frontiers | Frontiers in Environmental Science

Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages

Danu Kim[1,2], Jeongkyung Won[3], Eunji Lee[1,2], Kyung Ryul Park[4], Jihee Kim[2,5], Sangyoon Park[6], Hyunjoo Yang[3]* and Meeyoung Cha[1,2]*

[1]School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Korea, [2]Data Science Group, Institute for Basic Science, Daejeon, Korea, [3]School of Economics, Sogang University, Seoul, Korea, [4]Graduate School of Science and Technology Policy, Korea Advanced Institute of Science and Technology, Daejeon, Korea, [5]School of Business and Technology Management, College of Business, Korea Advanced Institute of Science and Technology, Daejeon, Korea, [6]Business School, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

The increasing frequency and severity of water-related disasters such as floods, tornadoes, hurricanes, and tsunamis in low- and middle-income countries exemplify the uneven effects of global climate change. The vulnerability of high-risk societies to natural disasters has continued to increase. To develop an effective and efficient adaptation strategy, local damage assessments must be timely, exhaustive, and accurate. We propose a novel deep-learning-based solution that uses pairs of pre- and post-disaster satellite images to identify water-related disaster-affected regions. The model extracts features of pre- and post-disaster images and uses the feature difference with them to predict damage in the pair. We demonstrate that the model can successfully identify local destruction using less granular and less complex ground-truth data than those used by previous segmentation models. When tested with various water-related disasters, our detection model reported an accuracy of 85.9% in spotting areas with damaged buildings. It also achieved a reliable performance of 80.3% in out-of-domain settings. Our deep learning-based damage assessment model can help direct resources to areas most vulnerable to climate disasters, reducing their impacts while promoting adaptive capacities for climate-resilient development in the most vulnerable regions.

KEYWORDS

natural disaster, computer vision, machine learning, daytime satellite imagery, damage detection, disaster response

1 Introduction

The widespread impacts of human-induced climate change have been observed as the frequency and intensity of extreme events, including floods, tornadoes, hurricanes, and tsunamis, increase (IPCC, 2022). Amid growing climate risk, the global adaptive capacity to deal with disasters has not progressed accordingly, although the Sustainable Development Goals call for a worldwide response (Field et al., 2012). The lack of timely, comprehensive, and accurate data tracking damage at a fine-grained geographical level is one of the main reasons for such an adaptation deficit (Amundsen et al., 2010; Moser and Ekstrom, 2010). For example, damage estimates are available only at the province level in the Emergency Events Database (EM-DAT), the largest international disaster database, making it difficult to pinpoint the worst-hit areas. Moreover, conventional damage assessments using field surveys are resource-intensive and time-consuming (Cao and Choe, 2020), which hinders comprehensive regional coverage (Bakkensen et al., 2018) and the rapid deployment of humanitarian assistance (Cao and Choe, 2020). Field surveys may also suffer from cognitive biases, such as reference dependence and recall errors (Guiteras et al., 2015).

Recent research in computer vision has combined high spatial resolution satellite images with machine learning to estimate disaster damage at a pixel or incident level (Potnis et al., 2019; Bai et al., 2020; Weber and Kané, 2020; Gupta and Shah, 2021; Wu et al., 2021). These approaches use the XBD dataset created by Gupta et al. (2019), the largest disaster damage dataset worldwide, providing pre-and post-disaster images with pixel-level damage labels. While the models developed using the dataset have made significant technical advances, they are fundamentally dependent on the existence of such fine-grained, complex damage labels in disaster-affected regions. Given that most high-quality ground-truth data comes from developed countries, a deep learning model that combines various forms of damage data and produces accurate local damage estimates would be helpful in on the ground disaster response efforts.

This paper presents a lightweight damage detection model based on deep learning and high spatial resolution satellite images. Our ground-truth data is less granular than the data employed by existing segmentation models. This feature is advantageous for developing countries lacking the statistical capacity and resources to produce quality local damage data. Evaluating our model against various water-related disasters from 2011 to 2019, our model achieved a performance of 90% in detecting disasters. In addition, the model showed a reasonable performance of 80% in regions not observed during training. Our case study on Providencia Island further demonstrates the generalizability of our

model, as it successfully distinguished local destruction caused by Hurricane Iota in 2021.

We focused on water-related disasters, such as floods, tornadoes, tsunamis, and hurricanes, among many others, given their increasing frequency and severity in the most vulnerable low-and middle-income countries (Hallegatte et al., 2013; Edmonds et al., 2020; Rentschler et al., 2022). Despite substantial losses caused by disasters, a significant population is unable to leave disaster-prone regions (Tellman et al., 2021) for socioeconomic and political reasons (Hunter, 2005; Raker, 2020; Lin et al., 2021; Henkel et al., 2022). This necessitates an effective post-disaster response; therefore, local damage estimates can be especially helpful in prioritizing relief efforts and climate-resilient redevelopment.

The high-resolution satellite images have several advantages over other spatial data sets (i.e., Google Street views and aerial images) used in the related literature (Fujita et al., 2017; Vetrivel et al., 2018). First, satellite imagery guarantees extensive spatial and temporal coverage. Second, satellite images do not always require resource-intensive damage labels for training, unlike other spatial input data used in earlier literature (Fujita et al., 2017; Vetrivel et al., 2018). Instead, they can combine various forms of damage labels (e.g., pixel-level labels, point coordinates of damaged properties, and district-level statistics) corresponding to their size.

The proposed model makes several methodological contributions to environmental damage detection. First, the model effectively identifies local destruction by employing binary damage labels corresponding to satellite images' size. This approach reduces the reliance of deep-learning models on fine-grained, complex ground-truth data, making the model more applicable to developing countries without access to such data. Second, to the best of our knowledge, our model is the first successful water-related disaster damage detection model. Third, our model is practical because it does not require extensive ground truth data specific to the damaged regions and its performance remains robust in unseen regions.

Our algorithm can help policymakers by identifying the ideal location for humanitarian assistance deployment and minimizing the time lag between the onset of a disaster and assistance responses. The model provides predicted labels to help determine a measure of centrality for the location where resources should be concentrated. As this machine learning-based assessment can be implemented faster and cheaper than conventional on-site inspections, development agencies would be able to deploy well-targeted humanitarian aids in less time with lower costs.

2 Related work

Recent works have explored the potential for remote sensing data to assess societies' average exposure to

disasters (Smith et al., 2019; Tellman et al., 2021). Addressing the limitations of conventional sources, some studies have combined remote sensing data with neural networks to measure disaster-incurred destruction directly. Fujita et al. (2017); Amit and Aoki (2017); Duarte et al. (2018) applied deep-learning-based models to pre- and post-disaster imagery data to classify regions into damaged or undamaged areas. The proposed classification models were trained with a large number of damage labels specific to their input image data. With methodological advances in machine learning techniques, semantic segmentation models built on detailed pixel-level ground-truth data have also emerged (Potnis et al., 2019; Bai et al., 2020; Weber and Kané, 2020; Gupta and Shah, 2021; Wu et al., 2021).

Segmentation-based damage assessment using satellite imagery is an environmental application of change detection (Janalipour and Taleai, 2017). Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times (Singh, 1989). Change detection models are either unsupervised or supervised. Unsupervised methods use clustering (Celik, 2009; Mehrotra et al., 2015), thresholding (Ghanbari and Akbari, 2015; Khanbani et al., 2021), or optimization (Kusetogullari et al., 2015) to find intrinsic differences between images without any prior information. On the other hand, supervised methods are more common and can achieve higher accuracy than unsupervised methods. However, they also require a large amount of pixel-level labels for training (Zhu, 2017; Zou et al., 2022), which tend to be the most expensive labels in the computer vision field, as they are costly and time-consuming to gather.

