Automating tephra fall building damage assessment using deep learning

Eleanor Tennant [1], Susanna F. Jenkins [2], Victoria Miller [3], Richard Robertson [4], Bihan Wen [5], Sang-Ho Yun [2], Benoit Taisne [2]

[1] Earth Observatory of Singapore @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore, 639798
[2] Earth Observatory of Singapore and Asian School of the Environment, Nanyang Technological University, Singapore, 639798
[3] GNS Science, P.O. Box 30368, Lower Hutt, 5040, Aotearoa New Zealand
[4] The UWI Seismic Research Centre, Saint Augustine, Trinidad, and Tobago
[5] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798

Correspondence to: Eleanor Tennant (eleanorm001@e.ntu.edu.sg)

In the wake of a volcanic eruption, the rapid assessment of building damage is paramount for effective response and recovery planning. Uninhabited aerial vehicles, UAVs, offer a unique opportunity for assessing damage after a volcanic eruption, with the ability to collect on demand imagery safely and rapidly from multiple perspectives at high resolutions. In this work, we established a UAV-appropriate tephra fall building damage state framework and used it to label ~50,000 building bounding boxes around ~2,000 individual buildings in 2,811 optical images collected during surveys conducted after the 2021 eruption of La Soufrière volcano, St Vincent and the Grenadines. We used this labelled data to train convolutional neural networks (CNNs) for: 1) Building localisation (average precision = 0.728); 2) Damage classification into two levels of granularity: No damage vs Damage (F1 score = 0.809); and Moderate damage vs Major damage, (F1 score = 0.838) (1 is the maximum obtainable for both metrics). The trained models were incorporated into a pipeline along with all the necessary image processing steps to generate spatial data (a georeferenced vector with damage state attributes) for rapid tephra fall building damage mapping. Using our pipeline, we assessed tephra fall building damage for the town of Owia finding that 22% of buildings that received 50-90 mm of tephra accumulation experienced at least Moderate damage. The pipeline is expected to perform well across other volcanic islands in the Caribbean where building types are similar, though would benefit from additional testing. Through cross validation, we found that the UAV look angle had a minor effect on the performance of damage classification models, while for the building localisation model, the performance was affected by both the look angle and the size of the buildings in images. These observations were used to develop a set of recommendations for data collection during future UAV tephra fall building damage surveys. This is the first attempt to automate tephra fall building damage assessment solely using post-event data. We expect that incorporating additional training data from future eruptions will further refine our model and improve its

Deleted: optical images

Deleted: D

Deleted: all of

Deleted: shapefile

Deleted: Our

Deleted:

applicability worldwide. To facilitate continued development and collaboration all trained models and pipeline code can be downloaded from GitHub.

Deleted: A

Deleted: to facilitate collaboration and development

1 Introduction

Tephra fall produced by explosive volcanic eruptions can have detrimental effects on buildings, which in turn affects the ability for a community to recover and rehabilitate. These effects range from surface-level issues such as corrosion of metal roofs (e.g., Rabaul, Papua New Guinea, Blong, 2003a) or damage to non-structural components (e.g., gutters: Ambae, Vanuatu, Jenkins et al., 2024) through to complete building collapse (e.g., Pinatubo, Philippines, Spence et al, 1996).

Deleted: under review

Deleted: in press

After, or during, an eruption, the collection of empirical data detailing the damage incurred is critical to guiding the planning and implementation of response and recovery efforts. This involves estimation of damages and losses, which are needed to determine the necessary funding for repair or reconstruction; along with an assessment of building functionality, which can inform temporary housing requirements. In addition to its use in post disaster recovery, the collection of damage data are key to the development of vulnerability models (Deligne et al., 2022), which relate hazard intensity to damage (e.g., Spence et al., 2005; Wilson et al., 2014; Williams et al., 2020), and can be used to inform resilient construction practises and/or for pre-event impact assessments.

Deleted: fragility functions

Deleted: ;

Deleted: Spence et al., 2021)

Deleted:

Post-event building damage assessments usually consist of ground surveys, whereby the amount of damage to each building is described using a quantitative or qualitative damage state (e.g., Spence et al., 1996; Blong 2003a; Jenkins et al. 2013; Jenkins et al. 2015; Hayes et al. 2019; Meredith et al. 2022). However, tephra fall damage can extend tens or even hundreds of kilometres away from a volcano (Spence et al., 2005) meaning that comprehensive ground based damage assessments can be both time consuming and costly. Furthermore, the uncertainty that is often associated with the end of an eruption may prevent the safe completion of a ground-based damage assessment before tephra is remobilised by winds and rain. This lag between the event itself and the completion of a damage assessment, can hinder recovery efforts and compromise the accuracy of data collected for the development of forecasting models.

Deleted: ground based

Given the need for, but also the challenges associated with, conducting post-event building damage assessments quickly, approaches that use remotely sensed (RS) data, either optical or Synthetic Aperture Radar (SAR) imagery have been developed in volcanology (e.g., Jenkins et al. 2013; Williams et al. 2020; Lerner et al. 2021; Biass et al. 2021; Meredith et al. 2022), and operationally by emergency management services (e.g., International Charter "Space and Major disasters", Copernicus Emergency Management Service, ARIA: Advanced Rapid Imaging and Analysis system) (Yun et al., 2015)). The use of optical imagery largely consists of visual inspection, which may be influenced by image resolution and is prone to subjectivity (Novikov et al. 2018). Furthermore, visual inspection of satellite optical imagery can still be time consuming without crowd sourcing (e.g., Ghosh et al. 2011) and is constrained by satellite recurrence intervals and cloud cover. Automated SAR based methods (e.g., Yun et al., 2015) are not limited by cloud cover, but they may lack the resolution required for building level damage assessment (30 m for damage proxy maps generated from Sentinel data using the ARIA system; https://aria-share.jpl.nasa.gov/20210409-LaSoufriere_volcano).

Deleted:), and

To our knowledge, only one study attempts to automate the assessment of building damage from volcanic hazards (Wang et al., 2024). In contrast, attention has been given to more commonly occurring hazards such as earthquakes and hurricanes, with the development of both mono- temporal (post-event imagery only) and multi-temporal (images taken at different times) approaches (Table 1). Early approaches at automation with optical imagery used image processing methods, often focusing on identifying changes in pixel values between pre- and post-event imagery (e.g., Bruzzone and Fernàndez Prieto 2000; Ishii et al. 2002; Zhang et al. 2003). Image processing methods are susceptible to user biases such as the choice of thresholds that equate to distinct levels of damage severity, or damage states, and may require recalibration when applied to a new dataset. As a result, image processing methods were succeeded by the application of traditional machine learning algorithms that use 'handcrafted' image features. These features are observable properties that can be extracted from the image such as shape, colour, texture, and statistical properties of the image (e.g., Li et al. 2015; Anniballe et al. 2018; Lucks et al. 2019; Naito et al. 2020). The success of a given machine learning approach is dependent on the selection of the best features for the job; for example, a texture-based feature might be good for classifying buildings as damaged or not damaged due to an increased number of edges in damaged buildings but less useful for a task such as differentiating between building roof types where the difference in textures between the classes

Deleted: While efforts to automate the assessment of building damage from volcanic hazards are minimal (to our knowledge there has been one study focusing on building damage from volcanic eruptions: Wang et al., 2024), attention has been given to more commonly occurring hazards such as earthquakes and hurricanes, with the development of both mono-temporal (post-event imagery only) and multi-temporal (uses pre- and post-event imagery) approaches (Table 1).

Deleted: cats from dogs

is less significant. Deep learning, in particular the use of convolutional neural networks (CNNs), removes this need for feature selection. A CNN is a network of layers comprising filters which are small matrices of values. When an image is passed through the network, at each layer the filters are convolved with the output from the previous layer to create a new representation of the image that is progressively more abstract with depth in the network. This process reduces the image's original spatial dimensions (X and Y) while increasing the number of channels, facilitating classification. During network training the filter values (known as weights) are optimised to reduce the loss between the predicted label for the image and the true label. Through this training a CNN learns the features of the images that are useful for classification. For a detailed background on deep learning see Aggarwal, (2018).

Thus far, deep learning models have been developed for optical image sets for hurricanes (Li et al. 2019a; Dung Cao and Choe 2020; Pi et al. 2020; Cheng et al. 2021; Khajwal et al. 2023); earthquakes (Nex et al. 2019; Xu et al. 2019; Duarte et al. 2020; Moradi and Shah-Hosseini 2020); wildfires (Galanis et al. 2021); volcanic hazards (Wang et al., 2024); and models that have been proposed for multiple hazards (e.g., Gupta and Shah 2020; Weber and Kané 2020; Shen et al. 2021; Bouchard et al. 2022) (Table 1). However, building damage caused by different hazards looks very different (e.g., damage caused by vertical loading from volcanic tephra fall vs ground shaking from an earthquake). These observable differences mean that an optical imagery multi-hazard damage classification model that performs consistently well across the different hazards is not yet achievable. Therefore, distinct models tailored for specific hazards are required (Nex et al., 2019, Bouchard et al., 2022). It follows that models may also benefit from being regionalised, given the differences in building typologies (construction material and styles) that can also affect the observable damage (Nex et al., 2019).

Deleted: ;

Many of the approaches for automating building damage assessment use both pre- and post-event imagery (Table 1), which makes the task more straightforward since any changes to the pre-event imagery can be considered damage. However, pre-event imagery at a high-enough resolution is not always available in post-disaster scenarios. The automated assessment of building damage from volcanic hazards using only post-event optical imagery has not yet been achieved in part due to absence of the large datasets that are needed in order to train models. The 2021 eruption of La Soufrière volcano, St Vincent and the Grenadines, provided unprecedented circumstances allowing for the collection of high-resolution UAV imagery

Deleted: an

Deleted: opportunity

enabling the development of fully automated models that can assess tephra fall building damage from post-event data only. With their growing ubiquity and low cost, UAVs have become an increasingly useful tool during and after volcanic eruptions (e.g., Andaru and Rau 2019; Gailler et al. 2021; Román et al. 2022). UAVs offer a distinct advantage over satellite imagery because they can be scheduled at any point, they do not suffer from cloud obscuring the images as they fly at relatively low altitude, and they capture imagery from multiple perspectives, which may lead to increased ability to capture damage information. In this study we used UAV optical imagery collected after the 2021 eruption of La Soufrière volcano to develop a methodology for tephra fall building damage assessment; the main contributions of our work are three-fold:

1. We have devised a UAV appropriate building damage state framework, laying the foundation for future tephra fall UAV building damage surveys.
2. We have developed a deep learning pipeline that consists of all trained models and image processing steps to rapidly output spatial damage data that can facilitate prompt, post-event response and recovery, and enable data collection prior to further changes by natural or human processes (tephra clean-up).
3. Imagery used in this work is diverse in terms of the flight altitude, time of acquisition after the event, and UAV vantage point. We have conducted extensive testing to understand the best practises for building damage surveys and to create a series of recommendations for the collection of future UAV surveys for building damage assessment.

Deleted: building damage maps

Table 1. A non-exhaustive list of works using deep learning on optical imagery for building damage assessment. Studies use different scores to evaluate performance: F1 scores are in italics, mean average precision scores are underlined, accuracy scores in ***bold****. For all scores, 1 represents a perfect model. A detailed explanation of the scores used for evaluation is provided in Section 2.3.3.*

Study	Hazard	Number of damage classes	Pre-disaster imagery	Data type	Building localisation	Damage classification
Li et al. (2019a)	Hurricane	2	No	airborne	0.448	
Weber and Kane, (2020)	Multi	4	Yes	satellite (xBD)	*0.835*	*0.697*

Deleted: Pre and post?

