Fay H. Johnston, Human–environmental drivers and impacts of the globally extreme 2017 Chilean fires, Ambio 48 (4) (2019) 350–362.
[10] T.J. Brown, B. Hall, A. Westerling, The impact of twenty-first century climate change on wildland fire danger in the western united, Climatic Change 62 (2004) 365–388.
[11] Susana Gómez-González, Fernando Ojeda, Paulo M. Fernandes, Portugal and Chile: longing for sustainable forestry while rising from the ashes, Environ. Sci. Pol. 81 (2018) 104–107. July 2017.
[12] Paul Fox-Hughes, Rebecca Harris, Greg Lee, Michael Grose, Nathan Bindoff, Future fire danger climatology for tasmania, Australia, using a dynamically downscaled regional climate model, Int. J. Wildland Fire 23 (3) (2014) 309–321.
[13] M.D. Flannigan, K.A. Logan, B.D. Amiro, W.R. Skinner, B.J. Stocks, Future area burned in Canada, Climatic Change 72 (1–2) (2005) 1–16.
[14] Sirio Modugno, Heiko Balzter, Beth Cole, Pasquale Borrelli, Mapping regional patterns of large forest fires in Wildland-Urban Interface areas in Europe, J. Environ. Manag. 172 (2016) 112–126.
[15] Michael Buxton, Rachel Haynes, David Mercer, Andrew Butt, Vulnerability to bushfire risk at melbourne's urban fringe: the failure of regulatory land use planning, Geogr. Res. 49 (1) (2011) 1–12.
[16] Raul Romero-Calcerrada, C.J. Novillo, J.D.A. Millington, I. Gomez-Jimenez, GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain), Landsc. Ecol. 23 (3) (2008) 341–354.
[17] Jennifer K. Balch, Bethany A. Bradley, John T. Abatzoglou, R. Chelsea Nagy, Emily J. Fusco, Adam L. Mahood, Human-started wildfires expand the fire niche across

the United States, in: Proceedings of the National Academy of Sciences of the United States of America, vol. 114, 2017, pp. 2946–2951, 11.
[18] Damage assessments for properties affected by the glass fire, Sonoma County Emergency and Preparedness Information, County of Sonoma, California, Oct 2020. https://socoemergency.org/recover/glass-wildfire/damage-assessments/.
[19] Jia-hang Liu, Xin-jian Shan, Jing-yuan Yin, Automatic recognition of damaged town buildings caused by earthquake using remote sensing information: taking the 2001 bhuj, India, earthquake and the 1976 tangshan, China, earthquake as examples, Acta Seismol. Sin. (Engl. Ed.) 17 (6) (2004) 686–696.
[20] Yalkun Yusuf, Masashi Matsuoka, Fumio Yamazaki, Damage assessment after 2001 Gujarat earthquake using landsat-7 satellite images, Journal of the Indian Society of Remote Sensing 29 (1–2) (2001) 17–22.
[21] Paolo Gamba, Fabio Casciati, Gis and image understanding for near-real-time earthquake damage assessment, Photogramm. Eng. Rem. Sens. 64 (1998) 987–994.
[22] Fuzhou Duan, Huili Gong, Wenji Zhao, Collapsed houses automatic identification based on texture changes of post-earthquake aerial remote sensing image, in: 2010 18th International Conference on Geoinformatics, IEEE, 2010, pp. 1–5.
[23] Colin Alstad, The xView2 AI Challenge, IBM, 2020. https://www.ibm.com/cloud/blog/the-xview2-ai-challenge.
[24] Ethan Weber, Hassan Kané, Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion, 2020. https://arxiv.org/pdf/2004.05525.pdf.
[25] Hanxiang Hao, Sriram Baireddy, Emily R. Bartusiak, Latisha Konz, Kevin LaTourette, Michael Gribbons, Moses Chan, Mary L. Comer, Edward J. Delp, An Attention-Based System for Damage Assessment Using Satellite Imagery, 2020. https://arxiv.org/abs/2004.06643.
[26] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, Matthew Gaston, xBD: A Dataset for Assessing Building Damage from Satellite Imagery, 2019. https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/Gupta_Creating_xBD_A_Dataset_for_Assessing_Building_Damage_from_Satellite_CVPRW_2019_paper.html.
[27] Laigen Dong, Jie Shan, A comprehensive review of earthquake-induced building damage detection with remote sensing techniques, ISPRS J. Photogrammetry Remote Sens. 84 (2013) 85–99.
[28] Fumio Yamazaki, Masashi Matsuoka, Remote sensing technologies in post-disaster damage assessment, Journal of Earthquake and Tsunami 1 (3) (2007) 193–210.
[29] J. Arn Womble, Kishor C. Mehta, Beverley J. Adams, Automated Building Damage Assessment Using Remote-Sensing Imagery, Forensic Engineering, 2007.
[30] Rohit Gupta, Mubarak Shah, RescueNet: Joint Building Segmentation and Damage Assessment from Satellite Imagery, 2020. https://ieeexplore.ieee.org/abstract/document/9412295?casa_token=sLEEDzWxrH8AAAAA:FNXAZLNtRzxxmqVRa5d6JRhZzGKZTcHKHFH4RVjyKDunC_gjf4Ozc8t5RYAx8efLhsU5LcqeKB0.
[31] Neal T. Anderson, Giovanni B. Marchisio, Worldview-2 and the evolution of the digitalglobe remote sensing satellite constellation: introductory paper for the special session on worldview-2, in: Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XVIII, ume 8390, International Society for Optics and Photonics, 2012, p. 83900L.