The xBD dataset constructed by Gupta et al. (2019)), significantly extended the coverage of pixel-level labels in the damage assessment field. The objective of the xBD challenge is a particular semantic segmentation task: first to locate a building's footprint and then estimate the damage to each building. The dataset contains pre- and post-disaster images along with pixel-level categorization for building damage. This dataset sparked the development of numerous segmentation-based damage detection models; for example, Weber and Kané (2020); Wu et al. (2021) proposed U-Net-based models, and Bai et al. (2020); Gupta and Shah (2021) used pyramid pooling modules. Both architectures are commonly used for image segmentation tasks.

While these segmentation models crucially rely on the pixel-level ground-truth data, as in the original dataset, our grid-level prediction model does not require such granular labeling data. Thus, we used a simplified version of the xBD dataset. We aimed to predict damage at rectangular-shaped grids of size $0.01 km^2$, which can still capture the localized nature of disaster damages, but at much less computational cost and with fewer data constraints. This lightweight feature can be especially helpful for rapid disaster responses in many countries.

One of the valuable global data sources in the damage assessment field is provided by UNOSAT, the operational satellite applications program of UNITAR[1]. This dataset assesses building damage using a five-point scale, ranging from No-Damage to Destroyed, with point locations. Despite the extensive coverage of the dataset, especially in developing countries, the models developed for semantic segmentation tasks cannot use this data as it does not include damage information at the pixel-level. One model that has been able to use the UNOSAT data set is Xu et al.'s (2019) state-of-the-art model. They utilized this data to build a binary classifier for detecting building damage at the grid level. We applied our disaster events to their model architecture and compared our model's performance relative to theirs.

3 Methods

We present a binary classification model that detects damage in a region from satellite image data. Our model identifies the rapid change mainly in building structures to distinguish the damage in target regions. Model input is a pair of images taken over the same geographic region before and after a disaster. Our model aims to determine whether a given region was substantially damaged due to a disaster. Due to the fixed temporal resolution of satellites, one great limit of the satellite-based approach is the difficulty of obtaining images right before and after a disaster. Such a time gap inevitably brings simple visual changes over the season and even general urban developments (e.g., the construction of buildings and roads). Our model needs to learn damage-specific features rather than those simple visual changes, to report disaster damage accurately.

We used transfer learning to train our model effectively, following the methods of Jean et al. (2016); Xie et al. (2016). Transfer learning consists of two steps: pre-training and fine-tuning. Pre-training helps the model learn general low-level features of the image with more straightforward tasks on a large dataset. After the pre-training, the model is fine-tuned to fit the objective of the target task. We followed this step and first pre-trained the network with a simple classification task for one satellite image. Only satellite imagery of non-disaster situations was used during pre-training, which is much easier to gather. After the pre-training, we fine-tuned our model to detect the damage from a pair of satellite images. The model can learn to determine whether the given region is destroyed effectively, taking advantage of the learned general geographical features from the pre-training that are
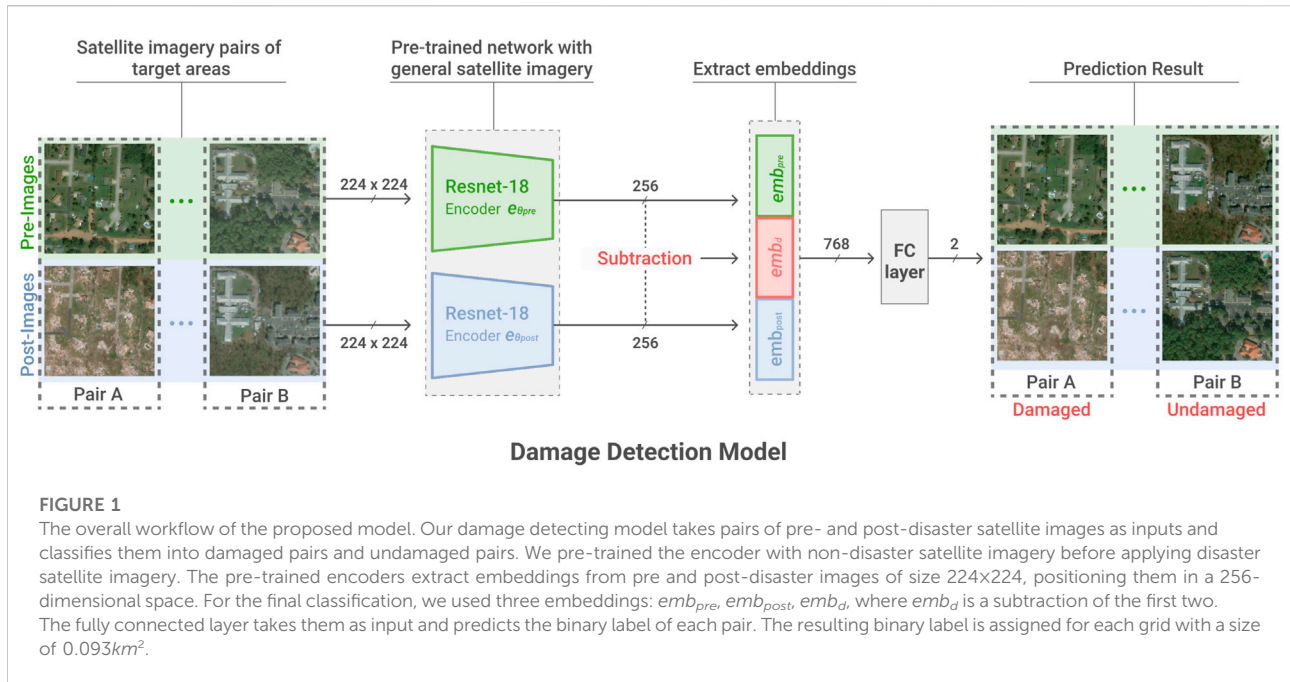

Damage Detection Model

FIGURE 1
The overall workflow of the proposed model. Our damage detecting model takes pairs of pre- and post-disaster satellite images as inputs and classifies them into damaged pairs and undamaged pairs. We pre-trained the encoder with non-disaster satellite imagery before applying disaster satellite imagery. The pre-trained encoders extract embeddings from pre and post-disaster images of size 224×224, positioning them in a 256-dimensional space. For the final classification, we used three embeddings: emb_{pre}, emb_{post}, emb_d, where emb_d is a subtraction of the first two. The fully connected layer takes them as input and predicts the binary label of each pair. The resulting binary label is assigned for each grid with a size of $0.093km^2$.

closely related to losses of properties and construction in the images.

3.1 Pre-training for general feature learning

Convolutional neural networks (CNN) for various tasks are often pre-trained with the ImageNet dataset (Deng et al., 2009). ImageNet-1000 is a large image classification dataset with over 1.2 million images with 1,000 classes. With such a large dataset, the model can learn mid-level visual features such as edges and corners and be used as a generic feature extractor (Oquab et al., 2014). We also started from the ImageNet-1000 pre-trained ResNet-18 model. The ResNet-18 network is a convolutional neural network with 18 layers (He et al., 2016). We wanted our model to get familiar with the bird's-eye viewpoint of satellite images before getting into the main task since it has never seen any satellite images. This approach is similar to that of Xie et al. (2016), which used a chain of transfer learning to train a model for poverty mapping.