Formatted Table

Deleted: P

Deleted: P & P

Dung Cao and Choe. (2020)	Hurricane	2	No	satellite	-	**0.972**
Pi et al. (2020)	Hurricane	2	No	UAV, airborne		0.745 (UAV) 0.807 (airborne)
Cheng et al. (2021)	Hurricane	5	No	UAV	0.656	**0.610**
Galanis et al. (2021)	Wildfire	2	No	satellite		*0.981*
Gupta and Shah (2020)	Multi	4	Yes	satellite (xBD)	*0.840*	*0.740*
Shen et al. (2021)	Multi	4	Yes	satellite (xBD)	*0.864*	*0.782*
Bouchard et al. (2022)	Multi	2	Yes	satellite (xBD)	*0.846*	*0.709*
Khajwal et al. (2023)	Hurricane	5	No	ground airborne	-	*0.650*
Singh and Hoskere, (2023)	Multi	5	No	satellite		**0.880**
Wang et al (2024)	Volcanic tephra	4	Yes	satellite	*0.868*	*0.783*

1.1 The 2020-2021 eruption of La Soufrière volcano St Vincent

La Soufrière St Vincent is an active stratovolcano standing at 1220 meters above sea level on the island of St Vincent. On 27th December 2020 a thermal anomaly was detected inside the summit crater by the NASA Fire Information for Resource Management System (FIRMS). This was confirmed by the Soufrière Monitoring Unit to be caused by a new dome growing within the crater. Dome growth continued for three months until 9 April 2021, when, following two days of heightened seismic activity and lava effusion rate, the ongoing effusive eruption of La Soufrière entered an explosive phase (Joseph et al. 2022). Between 9 – 22 April, a total of 32 distinct explosions occurred, with the tallest plumes reaching heights of up to 15 kilometres above the vent (Joseph et al. 2022). Throughout this explosive phase, tephra blanketed the island, resulting in a total deposit thickness of up to 16 centimetres in coastal communities to the north of the island (Cole et al. 2023) (Figure 1).

The explosive phase was anticipated, and an evacuation order was issued on 8 April 2021 for the ~16,000 residents in the northern part of the island (Joseph et al. 2022). As a result, there were no reported fatalities directly attributable to the eruption, nevertheless, the overall damage to infrastructure services and physical assets were estimated at XCD 416.07 million (equivalent to USD 153.29 million) (PDNA, 2022). Approximately 63% of this monetary impact was borne by the housing sector. In St. Vincent, residential buildings are typically single-story,

Deleted: P
Deleted: P
Deleted: P
Deleted: P
Formatted: Font: Cambria, 10 pt
Deleted: P & P
Deleted: P & P
Deleted: P & P
Deleted: P
Deleted: P
Deleted: P & P
Deleted: Our work ... [1]
Deleted: ion
Deleted:

detached structures, with the majority in the more impacted north of the island (census districts of Chateaubelair, Georgetown, and Sandy Bay: Figure 1) constructed using concrete and blocks (84% in Chateaubelair, 74% in Georgetown, 50% in Sandy Bay), with sheet metal roofs (90-92% of all buildings in these areas) (SVG population and housing census, 2012).

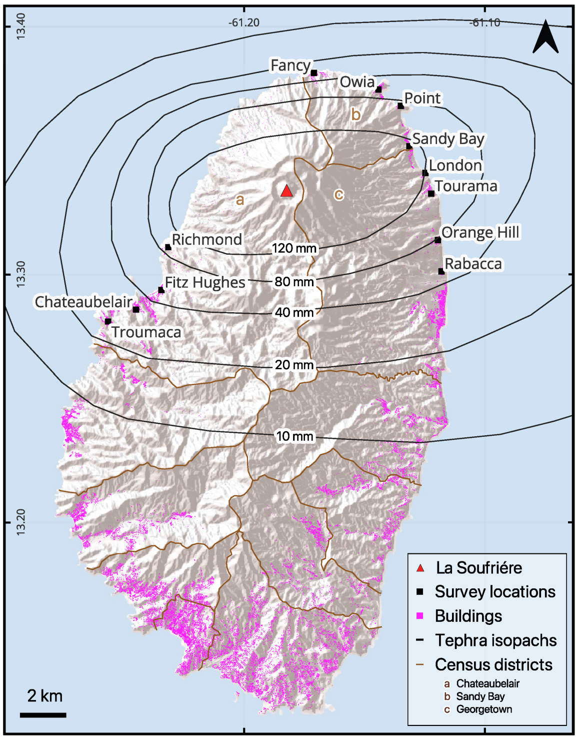
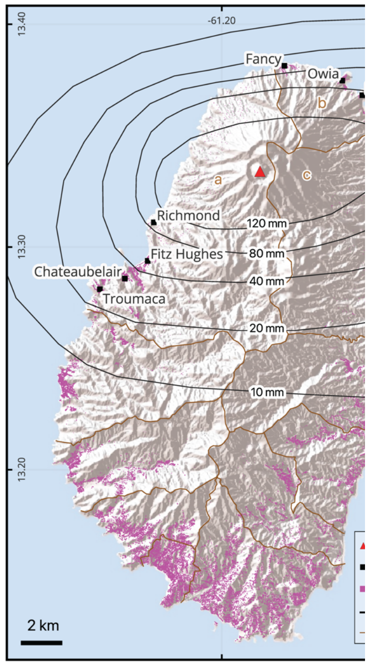


Figure 1. The island of St Vincent with UAV survey locations included in this work labelled and marked in black. Tephra isopachs (Cole et al., 2023) mark lines of constant total tephra thickness. Building footprints are marked in pink, data source: © OpenStreetMap contributors 2024. Distributed under the Open Data Commons Open Database License (ODbL) v1.0. Coordinate reference system: WGS 84 (EPSG:4326).

2 Method

After the 2021 eruption of La Soufrière three UAV optical imagery datasets were collected to assess the extent of the damage. These were collected by different parties at separate times after



Deleted:

Deleted: *Census districts referred to in the text are: a) Chateaubelair, b) Sandy Bay and c) Georgetown.*

the eruption. All UAV survey locations are shown in Figure 1, and representative examples of images can be found in Section S1 of the supplementary material.

Deleted: Figure 1

Formatted: Font: Not Italic

2.1 Dataset description

Dataset 1: April-May 2021 (UWI-TV)

Collected by UWI-TV at the request of The UWI Seismic Research Centre (SRC), this dataset consists of video footage for Chateaubelair, Fitz Hughes, Troumaca, and Sandy Bay acquired with a frame rate of 30 frames per second (fps) and a resolution of 1920 x 1080 pixels. Flight paths were not programmed, and the vantage point varies between at nadir (directly above buildings) and very off-nadir (showing the sides of buildings). Images do not contain GPS positioning or altitudes and were not manually georeferenced.

Deleted: .

Dataset 2: 12th – 14th May 2021 (GOV)

Collected by the Government of St Vincent and the Grenadines Ministry of Transport, Works, Lands and Surveys, and Physical Planning for the purpose of assessing the eruption impact. This dataset consists of video footage for Chateaubelair, London, Richmond and Sandy Bay acquired with a frame rate of 30 fps and a resolution of 1920 x 1080 pixels. Buildings are imaged at a nadir to off nadir vantage point with an altitude of ~ 200 m (above the ground). Buildings are lower resolution in this dataset when compared to the other two. Images contain GPS positioning and altitudes.

Dataset 3: August -September 2021 (SRC)

This is the most extensive dataset, collected by SRC for the purpose of assessing eruption impact. It consists of photos and videos for Belmont, Chateaubelair, Fancy, London (video only), Orange Hill (video only), Owia, Point, Rabacca (video only), Richmond, Sandy Bay, Tourama, Videos were acquired with a frame rate of 30 fps and have a resolution of 1920 x 1080 pixels, while photos are 4056 x 3040 pixels. Flight paths were programmed to follow a linear swath like trajectory. Buildings are captured from nadir between 55-290 m above the ground. Images contain GPS positioning and altitudes.

For all three datasets, image frames were extracted from the videos every two seconds, an interval chosen to reduce redundant homogeneous images, this resulted in a total of 7,956 image frames. Due to the UAV surveying approach (i.e., hovering in one place for a while) many near-identical images were generated. To avoid potentially biasing the training towards overrepresented buildings we manually filtered out duplicate images. After filtering, and the removal of images with no buildings present, the full combined dataset consisted of 2,811 image frames. We labelled all images by drawing bounding boxes around each building present and storing the bounding box positions. In total 49,173 building bounding boxes were drawn around ~2,000 individual buildings (with some buildings being present in multiple images). Given the absence of individual building location information, this number was approximated by overlaying Open Street Map building footprints with UAV GPS tracks where available. Bounding boxes were drawn by a team of five including the lead author, and all boxes were checked by the lead author. Each box was then assigned one of three damage states, which are described below. For consistency the damage states were assigned by the lead author. All labelling, modelling, and analysis were conducted using MATLAB 2023b.

2.2 Developing and applying a building damage state framework

The first tephra fall building damage state framework was developed after the eruption of Pinatubo, Philippines, 1991 (Spence et al., 1996), and was adapted from the macro seismic intensity scale used to evaluate seismic damage (Karnik et al., 1984). In the adapted framework damage ranges from DS0 – "no damage", through to DS5 – "complete roof collapse and severe damage to the rest of the building". Subsequent tephra fall building damage state frameworks were modified from the work of Spence et al., (1996) with changes in the wording made to reflect the characteristics of the case study (Table 2). In the damage state descriptions, damage to three critical aspects of a building is described: the roof covering, the roof structure, and the vertical structure (Blong 2003b; Hayes et al. 2019; Jenkins et al., 2024). In our study, most images depict buildings from an at nadir or close to nadir perspective making roof damage more discernible than damage to the vertical structure. Thus, we generated a damage state framework that is based on the proportion of observable damage to the roof, as in the work of Williams et al. (2020). Our final framework, which was developed over several iterations, classifies building damage into three classes: No observable damage to minor damage, Moderate damage, and Major damage (Table 3, Figure 2). Damage states are deliberately generic so that the range of possible damage to the range of different building types can be captured (Blong, 2003a). Our three classes are comparable to DS0-1, DS2, and DS3-5, respectively, of damage scales developed for ground surveys (Table 2). In the frameworks

Deleted: detailed building inventory information

Formatted: English (UK)

Deleted: Bounding boxes were drawn by a team of five including the lead author, and all boxes were checked by the lead author. smay. Nevertheless, this was not considered an issue since deep learning models for object localisation will quickly learn to ignore background pixels (Bouchard et al., 2022).

Deleted: modeling

Deleted: to describe damage from

Deleted: developed

Deleted: For tephra fall,

Deleted: D

Deleted: d

Deleted: consisted of the following classes:

Deleted: D1: "Light roof damage", D2: "Moderate roof damage", D3: "Severe roof damage and some damage to vertical structure", D4: "Partial roof collapse and moderate damage to rest of building",

Deleted: split damage into five damage states, plus one not damaged, based on

Deleted: d

Deleted: (Spence et al., 1996;

Deleted: under review

Deleted: in press

Deleted: Ground based damage state frameworks for tephra fall have previously split damage into five damage states, plus one not damaged, based on damage to three critical aspects of a building: the roof covering, the roof structure, and the vertical structure (Spence et al., 1996; Blong 2003b; Hayes et al. 2019; Jenkins et al., under review). Remote damage assessments are often less able to resolve the detailed resolution achievable on the ground, and so a coarser resolution damage state framework is needed....

Deleted: e or

Deleted: Classes are comparable to DS0-1, DS2, and DS3-5 of damage scales developed for ground surveys respectively (Table 2) . (

Deleted: (Table 2).

Deleted: damage state

presented in Table 2, DS1 describes light/minor damage or superficial damage to non-structural components. In our framework we included minor damage in the No damage class since the difference between the two can be subtle and not easily discernible through remote assessment. Furthermore, buildings with minor damage are typically habitable and unlikely to require costly repairs; therefore, from a response and recovery perspective, we considered them better grouped with undamaged buildings. Our Moderate damage class requires damage or collapse to up to 50% of the roof area, which closely fits with damage state 2 of Blong, (2003), Hayes et al., (2019) and Jenkins et al., (2024). The ground-based frameworks distinguish damage states 3 through 5 by increasing amounts of damage to the building walls (Table 2). However, the quantity and severity of impacted walls is not easy to differentiate in the majority of our UAV images, which show buildings from a nadir or close to nadir perspective. Therefore, in our framework, we grouped these states together under 'Major damage'.

Table 2. A comparison of tephra fall building damage state frameworks available to date.