[32] Keiko Saito, Robin Spencev, A. Terence, De C. Foley, Visual damage assessment using high-resolution satellite images following the 2003 Bam, Iran, earthquake, Earthq. Spectra 21 (SUPPL. 1) (2005) 309–318.
[33] Xiaohua Tong, Zhonghua Hong, Shijie Liu, Xue Zhang, Huan Xie, Zhengyuan Li, Sonlin Yang, Weian Wang, Feng Bao, Building-damage detection using pre- and post-seismic high-resolution satellite stereo imagery: a case study of the May 2008 Wenchuan earthquake, ISPRS J. Photogrammetry Remote Sens. 68 (1) (2012) 13–27.
[34] Turker MUSTAFA, B.T. San, Detection of collapsed buildings caused by the 1999 izmit, Turkey earthquake through digital analysis of post-event aerial photographs, Int. J. Rem. Sens. 25 (21) (2004) 4701–4714.
[35] The Department of Forestry and Fire Protection of California, Camp Fire Information, 2019. https://www.fire.ca.gov/incidents/2018/11/8/camp-fire/.
[36] Microsoft. US Building Footprints, 2019. https://github.com/Microsoft/USBuildingFootprints.
[37] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.
[38] Sasha Targ, Diogo Almeida, Kevin Lyman, Resnet in Resnet: Generalizing Residual Architectures, 2016. https://arxiv.org/abs/1603.08029.
[39] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105.
[40] Limin Wang, Sheng Guo, Weilin Huang, Yu Qiao, Places205-vggnet Models for Scene Recognition, 2015. https://arxiv.org/abs/1508.01667.
[41] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, Kurt Keutzer Squeezenet, Alexnet-level Accuracy with 50x Fewer Parameters and <0.5mb Model Size, 2016. https://arxiv.org/abs/1602.07360.
[42] Forrest Iandola, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell, Kurt Keutzer, Densenet: Implementing Efficient Convnet Descriptor Pyramids, 2014. https://arxiv.org/abs/1404.1869.
[43] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei Imagenet, A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[45] Joseph Z. Xu, Wenhan Lu, Zebo Li, Pranav Khaitan, Valeriya Zaytseva, Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks, 2019. https://arxiv.org/abs/1910.06444.
[46] Marios Galanis, Krishna Rao, Yi-Lin Tsai, Xinle Yao, Jonathan Ventura, G. Andrew Fricker, DamageMap Web Application, 2020. https://kkraoj.users.earthengine.app/view/damagemap.
[47] Eduardo Romera, Luis M. Bergasa, Jose M. Alvarez, Mohan Trivedi, Train here, deploy there: robust segmentation in unseen domains, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1828–1833.
[48] Alessandro Farasin, Luca Colomba, Paolo Garza, Double-step U-Net: a deep learning-based approach for the estimation ofwildfire damage severity through sentinel-2 satellite data, Appl. Sci. 10 (12) (2020) 1–22.
[49] Derek McNamara, William Mell, Alexander Maranghides, Object-based post-fire aerial image classification for building damage, destruction and defensive actions at the 2012 Colorado Waldo Canyon Fire, Int. J. Wildland Fire 29 (2) (2020) 174–189.
[50] George Vosselman, Maximilian Coenen, Franz Rottensteiner, Contextual segment-based classification of airborne laser scanner data, ISPRS J. Photogrammetry Remote Sens. 128 (2017) 354–371.
[51] Kourosh Khoshelham, Sander Oude Elberink, Sudan Xu, Segment-based classification of damaged building roofs in aerial laser scanning data, Geosci. Rem. Sens. Lett. IEEE 10 (5) (2013) 1258–1262.
[52] Compton J. Tucker, Red and photographic infrared linear combinations for monitoring vegetation, Rem. Sens. Environ. 8 (2) (may 1979) 127–150.
[53] Fawwaz T. Ulaby, David G. Long, Microwave Radar and Radiometric Remote Sensing, University of Michigan Press, Ann Arbor, 2014.
[54] Plank Simon, Rapid damage assessment by means of multi-temporal sar—a comprehensive review and outlook to sentinel-1, Rem. Sens. 6 (6) (2014) 4870–4906.
[55] Sebastian Martinuzzi, Susan I. Stewart, David P. Helmers, Miranda H. Mockrin, Roger B. Hammer, Volker C. Radeloff, The 2010 Wildland-Urban Interface of the Conterminous United States, United States Department of Agriculture, 2015, pp. 1–122.
[56] David Lallemant, Robert Soden, Steven Rubinyi, Sabine Loos, Karen Barns, Gitanjali Bhattacharjee, Post-disaster damage assessments as catalysts for recovery: a look at assessments conducted in the wake of the 2015 gorkha, Nepal, earthquake, Earthq. Spectra 33 (1_suppl) (2017) 435–451.
[57] World Bank, Safer Homes, Stronger Communities: A Handbook for Reconstructing after Natural Disasters, 2010.
[58] Carol C. Massarra, Hurricane Damage Assessment Process for Residential Buildings, Master's thesis, Louisiana State University and Agricultural and Mechanical College, 2012.
[59] Carol J. Friedland, Residential Building Damage from Hurricane Storm Surge: Proposed Methodologies to Describe, Assess and Model Building Damage, PhD thesis, Louisiana State University and Agricultural and Mechanical College, 2009.