We used a set of satellite image $\mathcal{I}$ from arbitrary regions and classified them into three classes according to their building density: more than 50%, under 50%, and 0%. We followed the method proposed by Han et al. (2020a), considering the limited size of the labeled dataset. Human annotators labeled a total of 1,000 randomly chosen images such that every image in $\mathcal{I}$ is labeled by five distinct annotators. For each image I_i^A, a label vector $y_i^A = [y_i^{high}, y_i^{low}, y_i^{zero}]$ ($0 \le y_i^{high}, y_i^{low}, y_i^{zero} \le 1$) is defined as the mean value of five one-hot building density response vectors from annotators. With this labeled pair set $L = \{(I_i^A, y_i^A)\}_{i=1}^{1000}$ and a larger

set of unlabeled image set $U = \{I_j^U\}_{j=1}^M$ with $M = 145921$, we pre-train ResNet-18 to conduct a three-class classification in a semi-supervised manner. The pre-trained network is used as an encoder in our damage classification model, which extracts the embeddings that contain semantic information about the images as high-dimensional vectors.

3.2 Damage classification

After pre-training the model to understand the general features of satellite imagery with large non-disaster images, we fine-tuned the model to detect damage after a disaster. Figure 1 describes the overall workflow of our model. The disaster image pair set $D = \{(I_i^{pre}, I_i^{post}, y_i)\}_{i=1}^N$ includes pairs with satellite images taken from the same pre-known damaged region before and after the disaster, and y_i, the binary label which indicates damage to the region. The structure of our model resembles the Siamese network, which has two branches with identical structures and parameters (Chicco, 2021). In our pseudo-siamese structure, however, the two identical branches do not share the same parameters, thereby giving the model greater flexibility (Zagoruyko and Komodakis, 2015). In addition, damage assessment differs from conventional change detection tasks because the target of interest in pre- and post-disaster images differs. The post-disaster encoder should focus on distinct features related to disasters, such as building wreckages, whereas the pre-disaster encoder should focus on ordinary buildings. Our model could effectively learn such distinctive characteristics using a pseudo-siamese structure with two separate encoders.

The pre-trained ResNet-18 in the previous step was fine-tuned as image encoders $e_{\theta_{pre}}$ and $e_{\theta_{post}}$ for pre-event and post-event image sets, respectively, to learn specialized features separately. Each encoder reduces the dimensions of given satellite imagery, transforming it into high-dimensional embeddings representing the area's vieual features. Then, we generate an embedding difference vector $emb_{di} \in \mathbb{R}^d$ and the final concatenated vector v_i for the i-th pair as follows:

$$emb_{di} = e_{\theta_{pre}}\left(I_i^{pre}\right) - e_{\theta_{post}}\left(I_i^{post}\right) \qquad (1)$$

$$v_i = Concat\left(e_{\theta_{pre}}\left(I_i^{pre}\right), emb_{di}, e_{\theta_{post}}\left(I_i^{post}\right)\right). \qquad (2)$$

We used the difference of embedding in the final concatenated vector to consider the change caused by the disaster in imagery. By feeding the embedding difference to the classifier, the model can learn the relationship between the features more effectively.

Finally, a damage classifier is trained to minimize the loss $\mathcal{L}$ defined between y_i and predicted value as follows:

$$\mathcal{L} = \frac{1}{|D|} \sum_{I_i^{pre}, I_i^{post}, y_i \in D} H\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right), \qquad (3)$$

where H is a binary cross entropy loss function. The damage classifier derives the predicted value $\hat{\mathbf{y}}_i = W \cdot \mathbf{v}_i$, where $W \in \mathbb{R}^{d \times \dim(y_i)}$ is a trainable weight matrix of the classifier.

3.3 Evaluation setting

We evaluated the model's performance by its classification accuracy to see whether the proposed model predicted the damage well. Furthermore, we checked the precision, recall, and F1-score considering the non-uniform distribution of class labels in our dataset. Precision and recall are calculated based on the number of true positives, false positives, and false negatives the model produces,

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \qquad (4)$$

where true and false indicate whether the prediction made by the model matches the ground truth label, and positive and negative refers to the predictions made by the model. F1 score is the harmonic mean of precision and recall,

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}, \qquad (5)$$

which varies from 0 to 1. A higher F1 score can be acquired only when precision and recall are both high. These metrics are frequently used when the class labels of the dataset are imbalanced.

3.4 Training details

We used the Ranger optimizer (Wright and Demeure, 2021) with a learning rate of 1e-4 and the CNN-based network ResNet-18 (He et al., 2016) that was pre-trained with the ImageNet dataset. All satellite imagery in a pair unit (pre- and post-disaster images) was rotated at random for augmentation. The embedding size of the ResNet-18 encoder is 256, and we trained the model for 100 epochs. Figure 2 shows the convergence of our model around 100 epochs.

4 Data

4.1 Satellite imagery dataset

We use the Zoom Level coordinate system to define satellite images' size, resolutions, and alignment. The purpose of using the system is to maintain consistency with other studies that combine satellite images and machine learning techniques (Jean et al., 2016; Han et al., 2020b). The system is a tile-based coordinate system that divides the entire world into non-overlapping square-shaped images. At z of 0, the entire world map is fitted to a single image tile and an increase of the zoom level by one results in half-sized image tiles. Thus, at z of 1, the world map is divided into 2×2 image tiles and hence has four times the resolution compared to $z = 0$. A higher zoom level divides the world map into more tiles, and each tile will cover a smaller geospatial area at a higher resolution. Economic development in regions is typically examined at a zoom level of 15 or more because one can start identifying building structures that are important for measuring population density at this zoom level. To clearly examine building outlines and smaller objects like vehicles, z of 17 or higher is typically used. However, processing images with higher zoom levels requires greater computational power and has limited regional coverage. Although they have a high spatial resolution, their temporal resolution is inevitably small as satellites with high zoom level cameras cover relatively narrow areas. Such low temporal resolution makes it harder to frequently track the changes in rural areas and countries that are still developing.

The zoom levels for input images of machine learning models are selected considering the task and data availability. In the case of damage detection models, the target area is small and requires detailed information. Thus, the existing damage detection models use relatively high spatial resolution images (z of 16–19) compared to the other models, such as poverty mapping or land classification models. This paper also employs satellite images at a zoom level of 17 (tile size of $0.093km^2$, 1.193m/pixel), allowing the model to consider the wreckage of buildings and roads. The RGB spectral bands are present in all of the images utilized in this research. For model pre-training, images of arbitrary regions were employed. We collected 146,921 satellite images spanning 2017 and 2018 from the ArcGIS World Imagery Wayback resource[2] for pre-training.
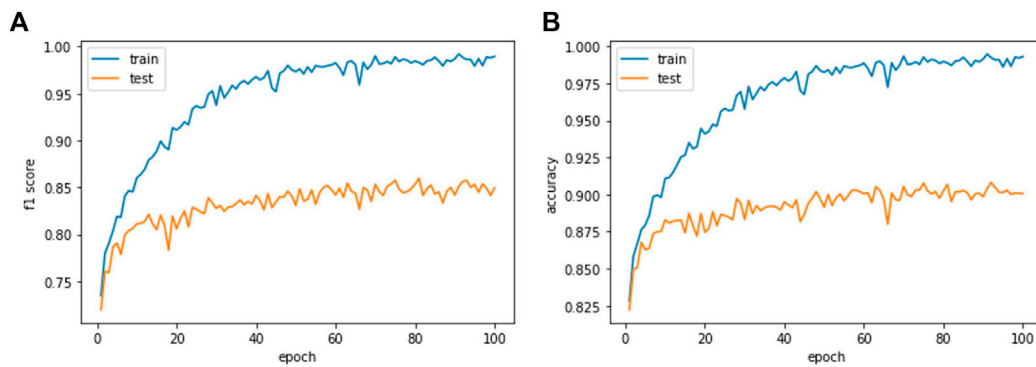
---