	Pinatubo, Philippines, 1991 Spence et al., (1996)	Rabaul caldera, Papua New Guinea, 1994 Blong, (2003)	Calbuco, Chile, 2015 Hayes et al., (2019)	Manaro Vuoi, Ambae island, Vanuatu, 2017-2018 Jenkins et al., (2024)
DS0	***No damage***		***No damage***	***No damage***
DS1	***Light roof damage:*** - Gutter damage. - Few tiles dislodged.	***Light damage:*** - Damage to gutters and/or water tanks. - Cleanup required	***Minor damage to non-structural elements:*** - Damage to gutters. - Few tiles dislodged. - Damage to fittings, e.g. air-conditioning units and appliances. - Damage to contents. - Dents in the roof covering.	**Light damage or damage to non-structural elements:** - Damage to gutters. - Damage to contents. - Dents or minor slumping in roof cover.
DS2	***Moderate roof damage:*** - Bending or excessive deflection of roof sheeting or purlins. - No damage to principal roofing supports.	***Moderate damage:*** - Bending or excessive damage to as much as half roof sheeting and/or purlins. - Damage to roof overhangs or verandas. - Slight roof structural damage possible. - Interior requires cleaning, repainting,	***Moderate damage but vertical structure and roof supports intact:*** - As above. - Bending or excessive (e.g., perforation, cracking) damage (with or without collapse) to up to half of roof covering, e.g. tiles, metal sheet.	***Moderate damage but vertical structure and roof supports intact:*** - As for DS1, plus: - Bending or excessive damage (without collapse) to up to half of the roof covering. - Little or no damage to roof support trusses and rafters.

Deleted: very

Deleted: W

Deleted: discernable

Deleted: therefore

Deleted: (

Deleted: in press

Deleted: ¶

Deleted: ¶
In some images tarpaulins can be seen partially or fully covering roofs (~30 buildings). These were potentially placed to cover damage that occurred during the eruption, including corrosion due to prolonged presence of tephra on metal roofs or, holes generated by nails lifted out through sub-optimal cleaning approaches (VM personal communication). Alternatively, tarpaulins may have been placed as a preventative measure to help shed tephra (e.g., Ambae Vanuatu, Jenkins et al., under review2024). Erring on the conservative side, we considered buildings with a tarpaulin to be damaged; we assessed the severity of the damage for each building based on the level of visible deformation. We assigned buildings with a tarpaulin and no visible deformation to the moderately damaged class and those with a tarpaulin and visible deformation to the major damage class. ¶

Deleted: *3*

Formatted: Centered

Formatted Table

Formatted: Centered

Formatted: Centered

Formatted: Centered

		and/or overhaul of electrical systems. - Solar heater needs replacing.	- Little to no damage to principal roof supports, i.e. rafters or trusses. - Damage to roof overhangs or verandas.	- Damage to roof overhangs or verandas. - Interior requires repair.
DS3	***Severe roof damage and some damage to vertical structure:*** - Severe damage or partial collapse of roof overhangs or verandahs. - Severe deformation of main roof sheeting. - Some damage to roof supporting structure, columns, trusses.	***Heavy damage:*** - Damage to roof structure and some damage to walls. - At least one wall damaged/misaligned. - Collapse of part of ceiling	***Severe damage to the roof and supports:*** - As above. - Bending or excessive (e.g., perforation, cracking) damage (with or without collapse) to over half of roof covering. - Damage to any single principal roof supports and some damage to walls. - Severe damage or partial collapse of roof overhangs or verandas.	***Severe damage to the roof and supports:*** - As for DS2, plus: - Bending or excessive damage (with or without collapse) to more than half of the roof covering. - Damage to any single principal roof supports and/or some damage to walls (less than half of walls affected). - Severe damage or partial collapse of roof overhangs or verandas.
DS4	***Partial roof collapse and moderate damage to rest of building:*** - Collapse of sheeting but not truss. - Partial collapse of sheeting and some truss failure. - Failure of supporting structure. - Moderate damage to other parts of building resulting from roof collapse.	***Severe damage:*** - Roof collapse and moderate to severe damage to rest of the building. - Failure of roof trusses and supporting structure. - At least half of the external walls and/or internal walls deformed or collapsed. - For two-storey buildings, collapse of external and internal walls of upper floor. - Plumbing and other services may be damaged.	***Partial or total collapse of the roof and supports:*** - As above - Collapse of roof covering and any single principal roof support(s). - At least half of the external walls and/or internal walls deformed or collapsed.	***Partial collapse of the roof and supports:*** - As for DS3, plus: - Collapse to less than half of roof covering and principal roof support(s). - At least half of external and/or internal walls deformed or collapsed.
DS5	***Complete roof collapse and severe damage to the rest of the building:*** - Collapse of roof and supporting structure over more than 50 percent of roof area.	***Collapse:*** - Collapse of roof and supporting external walls over more than 50% of floor area of building. - Internal walls collapsed. - Damage to floor and/or foundation. - Structure is irreparable, not	***Building collapse:*** - As above. - Collapse of roof, principal roof supports and/or supporting external walls over >50% of floor area of building.	***Building collapse:*** - As for DS4, plus: - Collapse of roof, principal roof supports and/or supporting external walls over more than half of floor area of building.

- Partition walls destroyed. - External walls destabilized.	salvageable, beyond economic repair.

Table 3. The damage state framework developed for our UAV optical imagery dataset

Damage state	Description of the damage
No damage to minor damage	- No visible damage/or - Up to 10% of the roof covering missing; and/or - No roof or structural collapse; and/or. - Visible damage to non-structural elements e.g., gutters or decorative elements (fascia). - *Comparable to DS0-1 (Table 2).*
Moderate damage	- Up to 50% roof area damaged (evidence of bending) or collapsed; may include light damage to vertical structure (e.g. wooden slats above windows broken). - *Comparable to DS2 (Table 2).*
Major damage	- More than 50% roof area damaged or collapsed; may include damage to the vertical structure including total building collapse. - *Comparable to DS3-5 (Table 2).*

Deleted: ¶

Deleted:

Deleted: *2.*

Deleted: *assessment*

Formatted Table

Formatted: Centered

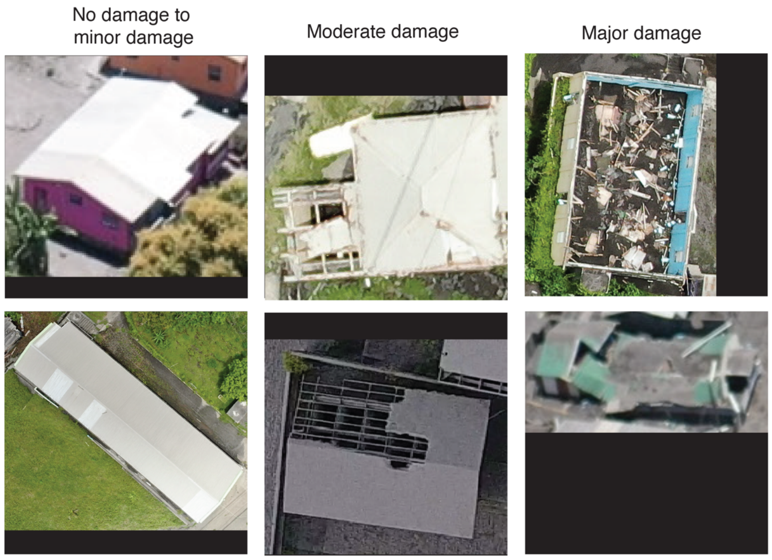


Figure 2. Example of the three damage states used in this work: No damage to minor damage, Moderate damage and, Major damage.

2.3 Model development

After labelling, we split the full combined image dataset (2,811 frames from the UWI-TV, GOV and SRC sets) into train/validation/test sets (Figure 3). Given that many images lacked GPS positions, we grouped images by location to ensure independence among the sets. The partitioning was chosen to include diversity in both the image sets (UWI-TV/GOV/SRC) and in the location, which affects the tephra fall thickness. We aimed for a standard data split of 80%/10%/10%, for train/validation/test, however given the above constraints, this produced a split of 80/8/12 (considering the number of bounding boxes and not the number of images). These datasets were used to develop our approach for building damage assessment. In line with studies shown in Table 1, we chose to split the damage assessment task into two subtasks: i) building localisation (i.e., identification of building bounding boxes within the images) and ii) damage classification. While it is possible to develop a model that can simultaneously locate and classify buildings with different levels of damage, model training under this approach can take significantly more time and resources to converge when compared to an approach that splits the tasks (Bouchard et al., 2022). Furthermore, decoupling the two tasks allows for

Formatted: Font: Italic

Formatted: Left

Deleted: 2

Formatted: Font: Cambria

Deleted: a sizable proportion of the data were not geotagged, images from each location were kept together to assure the train and test sets were independent.

Deleted: thickness of

Deleted: received

Deleted: with the majority of data assigned for training

Deleted: % train,

Deleted: 8% validation, and

Deleted: % test

greater flexibility; for example, if building locations are already known then only the classification can be run, speeding up the remote assessment.

In machine learning, the performance of a model and its optimal hyperparameters can be highly dependent on the characteristics of the dataset used for training, and hyperparameters that work well for one dataset may not work well for another. Therefore, it's common practice to optimise hyperparameters, model architectures, and training strategies to find the configuration that performs the best for a particular problem. For building localisation and damage classification we conducted a series of independent experiments using different image preprocessing approaches, CNN architectures, and combinations of hyperparameters with the aim of iterating towards the best experimental setup (Model selection: Section 3.1.1; Section 3.2.1). Each experiment consisted of three replicates of a given combination of these aspects. Replicates were conducted since the stochastic nature of the training process can cause models to converge at slightly different points (Aggarwal, 2018). For each experiment the replicate with the highest evaluation metric was the one compared against the other experiments.

Once we identified the best performing experimental setup for each task, we conducted K-fold cross validation on the combined training and validation sets to understand how the choice of these affects model performance (see Section 3.1.3, Section 3.2.2).

Following model selection and cross validation we calculated the performance of the best model identified for each task on the test set. Finally, to see if better performance could be achieved with more data available for training, we retrained the models on the combined training and validation data before evaluating on the test data (Evaluation on the test set: Section 3.1.3, Section 3.2.3). All stages of model development, including model selection, cross validation, and final evaluation, are shown in Figure 4 and more information about the specific experiments conducted for model selection is given in Section S3 of the supplementary material.

Past studies have trained deep learning algorithms on georeferenced images (i.e., each pixel has a geographical location attached) (Gupta and Shah, 2020; Shen et al., 2021; Bouchard et al., 2022) and non-georeferenced images (e.g., Li et al., 2019a; Pi et al., 2020; Cheng et al., 2021). In this work we labelled the non-georeferenced images and trained models on these. This was done firstly, to preserve the multiple viewing angles that we have of each building with each

Deleted: Most previous studies have split the damage assessment task into two subtasks: i) building localisation (i.e., identification of building bounding boxes within the images) and ii) damage classification (Table 1). Developing a model that can simultaneously locate and classify buildings with different levels of damage is feasible, however, model training under this approach can take a lot of time and resources to converge (Bouchard et al., 2022). Furthermore, decoupling the two tasks makes the approach more flexible in a post-disaster context, for example, if building locations are already known then only the classification can be run, speeding up the remote assessment. For these reasons, we split the building damage assessment task into two subtasks.

Deleted: : In line with the work of previous authors (Cheng et al. 2021; Bouchard et al. 2022), we split the building damage assessment task into two subtasks ... [2]

Deleted: deep

Deleted: experiment with different

Deleted: z

Deleted: settings

Deleted: each task in our damage assessment appro ... [3]

Deleted:

Deleted: One

Formatted: Not Highlight

Deleted: Three

Deleted: r

Deleted: ¶

Deleted: (building localisation, classification 1, ... [4]

Deleted: we combined the training and validation ... [5]

Deleted:

Deleted: To test the robustness to location, we train ... [6]

Deleted: final

Deleted: (building localisation, classification 1, ... [7]

Deleted: 4

Deleted: 3

Deleted: 2

Deleted: While studies have trained deep learning ... [8]

Deleted: Past studies have

Deleted: machine

Formatted: English (UK), Not Highlight

Formatted: English (UK), Not Highlight

Deleted: here we opted to

Deleted: train models

Deleted: on

Deleted: raw

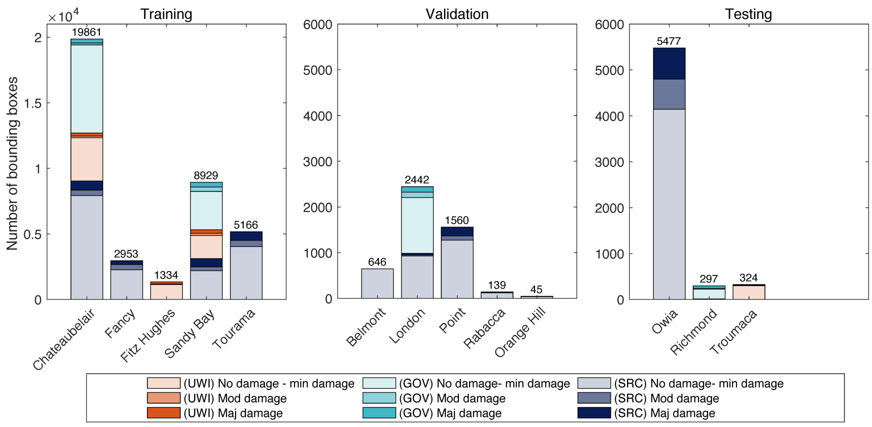
Deleted: non-rectified

Deleted: due to the absence of GPS location on a lar ... [9]

Deleted: ,

image counting as a different data point, and secondly, due to the absence of GPS locations on a large portion of the dataset. In an operational context, spatial information must be tied to the assessed damage. Therefore, beyond the creation of distinct models for each task, we designed a comprehensive, fully automated pipeline that integrates models for building localisation and damage classification. Our pipeline contains all the necessary processing steps to guide images through the separate models enabling them to operate on a georeferenced orthomosaic image (to be generated separately) or on non-georeferenced images. When applied to an orthomosaic image the output from the pipeline is a georeferenced vector dataset that can readily be plotted in a GIS to generate damage maps.