2   https://livingatlas.arcgis.com/wayback/

FIGURE 2
Learning curve of our model. **(A)** shows learning curve with F1 score. **(B)** shows learning curve with accuracy. The performance converges around 100 epoch.
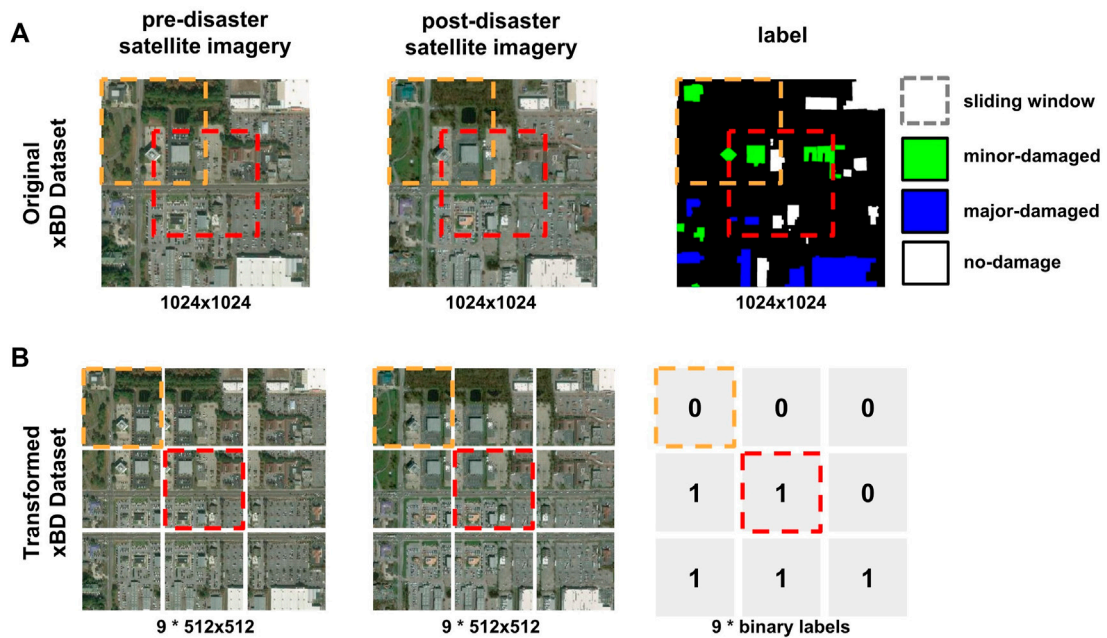

FIGURE 3
Method for simplifying the initial xBD dataset. The sliding window method crops each data sample **(A)** into nine different samples in **(B)**. The window size is 512×512, and the sliding unit is 256. Since the windows are half-overlapped, there are three windows on each side, resulting in 3×3 windows. The orange and red boxes show examples of our windows. **(A)** shows the satellite imagery in the original xBD dataset before and after a disaster at zoom level 16, with 1024×1024 pixels. The label has five classes: no building (black), no damage (white), minor damage (green), major damage (blue), and destroyed (red). **(B)** shows the result after cropping. Each image tile is now in zoom level of 17 with 512×512 pixels. Finally, to match the input size of the ResNet-18 encoder, we resized each image to 224×224 pixels. The cropped image is labeled 1 if it includes pixels of major damage or destruction; otherwise, it is labeled 0.

4.2 Disaster dataset

The model was trained with the xBD dataset (Gupta et al., 2019), which is the largest building damage assessment

dataset. The dataset includes pre- and post-images of various natural disasters, along with building annotations and damage scale labels. Damage labels span five types; four are related to damage scales of buildings (i.e., no-
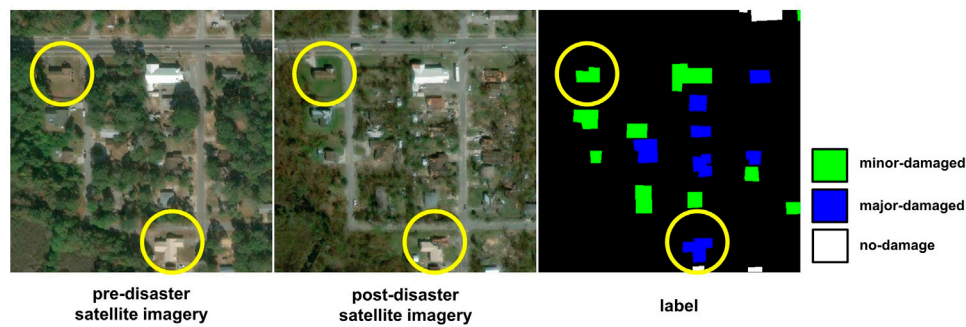
FIGURE 4
Examples of damage labels extracted from the xDB database. The pre- and post-disaster images of minor damaged structures are nearly identical, yet they show substantial visual differences for major damaged buildings.

damage, minor-damage, major-damage, and destroyed), and one is non-buildings. The dataset covers 22 natural disaster events of seven different categories. In this study, we targeted water-related disasters, including hurricanes, tornadoes, tsunamis, and floods. By focusing on them, our model can better learn the characteristics of water-related disasters, which are quite different from other disasters, like geological events. We chose water-related disasters, as their occurrences and impacts have been observed to grow among the most vulnerable countries to climate change.

The satellite imagery in the xBD collection has a zoom level of 16, each image tile covering $0.373 km^2$. We cropped each image tile into four half-sized image tiles. The resulting image tiles cover $0.093 km^2$ each, which fits the zoom level of our interest, $z = 17$. We also employed the sliding window method to overcome the limited training data size. With a window size of 512 and a sliding unit of 256, we could extract nine data samples per one xBD data sample. The method is depicted in Figure 3. After the cropping process, the images were re-scaled to 224×224 pixels to match the input size of the pre-trained ResNet-18 encoder.

Since our target is classification at an image level, the xBD dataset with labels at the pixel level cannot be directly applied to our model. Reducing the complexity of the data also brings positive effects such as minimizing noises in the original data, including the mismatch of building boundaries between the pre- and post-disaster images and the uneven distribution of damage class labels. We used a simple method to aggregate the information from each building polygon to derive the binary disaster label for each image. The image is classified as damaged if the maximum damage level of buildings in the image is greater than or equal to major-damaged. For example, Figure 4 is labeled as damaged since the maximum damage level in the image is major-damaged. However, if the maximum damage level in the image is less than or equal to minor-damaged, it is considered undamaged. In the case of minor-damaged buildings in the xBD dataset,

there is no discernible visual difference in the pre- and post-disaster images (Figure 4). In contrast, major-damaged buildings have a significant difference. Considering the characteristics of xBD labeling, we treated the images with only minor-damaged buildings as undamaged to reduce label noise and prevent the model from confusing small changes and the features of the damaged building.

After applying this rule, we acquired 12,241 damaged pairs and 25,109 undamaged pairs of satellite images. The data is summarized in Table 1. Among the total of 37,350 images, 11,018 did not include any buildings. The number of images, including major-damage, minor-damage, and destroyed, is 10,209, 10,046, and 6,413, respectively. The number of images with maximum damage levels of minor-damage, major-damage, and destroyed was 3,923, 5,828, and 6,413, respectively.

5 Results

5.1 Comparison with baseline models

The xView2 challenge[3] held in 2019 used the xBD dataset as a benchmark, and many models were proposed for damage assessment. However, our model cannot be compared to these models because the tasks and evaluation metrics are different. The xView2 challenge is defined at the building level, whereas our model detects damage at the grid level.

Before the xBD dataset was created, Xu et al. (2019) created a binary classification model which uses satellite imagery of pre- and post-disaster to detect the damage level. They built their own dataset covering only three disasters to train the model. Their model comes in four

---

3   https://xview2.org/

TABLE 1 Statistics of dataset with 11 events.

Disaster type	Event Name	Event Dates	Positive pairs	Negative pairs	All
Hurricane	Matthew	Oct 2018	1,996	1,649	3,645
	Harvey	Sep 2018	1,951	2,747	4,689
	Florence	Sep 2018	1,151	3,763	4,914
	Michael	Oct 2018	2,638	2,312	4,950
Tornado	Joplin	May 2011	595	746	1,341
	Moore	May 2013	324	1,719	2,043
	Tuscaloosa–Birmingham	Apr 2011	487	2,600	3,087
Flooding	Midwestern U.S.	Jan-May 2019	378	3,627	4,005
	South Asian	Jul-September 2017	1,861	3,710	5,571
Tsunami	Sunda Strait	Dec 2018	91	1,241	1,331
	Sulawesi	Sep 2018	769	995	1,764
Water-related	**All**	**2011–2019**	**12,241**	**25,109**	**37,350**

The bold text shows the statistics of all events.

TABLE 2 Performance comparison with baselines.

	Precision	Recall	F1 Score	Accuracy
CC	0.762	0.641	0.696	0.817
PO	0.716	0.621	0.663	0.794
TTC	0.741	0.669	0.703	0.815
TTS	0.759	0.648	0.697	0.816
Ours	**0.796**	**0.767**	**0.781**	**0.859**

The bold text indicates the best performance in each metric.

different versions: Channel Concatenate (CC), Post-disaster Only (PO), Twin-tower Concatenate (TTC), and Twin-tower Subtract (TTS). To the best of our knowledge, this work is the only research that has tried to detect the damage at a grid level. Therefore, we compared their models as a baseline.

The pre- and post-images are concatenated in the CC model before being fed to the AlexNet (Krizhevsky et al., 2012). Only post-disaster satellite images were used as input in the PO model. TTC and TTS models input both images to AlexNet's first convolution layer to extract the activation map at the lower level. TTC employs the concatenated activation map of two images as an input for the remaining convolution layers, while TTS uses the subtraction of the two activation maps. We implemented the model and trained it with our simplified xBD dataset.

Table 2 shows the baselines' worst and best performance compared to our model. The experiments were conducted using five different train and test splits. The bold text indicates the best performance. Both twin-tower baseline models, TTC and TTS, showed better performance than the single-tower ones, CC and PO. This result indicates that the model can capture damage better when pre- and post-event images are embedded separately. TTC and TTS showed similar performances, but TTC detected the damage slightly better since concatenation tends to lose less information in

training than subtraction. Compared to all of the four baseline models, the result from our model showed higher performance in all four evaluation metrics. In particular, the large gap in recall showed our model's ability to produce fewer false negatives on the damaged region. The proposed model utilizes embeddings of both images with a full encoder, which better preserves the information of the original image. We also took advantage of the benefits of TTC and TTS by putting together embeddings of pre- and post-images and subtracting them.

5.2 Ablation study

To check the role of each component, we conducted an ablation study where we removed each component from the model and evaluated the model performance. In this manner, we can check which components contribute the most and which may be removed.

- Without emb_d. To understand the impact of the embedding difference vector, emb_d, on model performance, we removed the emb_d and used only the two embedding vectors.
- Without sep. The encoder for pre and post-images is trained separately in the full model and does not share the parameters. We used the identical network for both images to test whether separating the two networks helps the model performance.
- Without emb_d, sep. We removed emb_d and shared one network for pre and post-images. This model can measure how the two components affect each other.
- Without pre-training. We trained the model from the random initialized state to understand how the model gets general information from the building density classification task.

TABLE 3 Precision, recall, F1 score, and accuracy of the ablation models.

Ablation Model	Precision	Recall	F1 Score	Accuracy
Without emb_d	0.788	0.765	0.776	0.855
Without sep	0.797	0.752	0.774	0.856
Without emb_d, sep	0.784	0.761	0.771	0.852
Without pre-training	**0.807**	0.755	0.780	**0.860**
Full model	0.796	**0.767**	**0.781**	0.859

The bold text indicates the best performance in each metric.

Table 3 compares the performance of each model. The experiments were conducted using five different train and test splits. The full model outperformed the others, demonstrating the value of each component. Interestingly, without pre-training model had the best overall performance among the ablation models, followed by without emb_d model. While the without emb_d model shows relatively high performance, the without emb_d, sep showed the poorest performance, demonstrating the need for independent encoders. This contradicts the findings of Weber and Kané (2020), which claimed that the model performs better when one network is shared for pre and post-disaster images.

The cause is likely due to the difference in task objective—classification vs object identification. The network sees the doubled training data when two images are shared, which can help with detailed object detection learning. Our model, on the other hand, is a classification model that considers the entire context rather than just a few data points. Separately focusing on the before and after contexts can help the model perform better. Also, the without sep model has the second best precision with the poorest recall among all ablation models. When the pre- and post-embeddings are extracted using the same encoder, the model is more likely to report observations as undamaged more frequently, resulting in higher precision with poor recall.

5.3 Qualitative analysis

We conducted a qualitative analysis to better understand the model's ability to identify the features of natural disaster damage. Figure 5 shows the success and failure cases and the ground truth labels from the xBD dataset. Our model successfully detected distinctive aspects of damage such as building ruins and missing trees in post-disaster images in the damaged pair of success cases in Figure 5A. Even though the buildings in the pre and post-images are not visually the same due to the shades, the model effectively distinguished the undamaged zone in the undamaged pair.

The failed cases in Figure 5B demonstrate the limitations related to the satellite image data. Our model classified the

damaged pair of failure cases as undamaged based on the status of nearly 50 buildings. However, this image also contained one building in the right-most top corner that was majorly damaged (pointed by the red arrow). In the bottom pair of failure cases, most of the farmland has submerged while the buildings appear to be unaffected. The ground truth label for the pair is undamaged because the xBD dataset only considers buildings. However, our model learned the features of the submerged area along with the damaged building during training and reported damage on that pair.

5.4 Out of domain setting

If a model is overfitted to the training dataset, it will not perform well on unknown samples and cannot be used in new disasters. We ran a cross-event test to check if the model could perform well on those unseen events (i.e., out of domain setting). Also, to see the effect of the mixture of multiple disaster types on the classification performance, we compared the model trained in a cross-event setting with all other events to the model trained with the same disaster type only.

Table 4 shows the model's prediction performance for four target events with two different cross-event settings. We choose the target event as the disaster with the smallest number of data samples in each disaster type. Due to the reduced size of the training data, the model was trained only for 50 epochs. We used the F1 score as the primary evaluation metric, considering the different label ratios of each event.

When the model was tested on the out-of-domain setting for each target, the performance was lower than the in-domain setting (random train:test = 8:2 split for the entire dataset) except for the Joplin tornado. Notably, the Midwestern U.S. Flooding and the Sunda Strait Tsunami showed an F1 score below 0.5. The reason for such low performance was the submerged buildings. The buildings were not visible in the post-disaster images, confusing our model. Also, the quality of post-disaster satellite imagery was low, with a high ratio of cloud coverage. The event had 91 positive image pairs, and the model predicted disaster for only 26 of them.