In Section 4 we apply the pipeline to assess building damage in Owia, St Vincent, which received 50-90 mm of tephra fall during the 2020-2021 eruption (Figure 1). Owia was selected out of the three possible test set locations (Figure 3) due to its large size and the existence of GPS locations that enabled the generation of a georeferenced orthomosaic image; for this we used Agisoft Metashape software. To compare the assessed building damage with tephra thickness, we used the TephraFits code (Biass et al., 2019) to identify the theoretical maximum accumulation using the isopachs from Cole et al., (2023). This maximum accumulation and the isopachs were interpolated using cubic splines and the surface was exported at a resolution of 10 m to provide a tephra thickness value for each building.



Deleted: which would otherwise be lost by converting and training on an orthomosaic image.

Deleted: I

Deleted: In a post-disaster context, the seamless functioning of models will benefit from a sequential workflow.

Deleted: B

Deleted: w

Deleted: all

Deleted: of

Deleted:

Deleted: allowing

Deleted:).

Deleted: that executess all optimized final models in turn.

Deleted: to produce an

Deleted: and the required processing steps to guide images through the models (Figure 3d). The pipeline runs on an orthomosaic image and generates spatial data in shapefile format

Deleted: our

Deleted: in

Deleted: is the islandSt Vincent and

Deleted: 8

Deleted: ; to do this we used

Deleted: which enabled the generation of an orthomosaic image. We generated an orthomosaic image using Agisoft Metashape software

Deleted: In the following sections we provide more detail on the algorithms and architectures used for each of the tasks, and how the performance of each task was evaluated.

Figure 3. The number of bounding boxes of each damage state in each UAV imagery dataset (UWI-TV, GOV, SRC) for each of the locations in this study. Imagery was divided into three groups: training, validation, and testing. The division of datasets between the three groups was chosen to incorporate diversity in the image sets (UWI-TV/GOV/SRC), whilst keeping images from the same location together and maintaining an approximate split of 80% training/10% validation/10% testing.

Deleted: *2*

Deleted: *2*

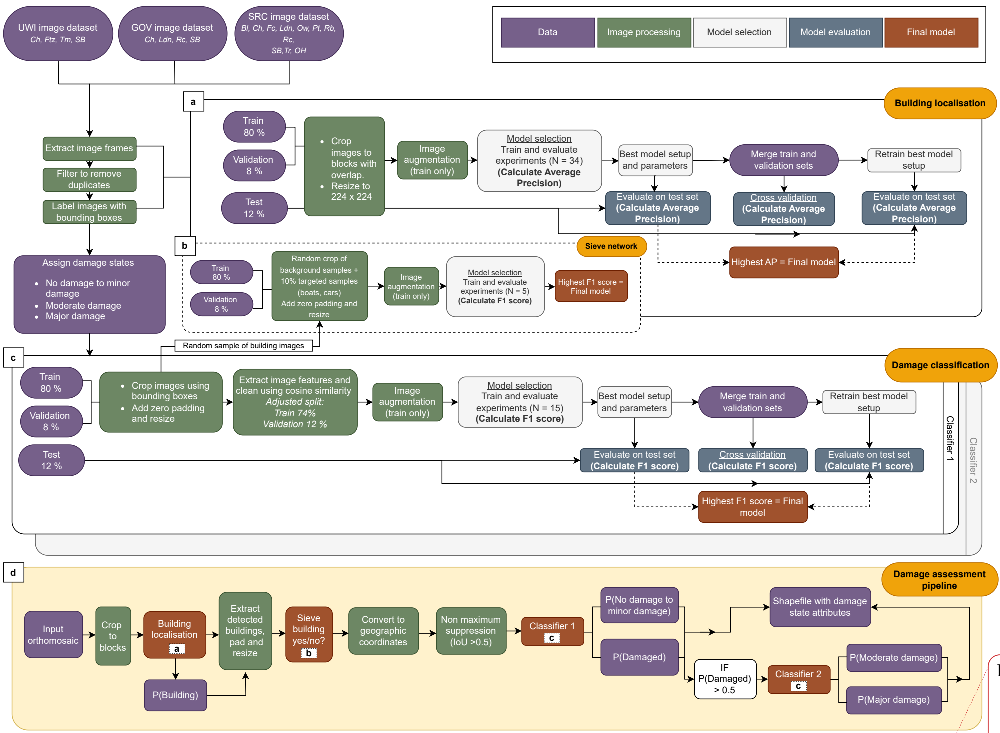
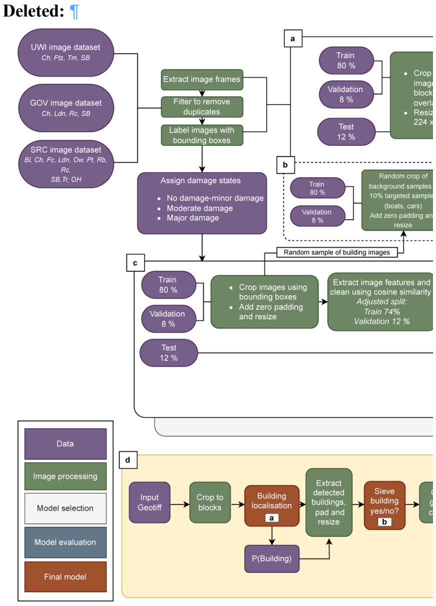


Figure 4. A schematic showing the full methodology for a) developing a model for building localisation, b) developing a sieve network, which acts as an add on to the building localisation model, c) developing a model for damage classification and d) the building damage assessment pipeline developed in this work. The pipeline operates on an orthomosaic image (to be generated separately) and incorporates the final trained models for building localisation and two stages of damage classification along with all the necessary processing steps to link the models. Dataset locations referred to are: Bl – Belmont, Ch – Chateaubelair, Fc – Fancy, Ftz – Fitz Hughes, Ldn – London, OH – Orange Hill, Ow – Owia, Pt – Point, Rb – Rabacca, Rc – Richmond, SB – Sandy Bay, Tr – Tourama, Tm- Troumaca. Pipeline schematic generated using draw.io.

Deleted: ¶



Deleted: 3.

2.3.1 Building localisation

For building localisation, we used the cutting edge two-stage object detector Faster R-CNN (Ren et al. 2017). When applied to a test image containing the relevant objects, Faster R-CNN outputs the positions within the image (X, Y, width, and height in pixels) of bounding boxes containing the object, and a confidence score for each box. As per customary practice (Zou et al. 2019) we used a confidence of > 0.5 meaning that only boxes with confidence greater than this are output.

For object detection, to reduce model training and inference time, full sized images were split into image blocks. Experiments conducted as part of building localisation model selection included variations in block size and the proportion of block overlap, along with the development of separate models for images captured with different viewing angles, training for only the SRC portion of the dataset (images mostly at nadir) and the combined UWI-TV-GOV portion (images mostly off-nadir). A total of 34 experiments were conducted to include all credible combinations of the varied hyperparameters and to find the best experimental setup (Table S2, supplementary material).

To improve the performance of the building localisation model we developed a sieve network that runs as an add on to the Faster R-CNN building detector. The sieve network reduces false positives which occur, when the detector predicts a bounding box that does not have an overlapping labelled building (i.e., detects a building when there is not one). More details on its development are provided in Section 3.2 of the supplementary material.

2.3.2 Damage classification

We chose to divide building damage classification into two separate classifications, Classifier 1 distinguishes between ‘No damage to minor damage’ versus the combined classes of ‘Moderate damage’ and ‘Major damage’, while Classifier 2 further differentiates between ‘Moderate damage’ and ‘Major damage’. A hierarchical approach to classification has been found effective when the number of samples is limited or classes are unbalanced (Li et al., 2019b; An et al., 2021). We conducted experiments separately for Classifiers 1 and 2. Experiments consisted of fine-tuning two different pretrained CNNs to determine which was better and should be used in the final models for each classifier: ResNet50 (He et al., 2015) trained on the ImageNet

Deleted: *Figure 3. A schematic showing the full methodology for a) developing a model for building localisation, b) developing a sieve network, which acts as an add on to the building localisation model, c) developing a model for building damage classification and d) the building damage assessment pipeline developed in this work. The pipelineand incorporates the final trained models for building localisation and two stages of building damage classification along with all the necessary processing steps to link the models. Dataset locations referred to are: Bl – Belmont, Ch – Chateaubelair, Fc – Fancy, Ftz – Fitz Hughes, Ldn – London, OH – Orange Hill, Ow – Owia, Pt – Point, Rb – Rabacca, Rc – Richmond, SB – Sandy Bay, Tr – Tourama, Tm- Troumaca.* ¶

Deleted: conducted experiments using

Deleted: Faster R-CNN is an improvement on the Fast R-CNN algorithm proposed by Girshick, (2015). The improvement comprises an initial region proposal network (RPN) which speeds up performance. InitiallyIn Faster R-CNN, image feature maps are extracted by passing the input image through a pretrained backbone CNN. The RPN then utilizes these features to generate proposals for potential object-containing areas, thi ... [10]

Deleted: patches

Deleted: the size of these patches blocks, and the amount of overlap between patchesblocks, and whether bl ... [11]

Deleted: fixedgivenFor each experiment t

Deleted: for building localisation

Deleted: see

Deleted: . More information on the values used for training can be found in the

Deleted: for details

Deleted: <#>Developing a sieve network¶

Deleted: small

Deleted: . Bounding boxes produced by the detector are passed to the sieve network to filter out detections ... [12]

Deleted: s

Deleted:

Deleted: Further information on the sieve networks

Deleted: can be found

Deleted: To develop the dataset used for training and evaluating the sieve network we randomly croppe ... [13]

Deleted: Building d

Formatted: Font: Times New Roman, Not Bold, English (UK)

Formatted: Normal

Deleted: .

Deleted: ¶
For building damage classification,

Deleted: w

Deleted: c

Deleted: This hierarchical approach to classification has been found effective when the number of samples ... [14]

dataset (Deng et al. 2009), and GoogleNet (Szegedy et al., 2015) trained on the places365 dataset (López-Cifuentes et al., 2019). Fine-tuning is a common approach to computer vision tasks where sufficiently large, labelled datasets are not available for the task at hand (typically hundreds of thousands of images are needed: Aggarwal, 2015). During fine-tuning, the high-level features that were learnt during the initial training on the large dataset can be leveraged for the new task. In addition to the different pretrained CNNs used, experiments also considered different ways of balancing the number of images for each damage state class (over-sampling the minority class, under-sampling the majority class and no balancing). When applied to a test building image, the trained classifier outputs the highest probability class and the associated probability. A total of 15 experiments were conducted for each of the classification tasks. For each experiment three replicates were conducted, each consisting of a grid search to find the best combination of learning rate, batch size and L2 regularisation. For more information on this see Section 3.3 of the supplementary material.

2.3.3 Model evaluation metrics

For building localisation Faster R-CNN experiments, we evaluated performance using the average precision (AP) at an intersection over union (IoU) threshold of 0.5, and the F1 score. AP, a common metric for evaluating object detection (Zou et al., 2019), measures how often the detector gets it right (true positives, TP) versus wrong (false positives, FP, and false negatives, FN). A TP occurs when a predicted box overlaps a labelled box by more than 50% (IoU > 0.5), a FP when there is no overlapping labelled box, and a FN when the detector misses a labelled box. When the detector is run on a test image a confidence score is output for each predicted box (0-1). Once the trained detector has been run over the full test set, the precision (TP/TP+FP), and recall (TP/TP+FN) are calculated at different confidence score thresholds and the area underneath the resulting precision-recall curve represents the AP. AP depicts the trade-off between precision and recall and provides an overall measure of detection performance. AP values range between 0-1, where a higher value indicates a better performance.

For building localisation, the F1 score was calculated at IoU and confidence thresholds of 0.5. The F1 score is calculated as: F1 = 2x (Precision x Recall)/ (Precision + Recall). To evaluate the performance of classification models, we used the macro-F1 score, which is the unweighted mean of the F1 scores calculated for each of the classes. Similarly to the AP, values of the F1 score range between 0-1, where a higher value indicates a better performance.

Deleted: the

Deleted: The AP is the most frequently used measure of an object detector's performance (Zou et al., 2019), and) and is calculated based on the number of times the detector gets it right (a true positive, TP) or wrong (a false positive, FP or a false negative, FN).

Deleted: A true positive occurs when the detector predicts a box that has an IoU with a labelled box of > 0.5.

Deleted: A false positive occurs when the detector predicts a bounding box that does not have an overlapping labelled box, while a false negative occurs where the detector fails to predict a box that is present in the labelled data

Deleted: The relative proportions of these are used to calculate the precision and recall, where precision is the number of things that were predicted as positive that were correct: Precision = TP/(TP+FP), and recall is the number of things that are truly positive that were identified: Recall = TP/(TP+FN).

Deleted: a

Deleted: (

Deleted:)

Deleted: (

Deleted:)

Deleted: which can be plotted against one another to form a curve. The AP is

Deleted:

Deleted: this

Deleted: ;

Deleted:

Deleted: it

Deleted: this

Deleted: ,

Deleted: ¶

3 Results

3.1 Building localisation

3.1.1 Model selection

The five experiments with the highest average precision are shown in Table 4, with the full list of experiments provided in Table S2 of the supplementary material. Average precisions across the 34 experiments ranged from 0.295 to 0.701 (Table 4 and Table S2). We found that block size played an important role in model performance; out of the 34 experiments conducted, the top three used a block size of 550 x 550 pixels, which was the middle of the sizes tested (450, 550, 650). We observed that models trained on the full dataset performed better than models trained separately for the nadir (SRC) and off-nadir imagery sets (UWI-TV and GOV sets combined) (Table 4 and Table S2).

Table 4. Hyperparameters for the five experiments with the highest average precision conducted for building localisation, ordered by average precision. The full table consisting of all 34 experiments is provided in the supplementary material. Columns marked with '' contain Yes/No information. Training dataset **: a= all, b= UWI-TV and GOV, c= SRC.*

Row ID	Block size	Mixed block size*	Block overlap	Block resized*	Training dataset **	Max Average Precision	F1 score
1	550	N	50%	Y	a	0.701	0.669
2	550	N	20%	Y	a	0.700	0.668
3	550	N	20%	Y	a	0.700	0.642
4	650	N	50%	Y	a	0.691	0.654
5	650	N	20%	Y	a	0.678	0.670

All trained sieve networks achieved macro and class F1 scores that were > 0.973 (Table S3, Supplementary material). The sieve networks efficacy at improving building localisation is demonstrated by comparing the results of the best detector when applied to the validation dataset pre-sieving (Table 4 row ID 1) with the post-sieving results. Pre-sieving there were a large number of false positive detections, resulting in a precision of 0.588, post-sieving these were reduced and the precision increased to 0.695 (Table 5).

Deleted: top

Deleted: (

Deleted:)

Deleted: conducted for building localisation

Deleted: 3

Deleted: 3

Deleted: 3

Deleted: More details on the results of experiments run for building localisation model selection can be found in Section S2.1 of the supplementary material.

Deleted: *3*

Deleted: *T*

Deleted: *highest scoring (average precision)*

Formatted Table

Deleted: All training/ UWI-TV&GOV/ SRC

Deleted: id

Deleted: ?

Deleted: ?

Deleted: all

Deleted: all

Deleted: all

Deleted: all

Deleted: all

Deleted: <#>Sieve Network

Deleted: <#>Table 5

Deleted: <#>4

Deleted: <#>The best performing sieve network experiment achieved a macro F1 score of 0.977.

Deleted: <#>The best detector identified through model selection (Table 43, row 1) achieved an F1 score of 0.669 (Table 76), with a precision and recall of 0.588 and 0.776, respectively, on the validation data. The lower value of precision is due to the substantial number of false positive detections. After the results of the detector were passed through the sieve network, the number of false positives was reduced, with an improved F1 score of 0.712 (Table 76).

Table 5. Comparing the performance of the best building localisation model when applied to the validation dataset before and after running the results through the sieve network.

	Precision	Recall	F1
Best detector pre-sieving	0.588	0.776	0.669
Best detector post-sieving	0.695	0.730	0.712

3.1.2 Cross validation

Cross validation was conducted for the single best performing building localisation model (without the sieve network) to understand how the choice of training and validation data affects performance. Analysing performance variations across different testing datasets can then inform recommendations for future data collection strategies (see Section 6).

We found that the performance of the selected object detector varied, depending upon the location (Figure 5a) or imagery dataset (Figure 5b) used for testing. For models tested on different locations average precisions in line with the AP achieved on the full validation set (0.701) were obtained for Point and Fancy (Figure 5a). The lowest AP values were for London (0.063) and Fitz Hughes (0.187). The standard deviation (SD) (Figure 5) shows the variability in performance between the three replicates that were trained for each test, which arises due to the stochastic nature of the training process. For models tested on the different imagery datasets individually the AP was low, with a mean value across all datasets of < 0.2 (Figure 5b). For all three locations (Chateaubelair, Sandy Bay, London), AP for models evaluated on the SRC dataset were lower than for the UWI-TV or GOV datasets.

Deleted: *4*

Deleted: *Experiments conducted for the sieve network, a small network that runs on the boxes produced by the object detector. Results are ordered from high to low by the Macro F1 score. ResNet50 and GoogleNet refers to the convolutional neural network architecture used in the experiment; the value after the underscore reflects the experiment ID where different IDs have different training parameters (see Section S2 of the supplementary material).*

Formatted: Justified

Deleted: ¶
Experiment ID ... [15]

Formatted: Font: Bold

Formatted Table

Formatted: Font: Bold

Formatted: Font: Bold

Deleted: , along with the potential for the model to generalize to a new dataset.

Deleted: s

Deleted: 4

Deleted: 4

Deleted: (Figure 4a)

Deleted: > 0.7 were

Deleted: 4

Deleted: in line with AP achieved on the full validation set (0.701)

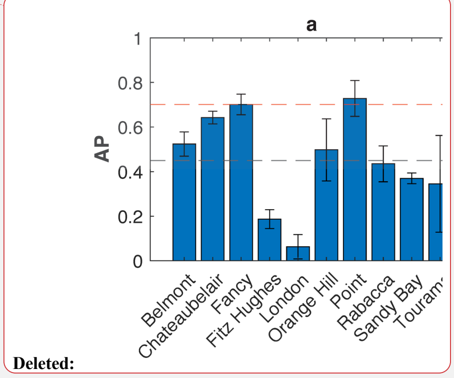
Deleted: 4

Deleted: (Figure 4b)

Deleted: 4

Deleted: higher

Formatted: Not Highlight



Deleted:

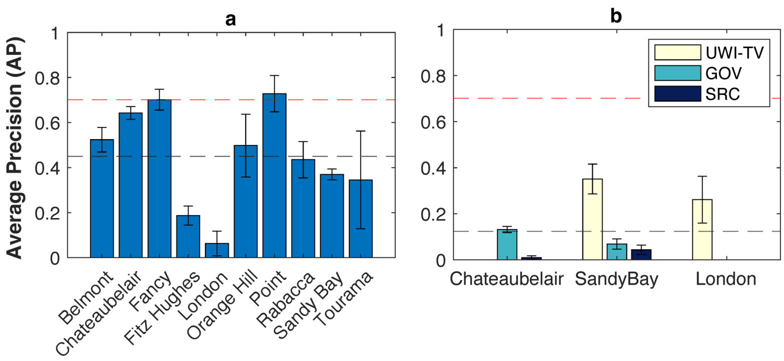


Figure 5. Cross validation of the best experimental setup for building localisation models which are trained to predict building box positions within the image. a) The effect of changing the location used as the test set on detector average precision (AP) and b) the effect of changing the imagery dataset (UWI-TV/GOV/SRC) used as the test set on AP. For b) cross validation of the imagery dataset, models are trained on all data from that location excluding the location used for testing as indicated by the bar. For London there is data from the GOV dataset, however the number of images in the SRC dataset is insufficient for training, so no bar is shown for GOV. The AP shown is the mean value from three trained models with the same setup while the error bars show the standard deviation. Black dashed lines show the mean AP value across all cross validation trained models; red dashed lines show the best AP from the experiments (0.701: Table 4).

3.1.3 Evaluation on the test set

Evaluation of the best detection model on the test set, which consists of completely unseen data from Owia, Richmond and Troumaca (Figure 3) produced an AP value that is the same as the value on the validation data (0.701) (Table 6). To understand if a better model could be achieved with more data available for training, we combined the training and validation data and used this to retrain the best experimental setup for the detector. Evaluation of the retrained model on the test set resulted in an average precision increase from 0.701 to 0.751 for the non-sieved detector, and from 0.668 to 0.728 for the sieved detector, showing that having more data available for training produced a better model (Table 6).

While the AP is higher for the retrained detector without the sieve, the addition of the sieve network creates a better balance between the precision and recall which is reflected in the

Deleted: *4.*

Deleted:

Deleted: *3*

Deleted: 2

Deleted: 4

Deleted: 3

Deleted: Retraining

Deleted:

Deleted: using the combined training and validation data caused the AP when applied to the test data to increase to 0.751 prior to sieving and 0.728 after sieving

Deleted: Comparing the precision and recall of the retrained detector and the retrained detector + sieve network shows that ...w
w

Deleted: w

Deleted:

higher F1 score (Table 6). For the present application equal importance is given to: 1) making correct predictions about building locations, and 2) identifying as many buildings as possible. Consequently, striking the balance between precision and recall is crucial. We therefore selected the retrained detector + sieve network as the final building localisation model and the model that is incorporated into the damage assessment pipeline (Table 6).

*Table 6. Comparison of the best building localisation models' performance when evaluated on the validation and the test sets. AP is average precision, P is precision, and R is recall. * Retrain models are trained on the combined training and validation sets. Results for the final model that is used in the damage assessment pipeline are in bold.*

	Validation set				Test set			
	AP	P	R	F1	AP	P	R	F1
Detector (0.5 conf)	0.701	0.588	0.776	0.669	0.701	0.604	0.776	0.679
Detector + Sieve (0.5 conf)	0.681	0.695	0.730	0.712	0.668	0.606	0.757	0.673
Detector retrain					0.751	0.642	0.816	0.719
Detector retrain +sieve					**0.728**	**0.710**	**0.782**	**0.744**

3.2 Damage classification

3.2.1 Model selection

The five experiments with the highest macro F1 score are shown in Table 7, with the full lists provided in Tables S4 and S5 of the supplementary material. For Classifier 1, Macro F1 scores across all 15 experiments ranged from 0.753 to 0.836, while for Classifier 2 scores ranged from 0.776 to 0.810 (Tables 7, S4, S5). Models trained to differentiate between the No damage to minor damage and Damaged classes performed better for the No damage to minor damage class, while those trained to differentiate between Moderate and Major damage performed better for the Major damage class (Table 7). The best performing models for both classifiers used the ResNet50 architecture rather than GoogleNet with an unbalanced dataset. For Classifier 1 the best model had F1 = 0.962 for the No damage to minor damage class and F1 = 0.710 for the Damaged class. While for Classifier 2 the Moderate damage class had F1 = 0.770 and Major damage F1 = 0.851.

Deleted: , reflected in a higher F1 score.
Deleted: 7
Deleted: 6
Formatted: Caption, Left, Keep with next
Formatted Table
Deleted: Building
Deleted: d
Deleted: top
Deleted: (
Deleted:)
Deleted: conducted for building damage classification
Deleted: 6
Deleted: 5
Deleted: 3
Deleted: 4
Deleted: and 0.776 to 0.810 for
Deleted: c
Deleted: 1 and 2 respectively
Deleted: 6
Deleted: 5
Deleted: 3
Deleted: 4
Deleted: Not Damaged

Table 7. The top five experiments conducted for each of the building damage classifiers, ordered by the macro F1 score. The full list consisting of all 15 experiments for each classifier is provided in Tables S4 and S5 of the supplementary material.