The model performed better when trained on all other events for three disasters, Hurricane Matthew, the Joplin tornado, and Midwestern U.S. flooding. The F1 score of the Joplin tornado was even higher than the in-domain setting when trained with all other events. The performance gap between the two different cross-event settings, all other events and only the events of the same disaster type, was the largest in hurricane Matthew, even though the amount of hurricane data was the largest. We speculate that the dataset's diversity led to better performance in all other event settings,
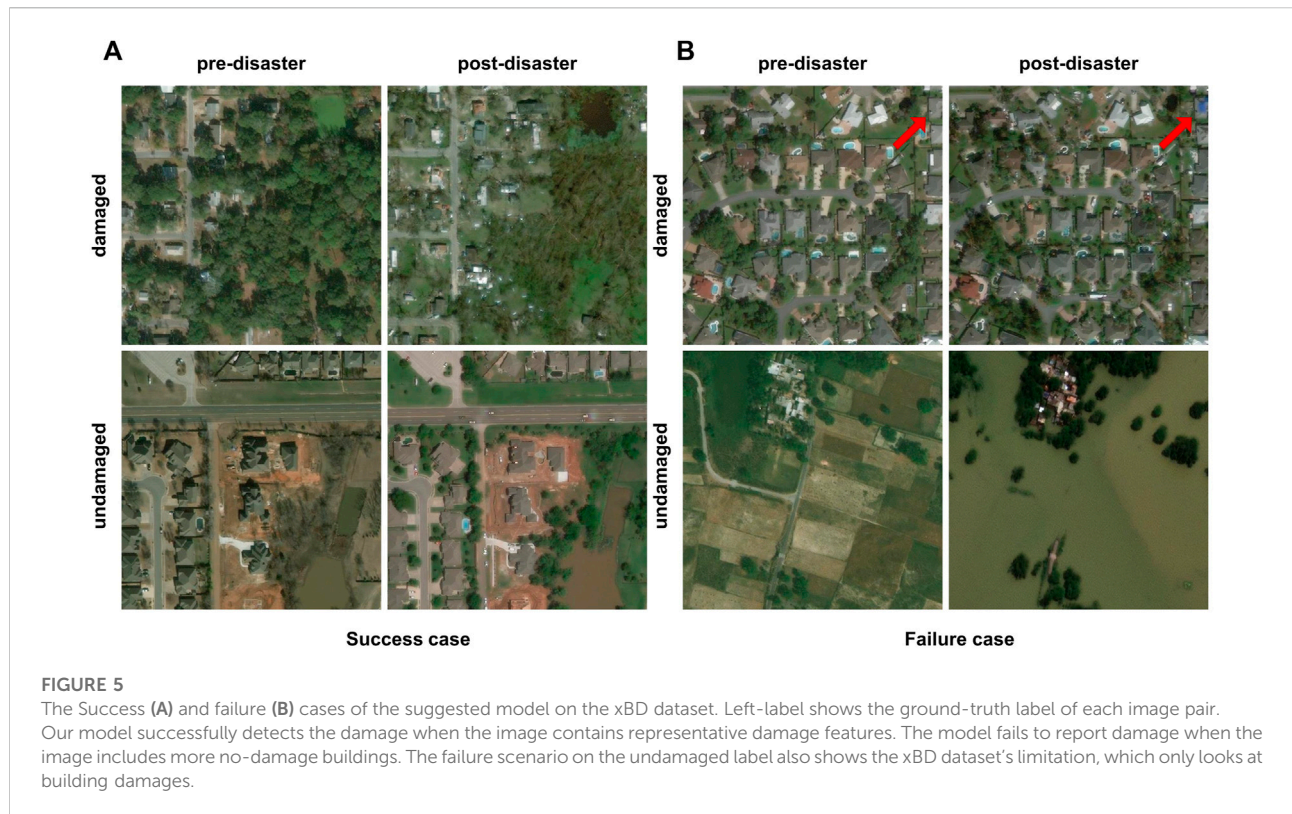
FIGURE 5
The Success **(A)** and failure **(B)** cases of the suggested model on the xBD dataset. Left-label shows the ground-truth label of each image pair. Our model successfully detects the damage when the image contains representative damage features. The model fails to report damage when the image includes more no-damage buildings. The failure scenario on the undamaged label also shows the xBD dataset's limitation, which only looks at building damages.

TABLE 4 Performance of our model in the out-of-domain setting.

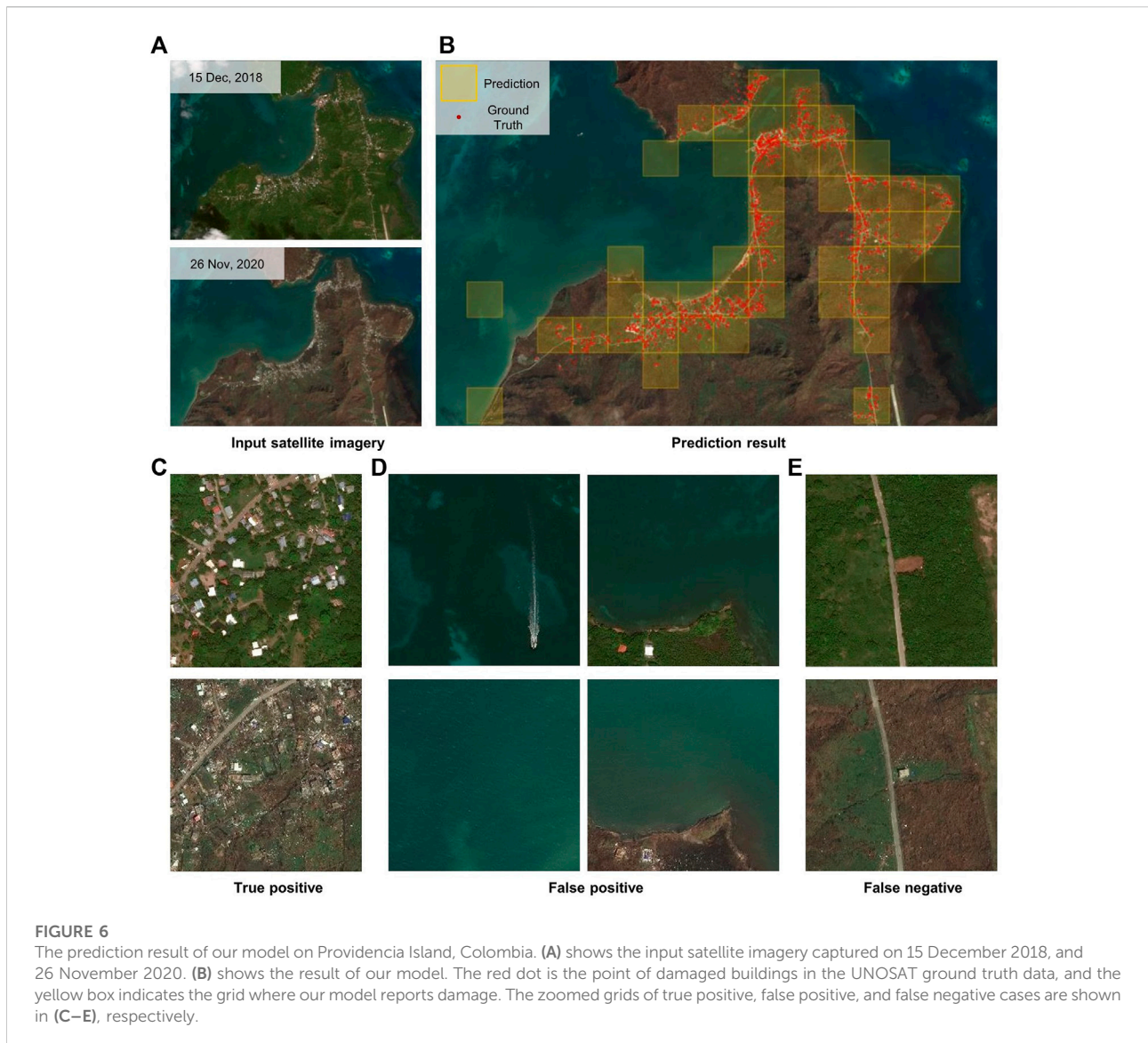| Target event	Training data	Precision	Recall	F1 Score	Accuracy
Hurricane Matthew	All events	0.820	0.513	0.631	0.672
	Hurricanes only	0.690	0.241	0.358	0.525
Joplin Tornado	All events	0.918	0.886	0.902	0.915
	Tornadoes only	0.847	0.854	0.850	0.867
Midwestern U.S. Flooding	All events	0.587	0.429	0.495	0.918
	Flooding only	0.234	0.492	0.317	0.800
Sunda Strait Tsunami	All events	0.115	0.286	0.164	0.801
	Tsunami only	0.459	0.308	0.368	0.928

preventing the model from becoming too specific and allowing it to learn general features.