Deleted: *6*
Deleted: *5*
Deleted: *3*
Deleted: *4*

Row ID	Architecture	Class balancing: Not Balanced/ under-sampled/ over-sampled	F1 No damage to minor damage	F1 Damaged	F1 Macro
Classifier 1					
1	Resnet50	not	0.962	0.710	0.836
2	Resnet50	not	0.960	0.696	0.828
3	Resnet50	not	0.957	0.699	0.828
4	Resnet50	not	0.962	0.692	0.827
5	Resnet50	under	0.951	0.646	0.799

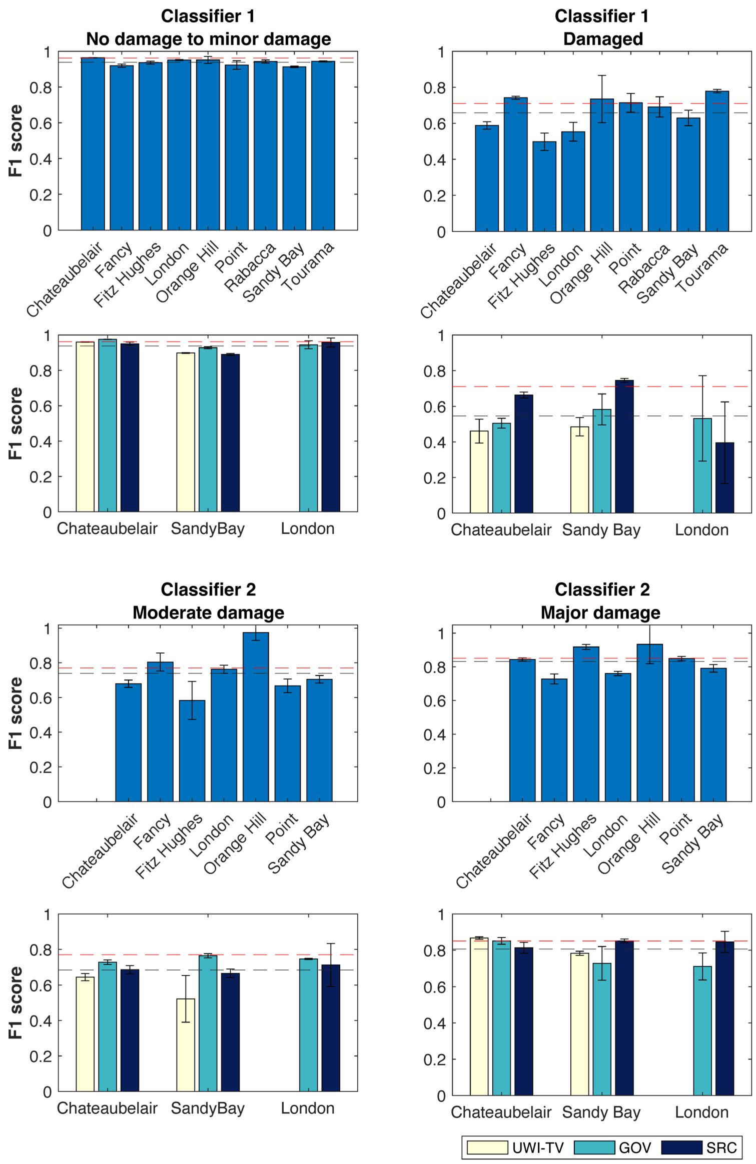
Row ID	Architecture	Class balancing: Not Balanced/ under-sampled/ over-sampled	F1 Mod damage	F1 Maj damage	F1 Macro
Classifier 2					
1	Resnet50	not	0.770	0.851	0.810
2	GoogleNet	over	0.737	0.848	0.793
3	Resnet50	over	0.749	0.835	0.792
4	Resnet50	not	0.749	0.835	0.792
5	Resnet50	under	0.735	0.845	0.790

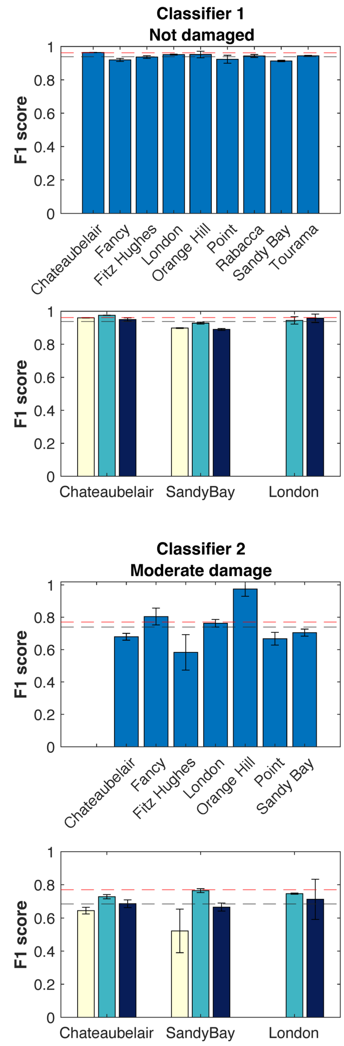
Formatted Table
Deleted: t
Deleted: D
Deleted: d

3.2.2 Cross validation

Cross validation was conducted for both of the single best performing models for Classifiers 1 and 2 identified through model selection. As was the case for the best building localisation model, this was done to understand how the choice of training and validation datasets affected model performance and to understand how our model might perform on a new dataset.

Deleted: the potential for our model to generalize to





Deleted:

Figure 6, Cross validation for Classifiers 1 and 2. For rows 1 and 3 the best experimental setup was retrained on all the data from locations in the combined training and validation data and evaluated on the location shown. For rows 2 and 4 the best experimental setup was retrained on

Deleted: *5*

all the data from the location shown and evaluated on each dataset (UWI-TV/GOV/SRC) separately. Each training was conducted three times, the value plotted is the mean, and the error bars show the standard deviation. Black dashed lines show the mean F1 score across all cross validation trained models, red dashed lines show the best F1 score for each class from the experiments (Table 6).

The performance of Classifier 1 for the No damage to minor damage class is consistent across the distinct locations and datasets used for evaluation with mean F1 scores between 0.913-0.983 for locations and 0.898-0.976 for datasets (Figure 6). For the Damaged class there is more variety in the performance across the locations and datasets used for evaluation. The mean F1 scores for the separate locations range from 0.588 (Fitz Hughes) to 0.779 (Tourama) while for the different datasets the range is 0.393 (London-SRC) to 0.745 (Sandy Bay-SRC).

For Classifier 2, the Moderate damage class is more sensitive to the choice of location and dataset used for the evaluation than the Major damage class (Figure 6). For the different locations the mean F1 score ranged from 0.583-0.974. Similarly to Classifier 1, the location with the lowest mean F1 score is Fitz Hughes, whereas the highest score was produced for Orange Hill. For the different datasets the range for the Moderate damage class is between 0.522-0.746. For the Major damage class F1 scores for the distinct locations are between 0.728-0.933 while for the different datasets the range is between 0.711-0.867.

3.2.3 Evaluation on the test set

Evaluation of the single best models for Classifier 1 and Classifier 2 on the unseen test set produced Macro F1 scores that were comparable with the scores for the validation set: 0.829 for Classifier 1 and 0.791 for Classifier 2 (Table 8). For Classifier 2, retraining the model on the combined training and testing data increased the Macro F1 score from 0.791 to 0.838. Whereas for Classifier 1 retraining produced a slightly lower Macro F1 score (0.809 compared to 0.829). Nevertheless, the retrained model for Classifier 1 achieved a higher recall on the Damaged class than the non-retrained model. In an operational setting it's desirable to correctly classify as many of the damaged buildings as possible, since in our pipeline these will be passed onto Classifier 2, therefore we took the retrained models for both classifiers as the final models and the models that are incorporated into the damage assessment pipeline.

Deleted: 5

Deleted: Figure 5 shows that the

Deleted: t

Deleted: d

Deleted: testing

Deleted: 5

Deleted: the choice of

Deleted: validation

Deleted: 5

Moved (insertion) [1]

Deleted: .

Moved up [1]: For the Major damage class F1 scores for the distinct locations are between 0.728-0.933.

Deleted: For the Moderate damage class, the mean F1 score ranged from 0.583-0.974. Similarly to Classifier 1, Fitz Hughes produced the lowest mean F1 score, whereas the highest score was produced for Orange Hill. For the Major damage class F1 scores for the distinct locations are between 0.728-0.933.For Classifier 2 the sensitivity to the choice of dataset (UWI-TV/GOV/SRC) for the Moderate damage class is greater than for the Major damage class. For Moderate damage, the range is between 0.522-0.746, while for Major damage the range is from 0.711-0.867.¶

Deleted: c

Deleted: ication

Deleted: classification

Deleted: Table

Deleted: 7

Deleted: 6

Deleted: d

Deleted: The confusion matrices for both final models are plotted in Figure 6, these show class accuracy i.e., how many of the true class were correctly classified. For Classifier 1 89% of the Not damaged buildings were correctly classified, and 73% of the Damaged buildings were correctly classified. For Classifier 2 81% of the moderately damaged buildings were correctly classified, while 87% of the buildings with major damage were correctly classified. ...

*Table 8. Comparison of the best damage classification models' performance when evaluated on the validation and the test sets. AP is average precision, P is precision, and R is recall. * Retrain models are trained on the combined training and validation sets. Results for the final models that are used in the damage assessment pipeline are in bold.*

	Validation set							Test set						
	No damage to minor damage			Damaged				No damage to minor damage			Damaged			
	P	R	F1	P	R	F1	F1 Macro	P	R	F1	P	R	F1	F1 Macro
Classifier 1	0.950	0.976	0.962	0.793	0.643	0.710	0.836	0.891	0.940	0.915	0.809	0.689	0.744	0.829
Classifier 1 retrain								**0.899**	**0.894**	**0.896**	**0.717**	**0.728**	**0.722**	**0.809**
	Mod Damage			Maj Damage				Mod Damage			Maj Damage			
Classifier 2	0.769	0.660	0.770	0.852	0.825	0.851	0.810	0.903	0.663	0.765	0.730	0.927	0.817	0.791
Classifier 2 retrain								**0.861**	**0.809**	**0.834**	**0.817**	**0.866**	**0.841**	**0.838**

4 Application of the full damage assessment pipeline: Assessing tephra fall building damage in Owia

In this work we have developed separate models for building localisation and two stages of damage classification. However, in an operational context models need to work sequentially, this led to the development of our damage assessment pipeline (outlined in Figure 4d). The pipeline operates on an orthomosaic image and outputs a georeferenced vector set, with the following *attributes* for each building that is detected: *detection* (box confidence score), *ClassPred_1* (output class from Classifier 1, Damaged or No damage to minor damage), *ClassProb_1* (the probability of that class), *ClassPred_2* (output class from Classifier 2, Moderate damage or Major damage, this is only run if Classifier 1 outputs damage), *ClassProb_2* (the probability of the class output by Classifier 2), *damageState* (the final damage state).

The tephra fall building damage map shown in Figure 7a was produced by overlaying the georeferenced vector that was output by the pipeline with the orthomosaic image in QGIS. Our remote damage assessment pipeline identified 442 buildings. Of these, 78% (N = 343) were classified as having No damage to minor damage, 9% (N = 40) as having Moderate damage and 13% (N = 59) as having Major damage. We observed that the two upper tephra fall thickness

Deleted: *7*
Deleted: *6*
Deleted: *'*
Deleted: ¶ AP ... [16]
Deleted: t
Deleted: d
Deleted: t
Deleted: d
Deleted: d
Deleted: ¶ ¶ ¶ ¶ ... [17]
Deleted: **Example**
Deleted: **a**
Deleted: In order to
Deleted: 3
Deleted: , which can easily be generated using soft ... [18]
Deleted: s
Deleted: shape file
Deleted: c
Deleted: d
Deleted: n
Deleted: t
Deleted: d
Deleted: c
Deleted: m
Deleted: m
Deleted: c
Deleted: c
Deleted: .
Deleted: Figure 7 shows an example
Deleted: running the pipeline on an orthomosaic i ... [19]
Deleted: pipeline orthomosaic and output
Deleted: shapefile
Deleted: The example which consists of 417 buildi ... [20]

bins (70-80 mm and 80-90 mm), both had a higher proportion of buildings with Major damage compared to the lower thickness bins (Figure 7b, c), indicating a correlation between tephra fall thickness and building damage though it is not very pronounced. These findings are discussed in Section 5.3.

The full pipeline took 1 hour to run on a standard 16GB RAM 2021 MacBook Pro, with an M1 Pro chip. Most of the inference time was attributed to the building localisation module in the pipeline, which may be bypassed if building footprints are already available. When only the classifiers were run the time taken to run was reduced to < 5 mins.

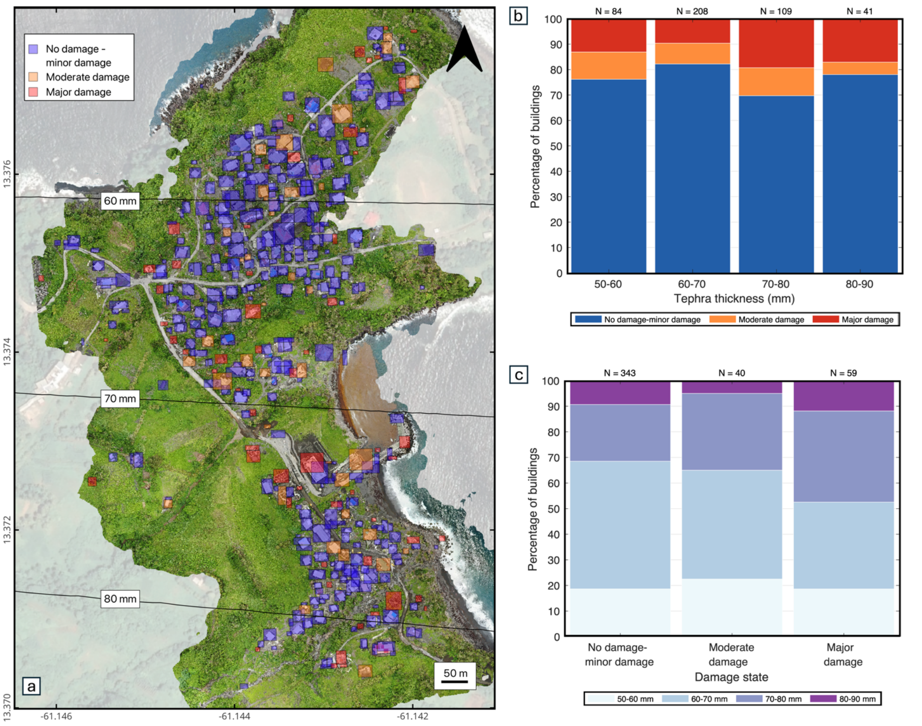


Figure 7. Application of our remote tephra fall building damage assessment pipeline to Owia, located in the north of St. Vincent. a) The damage map produced by overlaying the spatial data generated by our pipeline onto the orthomosaic image, black lines are tephra isopachs interpolated from Cole et al., 2023; b) the proportion of damage states with increasing tephra

Deleted: *7*

Deleted: *the full*

Deleted: *on the orthomosaic image for*

Deleted: *.*

Deleted: *,*

Deleted: *t*

thickness; c) the proportion of tephra thickness bins with increasing damage state. Coordinate reference system: WGS 84 (EPSG:4326). Satellite basemap © Google Maps 2024.

5 Discussion

In this work we have developed models for building localisation, and two levels of damage classification for building damage resulting from tephra fall. Our final models demonstrate strong performance for both building localisation (AP = 0.728; F1 = 0.744) and damage classification (Classifier 1, F1 = 0.809, Classifier 2, F1 = 0.838). Despite using post-event imagery only, which makes the task more challenging than approaches using multi-temporal imagery, our results are comparable to existing optical imagery building damage assessments developed for various hazards that use both mono-temporal and multi-temporal images (F1 scores are between 0.656-0.868 for building localisation and 0.650-0.981 for damage classification, Table 1).

Deleted: AP =
Deleted: building
Deleted: that use
Deleted: both pre- and post-event
Deleted: for building localisation
Deleted: :
Deleted: ,
Deleted: damage classification:

5.1 Building localisation

Through running our building localisation experiments we found that the pre-processing of images before detector training (particularly the block size) significantly influenced detector performance. The block sizes tested were chosen as a trade-off between reducing image size sufficiently to reduce computational cost, and retaining a large enough size such that buildings were not dissected unnecessarily. Given that the optimum block size was the middle size of the range tested, we are confident that this balance was achieved. Cross-validation results demonstrated variability in average precision (AP) for models trained on different locations and imagery datasets (UWI-TV/GOV/SRC) (Section 3.1.2; Figure 5). Deep learning models are known to perform well when the data they are evaluated on have similar characteristics to the data they were trained on, though have more difficulty when working with 'out of distribution' samples (Ben-David et al., 2010). Given the relatively consistent building typology across locations (most buildings observed are detached single storey buildings with either a gable or hip shaped metal sheet roof; a lesser proportion have flat concrete roofs), the differences in AP are likely due to observable variations in UAV altitude, off-nadir angles, tephra thicknesses, and varying training sample sizes.

Deleted: 3
Deleted: 4
Deleted: data come from the same distribution
Deleted: ,
Deleted: the majority of

The cross-validation AP was notably lower for the London and Fitz Hughes datasets (Section 3.1.2). For the London images (from SRC and GOV datasets) this is likely caused by the smaller apparent size of buildings in these images compared to the other locations, due to the higher UAV altitude. Variations in object size within the training and testing data has been found to affect the performance of deep learning models developed for building localisation, with models often performing better for objects that are the same size as those in the training data (Nath and Benzadan, 2020; Cheng et al., 2021; Bouchard et al., 2022). Fitz Hughes images were all from the UWI-TV image dataset which contributed just 17% to the combined training and validation set used for cross validation. This dataset was collected closer in time to the eruption, therefore as a whole had more tephra on the ground than the SRC and GOV datasets, which affects background colour. Furthermore, the UWI-TV dataset viewed buildings mostly from an off-nadir perspective, while the other datasets were predominantly nadir images. The effect of image background colour on localisation performance is expected to be minor, Cheng et al., (2021) found that for the same event localisation AP dropped from 65.6 to 63.3 when their model was tested on images containing buildings surrounded by vegetation compared to buildings with an ocean backdrop. While Bouchard et al., (2022) suggested that models quickly learn to ignore background pixels. On the other hand, variation in off-nadir angles is a widely acknowledged challenge of working with UAV or aerial images (Cotrufo et al., 2018; Nex et al., 2019; Pi et al., 2020). Under representation of the mostly off-nadir UWI-TV images in the training data may have impacted the model's ability to recognise such instances in the test data. During model development we experimented with different models for the different datasets (UWI-TV, GOV, SRC), but found that models developed on the combined dataset performed better than those developed on the separate datasets and a combined model was the one selected and used for cross validation. Rather than suggesting that variations in off-nadir angle are not important, this finding likely reflects the smaller size of the individual datasets compared to the combined datasets, meaning that less information was available to learn from. The application of sampling approaches like those used for the damage states in the classification model development (over or under sampling) could have been applied to balance the data. However, the SRC dataset is much larger than either of the UWI-TV and GOV sets (Figure 3), therefore we considered that oversampling would introduce significant bias towards the specific examples in the under-represented dataset, whereas through under sampling we would lose a large amount of the data that are available to learn from. Given these factors, we did not use sampling approaches. Future work might consider the application of generative AI

Deleted: The

Deleted: and

Deleted: (UWI-TV) exhibited the lowest average precisions (Figure 4). Both London datasets featured smaller buildings than the rest of the locations, evident in Section S3 of the supplementary material, while the

Deleted: s

Deleted: a

Deleted: nd,

Deleted: The training data was, predominantly nadir images from the SRC dataset with fewer UWI-TV examples , collected further in time after the eruption had fewer UWI-TV examples which are off-nadir and, collected more closely in time after the eruption, meaning more less tephra was present in images and, had.

Deleted: Images captured with different

Deleted: differences

Deleted: s

Deleted: .

Deleted: This

Deleted: U

Deleted: u

Deleted: z

Deleted: training and testing

Deleted: images

Deleted: was

Deleted: better than

Deleted: doing this separately

Deleted: T

Deleted: is likely related

Deleted: to

Deleted: , rather than the difference in the off-nadir angles...

Deleted: ,

Deleted: h

Deleted: 2

Deleted: we did not use this approach as algorithms such as generative adversarial networks (GANs) to expand the dataset (e.g., Yi et al. 2018; Yorioka et al., 2020), although more work needs to be done to quantify the diversity in the generated data.

The variability in cross-validation results for the building localisation model likely comes from a combination of the above factors (differences in UAV altitude, off-nadir angles, tephra thickness, and varying training sample sizes), and suggests that there was insufficient information in the training data for our detection models to perform well across the range of characteristics present. This is supported by the increased performance when the best localisation model was retrained on the combined training and validation data. However, further investigation is required to separate the unique effect of each aspect.

5.2 Damage classification

The final classification models achieved better performance than the final localisation model with macro F1 scores of 0.809 and 0.838 on the test data (Table 8). Cross-validation showed that classification models were less sensitive than the localisation model to the choice of datasets used for training and evaluation (Section 3.2.2). We found that class wise our models performed better on the No damage to minor damage class followed by the Major damage class. This agrees with other multi-class studies that have found the extremities of the damage state scheme applied easier to classify than the intermediate ones (Kerle et al., 2019, Valentijn et al., 2020).

5.3 Application of the full damage assessment pipeline: Assessing tephra fall building damage in Owia

Application of our remote damage assessment pipeline to the town of Owia found that 22% of buildings that received tephra accumulation in the range of 50-90 mm experienced Moderate damage or Major damage. Within this range, the relationship between tephra thickness and building damage was not as pronounced as in other studies (Blong, 2003b; Hayes et al., 2019; Jenkins et al., 2024). This may be attributed to the small geographic area and therefore small range of tephra thicknesses considered in our application when compared to other studies. In the damage assessments of Blong, (2003b), Hayes et al., (2019) and Jenkins et al., (2024) buildings received ~100 to 950 mm, trace to 600 mm and, trace to >220 mm respectively.

Deleted: (Section 3.1.3)

Deleted: this requires

Deleted: Building

Deleted: d

Deleted: 7

Deleted: 6

Deleted: n

Deleted: t

Deleted: d

Deleted: m

Deleted: is in agreement

Deleted: used

Deleted: to be

Formatted: Font: 12 pt, Bold, English (UK)

Formatted: Line spacing: 1.5 lines

Deleted: <#>Application of the pipeline¶

Spence et al., (1996) assessed building damage over a similarly narrow range of tephra thicknesses to this work (~150-200 mm) and found that there was considerable variation in the level of damage despite the majority of buildings having a metal sheet roof. The spacing between the principal roof supports (roof span) was found to be important for the amount of damage observed, with long span buildings experiencing higher levels of damage than short span ones (Spence et al., 1996). There are limited long span buildings in the Owia case study, however additional characteristics such as construction style and material, building layout, age, condition, height, and roof pitch can all affect a buildings ability to withstand tephra loading (Spence et al., 1996; Pomonis et al., 1999; Blong, 2003b; Jenkins et al., 2014). Variation in these characteristics across Owia could be responsible for the observed variation in building damage over the narrow range of thicknesses considered.

If we convert tephra thickness to loading, we can compare the results of our assessment with existing relationships between tephra loading and damage for similar building types. Using a density of 1500 kg/m^2 (Cole et al., 2023) suggests that a loading of at least 75-135 kg/m^2 was applied to buildings for the range of thicknesses considered (50 mm-90 mm). Census data for Owia states that 90 % of buildings have metal sheet roofs (SVG population and housing census, 2012), with the remaining 8% comprised of reinforced concrete roofs and 2% ‘other material’. Given the higher resistance of the 8% of non-metal sheet roof buildings in Owia, we might expect vulnerability models developed for metal sheet roofs to overestimate damage in the town. Fragility functions developed for Indonesian style buildings with metal sheet roofs (Williams et al., 2020), calculate a 48-80% probability of Owia buildings experiencing damage exceeding Damage State 2, higher than the 22% experiencing Moderate or Major damage in our study. Fragility curves for roof failure (Major damage) of old or poor condition metal sheet roofs (Jenkins et al., 2014), calculate that just over 10% of buildings in Owia would experience sufficient loading for roof collapse, comparable to the 13% observed in our study. These comparisons highlight some of the challenges associated with using vulnerability models developed for different locations. Moreover, they reiterate the need for the collection of both post-event impact data and building typology information that can be used to increase the amount of empirical data available for vulnerability model development and allow regional vulnerability models to be developed for specific building types.