5.5 Case study on Hurricane Iota

Having seen the potential for using computer vision techniques to assess disaster damage, we now introduce one case study of Hurricane Iota, which hit Colombia's Providencia Island in November 2020. The xBD dataset does not include Hurricane Iota. Therefore, it is valuable to test our detection model for the northern part of the island to demonstrate the applicability of the model. According to IFRC (2021), a substantial proportion of the

island's infrastructure, an estimated 98%, was destroyed, and 95% of its population was affected. We compared our model's prediction result to the ground truth label generated by UNITAR (2020).

Figure 6 shows the input images and the prediction results. We utilized Maxar SecureWatch to access pre- and post-disaster satellite imagery of the target area. The pre-disaster image was taken on 15 December 2018, and the post-disaster image was taken on 26 November 2020. The disaster occurred in November 2020, yet we could only find a cloud-free pre-disaster image of the region from December 2018, as shown in Figure 5A. Both images contain RGB spectral bands, and we processed the images to match the specifications for the model training dataset described in Section 4.1. We cropped the images into tiles that do not

FIGURE 6
The prediction result of our model on Providencia Island, Colombia. **(A)** shows the input satellite imagery captured on 15 December 2018, and 26 November 2020. **(B)** shows the result of our model. The red dot is the point of damaged buildings in the UNOSAT ground truth data, and the yellow box indicates the grid where our model reports damage. The zoomed grids of true positive, false positive, and false negative cases are shown in **(C−E)**, respectively.

overlap, each covering $0.093 km^2$ (zoom level of 17). The model successfully detected damage in the disaster-affected area as shown in Figure 6B, which shows ground truth labels (red dots) and damage prediction (shaded tiles). Even though the two images have different color compositions, the overlap between ground truth labels and predictions indicates that the model successfully distinguished the characteristics of damaged buildings from other changes (with the 2020 image missing green space). Our model had a 97.5% accuracy and an F1 score of 0.851, with 43 true positives, six false positives, and nine false negatives out of 595 grids.

An example pre- and post-image pair of a damaged region is shown in Figure 6C. We also find that some regions were falsely incorrectly labeled as damaged when they were not affected or vice versa (i.e., a false positive) and that some

damage was overlooked (i.e., a false negative). The left image pair in Figure 6D shows a false positive case due to a modeling error. The majority of the training data comes from inland images. Hence, a passing ship captured in the pre-disaster image served as a noise source. Filtering out the sea using shoreline data could be one way to avoid faulty detection caused by moving objects. The right image pair in Figure 6D shows another false positive case due to erroneous human labels. The two buildings in the image appear to be damaged, but the ground truth labels were missing in the region due to a lack of human resources. Our model correctly identified damage in the area, demonstrating its ability to recognize damaged structures.

The image pair in Figure 6E shows a false negative case due to limited temporal resolution. The only satellite imagery available in

the area prior to the disaster was from 15 December 2018, nearly 2 years before the disaster. During the period the two satellite images were taken, a new building was being built that was destroyed by the hurricane. Given that there is no visible building structure in the pre-disaster image, the model failed to detect damage in the area, resulting in a false negative case. This limitation in temporal resolution, however, is likely to be resolved thanks to the increasing availability of small satellites.

6 Conclusion

This study demonstrated how computer vision techniques can be used to develop data-driven disaster response strategies. Our lightweight model could successfully identify the damaged areas from water-related disasters with only pre- and post-satellite images and simple damage labels. Since our model is less constrained by the need for highly detailed input and label data, it can help disaster response efforts in many settings where the previous deep-learning-based detection models were not successful. For example, a development agency could prioritize their resources based on the sum of our binary labels and locate where to deploy the tents and shelters within their administrative with precision. This timely and accurate way of estimating damages shows how our model could help promote the adaptive capacities of many vulnerable countries.

Experimental results demonstrated that our model outperformed existing baselines in detecting water-related damage, achieving high accuracy of 91.4%. Our model successfully identified the damaged areas even with sparse damage labels. Evaluation of our model through an out-of-domain setting and case study demonstrated the model's robustness. This was especially apparent from the case study result which suggested that our model is applicable to real-world responses with a degree of high accuracy. The ablation study confirmed that the unique embedding method applied to pre- and post-disaster images via two separate encoders was critical to the performance. Moreover, pre-training our model with non-disaster satellite imagery before learning disaster-specific features was another critical, but far more challenging, factor to the model's success.

Future work may improve our study in several aspects. First, integrating other socioeconomic indicators into the model can help find more socioeconomically vulnerable regions. For example, if two areas with the same magnitude of destruction have starkly different socioeconomic vulnerabilities, the model should indicate more severe damage in the more vulnerable area. Secondly, our approach can be further expanded to identify damage from man-made disasters. Since these disasters typically affect more densely populated civilian buildings compared to natural disasters, future work could focus on producing more detailed damage estimates. Lastly, future innovations could work on integrating increasingly diverse satellite sources to produce more robust damage estimates. This approach would best support responses to disasters where it is often impractical to rely on one particular type of satellite imagery to make an effective damage assessment.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

All authors contributed to the conception and design of the study. JW initiated the idea. DK and EL preprocessed the data. DK implemented the model and ran experiments. EL implemented the baseline model. DK, JW, and EL wrote the first draft of the manuscript. (DK: related works, data, and experiments, JW: introduction and discussion, EL: methods). KP, JK, SP, HY, and MC gave critical comments on the direction of the study and revised the manuscript.

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amit, S. N. K. B., and Aoki, Y. (2017). "Disaster detection from aerial imagery with convolutional neural network," in *2017 international electronics symposium on knowledge creation and intelligent computing (IES-KCIC)* (IEEE), 239–245.

Amundsen, H., Berglund, F., and Westskog, H. (2010). Overcoming barriers to climate change adaptation—a question of multilevel governance? *Environ. Plann. C. Gov. Policy* 28, 276–289. doi:10.1068/c0941

Bai, Y., Hu, J., Su, J., Liu, X., Liu, H., He, X., et al. (2020). Pyramid pooling module-based semi-siamese network: A benchmark model for assessing building damage from xbd satellite imagery datasets. *Remote Sens.* 12, 4055. doi:10.3390/rs12244055

Bakkensen, L. A., Shi, X., and Zurita, B. D. (2018). The impact of disaster data on estimating damage determinants and climate costs. *Econ. Disaster. Clim. Chang.* 2, 49–71. doi:10.1007/s41885-017-0018-x

Cao, Q. D., and Choe, Y. (2020). Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. *Nat. Hazards (Dordr).* 103, 3357–3376. doi:10.1007/s11069-020-04133-2

Celik, T. (2009). Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering. *IEEE geoscience remote Sens. Lett.* 6, 772–776.

Chicco, D. (2021). *Siamese neural networks: An overview.* New York, NY: Springer US, 73–94. doi:10.1007/978-1-0716-0826-5_3

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

Duarte, D., Nex, F., Kerle, N., and Vosselman, G. (2018). "Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach," in *ISPRS annals of photogrammetry, remote sensing & spatial information sciences*, 4.

Edmonds, D. A., Caldwell, R. L., Brondizio, E. S., and Siani, S. M. (2020). Coastal flooding will disproportionately impact people on river deltas. *Nat. Commun.* 11, 4741–4748. doi:10.1038/s41467-020-18531-4

Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q. (2012). *Managing the risks of extreme events and disasters to advance climate change adaptation: Special report of the intergovernmental panel on climate change.* Cambridge University Press.

Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., and Nakamura, R. (2017). "Damage detection from aerial images via convolutional neural networks," in *2017 Fifteenth IAPR international conference on machine vision applications (MVA)* (IEEE), 5–8.

Ghanbari, M., and Akbari, V. (2015). "Generalized minimum-error thresholding for unsupervised change detection from multilook polarimetric sar data," in *2015 IEEE international geoscience and remote sensing symposium (IGARSS)* (IEEE), 1853–1856.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., et al. (2019). "Creating xbd: A dataset for assessing building damage from satellite imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 10–17.

Gupta, R., and Shah, M. (2021). "Rescuenet: Joint building segmentation and damage assessment from satellite imagery," in *2020 25th international conference on pattern recognition (ICPR)* (IEEE), 4405–4411.

Guiteras, R., Jina, A., and Mobarak, M. A. (2015). Satellites, self-reports, and submersion: Exposure to floods in Bangladesh. *Am. Econ. Rev.* 105 (5), 232–236. doi:10.1257/aer.p20151095

Hallegatte, S., Green, C., Nicholls, R. J., and Corfee-Morlot, J. (2013). Future flood losses in major coastal cities. *Nat. Clim. Chang.* 3, 802–806. doi:10.1038/nclimate1979

Han, S., Ahn, D., Cha, H., Yang, J., Park, S., and Cha, M. (2020a). Lightweight and robust representation of economic scales from satellite imagery. *Proc. AAAI Conf. Artif. Intell.* 34, 428–436. doi:10.1609/aaai.v34i01.5379

Han, S., Ahn, D., Park, S., Yang, J., Lee, S., Kim, J., et al. (2020b). "Learning to score economic development from satellite imagery," in *Proceedings of the CM SIGKDD international conference on knowledge discovery & data mining (KDD)*, 2970–2979.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

Henkel, M., Eunjee, K., and Magontier, P. (2022). The unintended consequences of post-disaster policies for spatial sorting.

Hunter, L. M. (2005). Migration and environmental hazards. *Popul. Environ.* 26, 273–302. doi:10.1007/s11111-005-3343-x

IFRC (2021). *DREF final report Colombia - hurricane Iota (MDRCO017) (international federation of red cross and red crescent societies).*

IPCC (2022). *Climate change 2022: Impacts, adaptation, and vulnerability.* Cambridge University Press. In Press.

Janalipour, M., and Talei, M. (2017). Building change detection after earthquake using multi-criteria decision analysis based on extracted information from high spatial resolution satellite images. *Int. J. remote Sens.* 38, 82–99. doi:10.1080/01431161.2016.1259673

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353, 790–794. doi:10.1126/science.aaf7894

Khanbani, S., Mohammadzadeh, A., and Janalipour, M. (2021). A novel unsupervised change detection method from remotely sensed imagery based on an improved thresholding algorithm. *Appl. Geomat.* 13, 89–105. doi:10.1007/s12518-020-00323-6

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. neural Inf. Process. Syst.* 25.

Kusetogullari, H., Yavariabdi, A., and Celik, T. (2015). Unsupervised change detection in multitemporal multispectral satellite images using parallel particle swarm optimization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 8, 2151–2164. doi:10.1109/jstars.2015.2427274

Lin, Y., McDermott, T. K., and Michaels, G. (2021). *Cities and the sea level.*

Mehrotra, A., Singh, K. K., Nigam, M. J., and Pal, K. (2015). Detection of tsunami-induced changes using generalized improved fuzzy radial basis function neural network. *Nat. Hazards (Dordr).* 77, 367–381. doi:10.1007/s11069-015-1595-z

Moser, S. C., and Ekstrom, J. A. (2010). A framework to diagnose barriers to climate change adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 107, 22026–22031. doi:10.1073/pnas.1007887107

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Potnis, A. V., Shinde, R. C., Durbha, S. S., and Kurte, K. R. (2019). "Multi-class segmentation of urban floods from multispectral imagery using deep learning," in *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium* (IEEE), 9741–9744.

Raker, E. J. (2020). Natural hazards, disasters, and demographic change: The case of severe tornadoes in the United States, 1980–2010. *Demography* 57, 653–674. doi:10.1007/s13524-020-00862-y

Rentschler, J., Salhab, M., and Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nat. Commun.* 13, 3527. doi:10.1038/s41467-022-30727-4

Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *Int. J. remote Sens.* 10, 989–1003. doi:10.1080/01431168908903939

Smith, A., Bates, P. D., Wing, O., Sampson, C., Quinn, N., and Neal, J. (2019). New estimates of flood exposure in developing countries using high-resolution population data. *Nat. Commun.* 10, 1814–1817. doi:10.1038/s41467-019-09282-y

Tellman, B., Sullivan, J., Kuhn, C., Kettner, A., Doyle, C., Brakenridge, G., et al. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature* 596, 80–86. doi:10.1038/s41586-021-03695-w

UNITAR (2020). "Damage assessment in Providencia island," in *Providencia and santa catalina department united nations Institute for training and research.* Colombia. as of 21 November 2020.

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., and Vosselman, G. (2018). Disaster damage detection through synergistic use of deep learning and 3d point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning. *ISPRS J. photogrammetry remote Sens.* 140, 45–59. doi:10.1016/j.isprsjprs.2017.03.001

Weber, E., and Kané, H. (2020). *Building disaster damage assessment in satellite imagery with multi-temporal fusion.* arXiv preprint arXiv:2004.05525.

Wright, L., and Demeure, N. (2021). *Ranger21: A synergistic deep learning optimizer.* arXiv preprint arXiv:2106.13731.

Wu, C., Zhang, F., Xia, J., Xu, Y., Li, G., Xie, J., et al. (2021). Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets. *Remote Sens.* 13, 905. doi:10.3390/rs13050905

Xie, M., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2016). "Transfer learning from deep features for remote sensing and poverty mapping," in *Thirtieth AAAI conference on artificial intelligence*.

Xu, J. Z., Lu, W., Li, Z., Khaitan, P., and Zaytseva, V. (2019). *Building damage detection in satellite imagery using convolutional neural networks*. arXiv preprint arXiv:1910.06444.

Zagoruyko, S., and Komodakis, N. (2015). "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4353–4361.

Zhu, Z. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS J. Photogrammetry Remote Sens.* 130, 370–384. doi:10.1016/j.isprsjprs.2017.06.013

Zou, L., Li, M., Cao, S., Yue, F., Zhu, X., Li, Y., et al. (2022). Object-oriented unsupervised change detection based on neighborhood correlation images and k-means clustering for the multispectral and high spatial resolution images. *Can. J. Remote Sens.* 48, 441–451. doi:10.1080/07038992.2022.2056434