Deleted: other
Deleted: 10
Deleted: ?
Deleted: 10
Deleted: fragility curves
Deleted: village
Deleted: (to change depending on what the other 10% is)
Deleted:
Deleted:
Deleted: locations, and
Deleted: Furthermore
Deleted: It
Deleted: building typology and

Like the studies presented in Table 1, our pipeline consists of separate models for localisation and damage classification. One of the benefits of this is that in locations where precise building location information is available for the assessment area, the localisation step can be bypassed and only the classifiers run. This not only enhances overall performance but also significantly reduces computation time. Furthermore, either of the classifiers can be run independently and/or combined with other damage assessment procedures; for example, an initial synthetic aperture radar (SAR) based assessment (e.g., Yun et al. 2015, Jung et al., 2016), could be followed with our Classifier 2 to provide additional granularity on the severity of the damage at a building level rather than a pixel level.

Deleted: O

Deleted: building

Deleted: c

5.4 Generalisability to other locations

Our models have performed well for images collected on the island of St Vincent where building typologies are relatively consistent. We therefore expect that our models will perform well in other locations with similar building types, such as the other islands in the Lesser Antilles. This hypothesis should be validated through further testing. In absence of additional UAV datasets that include damaged buildings, testing can be done by conducting pre-event surveys to test the performance of the building localisation model and Classifier 1 for the No damage to minor damage class. While this is unable to assess the ability of our approach to classify damage, it would provide *some* indication of performance following an event in a new location.

Deleted: c

Deleted: t

Deleted: d

To develop a model that is robust to the diverse building types found across the world necessitates assembling diverse datasets showcasing potential variations in building types and the associated tephra fall damage. To our knowledge the UAV datasets described in this work are the first of their kind. However, the increasing utilisation of UAVs during and after volcanic events suggests the possibility of the emergence of more datasets in the years to come. Our model represents a crucial initial step towards the operational implementation of this approach globally. The compilation of global tephra fall building damage UAV datasets will facilitate the ongoing refinement of building damage assessment approaches, including the one presented here. In pursuit of this objective, our models stand ready for retraining as more data becomes available. While our approach leverages images captured under a spectrum of flight conditions (off-nadir angle, altitude, flight trajectory), our investigation has both pinpointed specific

Deleted: However

conditions that are best suited for capturing building damage, which are detailed in Section 6, and highlighted the importance of consistency in data collection.

5.5 Improving model performance and future perspectives

The advantages of acquiring additional UAV datasets both before and after an event have been outlined in Section 5.4. In addition to this, pre-event imagery can be used to construct building inventories manually or using machine learning methods (e.g., Iannelli and Dell'Acqua, 2017; Gonzalez et al., 2020; Meng et al., 2023). Prior to an eruption, information about how the building typologies present will respond under certain tephra loadings (i.e., the forecasted damage state) can be obtained through the application of fragility functions. This information could enhance our model by serving as prior information that is updated with outputs from our remote damage assessment using Bayesian statistics. A similar approach has been suggested for updating the United States Geological Survey's (USGS) Prompt Assessment of Global Earthquakes for Response (PAGER) system (Noh et al., 2020). The framework provides a structured way of incorporating the PAGER forecasted loss with the potentially noisy and incomplete observations of loss in the early stages of response.

Alternatively, with ample individual building inventory data available, tailored damage classification models for specific building typologies could be developed and applied. The rationale is that a model dedicated to a specific building type is expected to outperform a generic multi-typology model.

In this work, we established a three-class damage state framework. Existing frameworks that were developed for ground based tephra fall damage assessment split damage into five damage states classes and one non-damage class (Spence et al, 1996; Blong, 2003; Hayes et al., 2019; Jenkins et al., 2024, Table 2) however in our preliminary analyses we found that: 1) in many images we were unable to confidently apply a six-class scheme due to only being able to see one side of the building, and 2) there were not enough examples of each damage state class to be able to train a six-class model. With the addition of future tephra fall building damage datasets it may be possible to apply a finer resolution damage state framework that can provide more detail on the observable damage. However, it is unlikely that the resolution of ground-surveys can be achieved using optical imagery, since lower damage states are still difficult to resolve

Deleted: 3

Deleted: surveys

Moved up [2]: or using machine learning methods such as the work of Meng et al., (2023).

Moved (insertion) [2]

Deleted: interrogated manually or using machine learning methods to construct building inventories that contain information such as construction materials and styles (e.g., Iannelli and Dell'Acqua, 2017; Gonzalez et al., 2020; Meng et al., 2023).

Deleted: may be particularly beneficial in constructing building inventories, which or using machine learning methods such as the work of Meng et al., (2023). include details about building typologies such as construction materials and styles. Surveys can be interrogated manually to extract building attributes

Deleted: given knowledge

Deleted: , an idea about how the buildings

Deleted: . The forecasted damage state could be subsequently refined through Bayesian updating

Deleted: based on our damage assessment models output

Deleted: based on our damage assessment models output

Deleted:

Deleted: three class

Deleted: in review

Deleted: Nevertheless, the damage states developed in our work can be equated to existing damage states generated for ground surveys such that: No damage – to minor damage = DS0-DS1, Moderate damage = DS2 and Major damage = DS3-5.

Deleted: may be applied

Deleted: is capable of providing

Deleted: confidently

even with very high-resolution images (Cotrufo et al., 2018). Some studies have incorporated 3D point-cloud information into analyses (Cusicanqui et al., 2018; Vetrivel et al., 2018). While these approaches have shown potential, and could potentially be used to provide additional granularity to our damage states, we opted against integrating point cloud analyses into our model, due to the considerably longer processing times associated with such an approach. Longer processing times would undermine the swift processing requirement inherent in our methodology.

5.6 Caveats

During the assignment of building damage states, uncertainties arose, particularly concerning the interpretation of tarpaulins and, pre-existing damage. For tarpaulins, the ambiguity arose from whether these were either strategically placed prior to the eruption as preventative measures to cause tephra to slide off the roof more easily; or they were placed post event to cover damage caused by tephra fall. Additionally, in certain instances, distinguishing between a collapsed roof and a section of the building initially lacking roofing material—possibly functioning as a walled storage area —proved challenging. Pre-existing damage not related to volcanic activity or buildings that were under construction at the time of image acquisition were considered as damaged and classified accordingly. The presence of buildings under construction at the time of image acquisition has been recognised as a challenge in studies using mono-temporal imagery (Nex et al., 2019; Cheng et al., 2021). Pre-event imagery would have provided clarity on both of these matters, however this was not available at high enough resolution for this region.

The majority of images used for training and evaluating our models came from the SRC dataset, which was collected several months after the eruption. As a result, the majority of images do not have much tephra present. In an operational context, to expedite the recovery process, data would ideally be collected as quickly after the eruption as it is safe to do so, therefore more tephra would be present in the images. Given the compound effects of variations in flight angle, image lighting, resolution and also the presence of tephra, we do not have enough information to test the effect of tephra thickness on model performance, and caution should be taken when using the model on data collected at different times after the eruption.

Deleted: We developed our approach using deep learning on 2D optical imagery, while s

Deleted: used

Deleted: , or combined point cloud information with deep learning on optical imagery for damage level classification (Vetrivel et al., 2018).

Deleted: the use of 3D spatial data has shown potential, andpotential

Deleted: may

Deleted: . This decision was motivated by

Deleted: ,

Deleted: which

Deleted: ¶

Deleted: by several authors

Deleted: result

6 Recommendations for UAV building damage assessment data collection

In the future we advocate for the adoption of a standardised protocol for data collection for the purpose of UAV damage assessment. While our model was developed using a diverse dataset, there were some disparities in performance across distinct data types. Consequently, the standardisation of image collection serves two purposes, 1) to allow the best results to be achieved when implementing our models, and 2) to collect data that is rich in information useful for damage assessment with the aim of working towards the development of global datasets for tephra fall damage. For best results we have the following recommendations:

- The bulk of our dataset was collected several months after the eruption of La Soufrière however, for generating a global dataset that can be used for response and recovery, models should ideally be trained on images collected shortly (days to weeks) after an event.
- Flight paths should be pre-programmed to ensure comprehensive coverage of the area and limit bias associated with overrepresentation of certain buildings. Ideally two flights would be conducted with two sets of perpendicular flight lines to capture buildings from a different perspective. GPS positioning should be enabled.
- A fixed altitude of 50-80 m above the ground should be maintained where possible. This is appropriate to capture sufficient data for accurate damage classification based on the established framework and strikes a balance between detailed information capture and overall coverage. In mountainous areas this may not be achievable for some UAV types. In which case a uniform height should be maintained such that the size of buildings is consistent across image frames.
- We suggest a slightly off-nadir camera positioning (~5-15°), which is sufficient to capture any bending in the roof that may not be captured from a nadir perspective.
- Overlap between images should be enough to generate orthoimages, 80% forward and 70% lateral overlap is sufficient.

In addition to the development of optimum post-event data collection practises we advocate for the collection of pre-event UAV datasets. Ideally, pre- and post-event imagery is collected using the same flight paths, altitudes, and camera positioning. Pre-event datasets serve multiple purposes:

Deleted: size

- Facilitates the creation of building inventories.
- Enables precise comparison of pre- and post-event imagery, reducing uncertainty regarding initial building conditions.
- Supports the development of high-resolution change detection models potentially yielding more accurate results than relying solely on post-event imagery.
- Provides an opportunity for UAV pilots to gain experience in capturing building datasets during ‘quiet times’.

7 Conclusions

Following a large tephra fall event, building damage assessment needs to be conducted rapidly for the purpose of response and recovery, and for the collection of data that can be used to forecast building damage from future events. By leveraging post-event optical imagery obtained after the 2021 eruption of La Soufrière volcano on the island of St Vincent, and convolutional neural networks, we have developed an automated tephra fall building damage assessment pipeline. The pipeline incorporates models for building localisation and two distinct levels of damage classification: distinguishing between No damage to minor damage and damage, as well as between Moderate and Major damage, which were trained and evaluated separately. When provided with UAV optical imagery, our pipeline can rapidly generate spatial building damage information. Our models perform well for the St Vincent datasets and are anticipated to perform well in locations where building typologies are similar, but this requires more testing to understand the limits of their application.

Building localisation model cross validation results underscore the influence of factors such as UAV altitude, off-nadir angles, tephra thickness, and training sample sizes on model performance, while results show that damage classification models were affected by these factors to a lesser extent. We acknowledge the challenges posed by diverse datasets and by limited data, and we propose a series of recommendations to guide the collection of future UAV building damage datasets. In addition to the collection of post-event datasets we advocate for the collection and incorporation of pre-event datasets, which can be used to support the advancement of change detection models; to partially evaluate the models presented here during quiescent times, and to develop building inventories that can be used along with fragility functions for forecasting building damage.

Deleted: building

Our research marks a step forward in tephra fall building damage assessment, offering a versatile and effective pipeline with the potential for regional applicability. As the field of UAV-based damage assessment in volcanology continues to evolve, our work lays a foundation for further advancements, contributing to the resilience of communities in the face of volcanic eruptions.

8 Author contributions

Conceptualization: SFJ, RR, ET, VM. Data collection: RR and VM. Development of the methodology: ET, SFJ, BW. Software: ET. Formal analysis: ET. Supervision: SFJ. Writing – original draft: ET. Writing-Reviewing & Editing: ET, SFJ, VM, RR, BW, BT, SHY.

9 Competing interests

The authors declare no competing interests.

10 Acknowledgements

We are indebted to Monique Johnson: The UWI Seismic Research Centre, Javid Collins: UWITV, Nikolai Lewis and Marla Mulraine: The Government of St Vincent and the Grenadines Ministry of Transport, Works, Lands and Surveys, and Physical Planning, for sharing their UAV data and collaborating on this work. All images and data provided in this study have been approved for publication by the local agency responsible for monitoring geohazards in St Vincent: The UWI Seismic Research Centre. We are very grateful to Chee Jain Hao Denny, Sim Yu Yang, Isaiah Loh Kai En, Huang Wanxin for their assistance with data preparation, and to Vanesa Burgos, Elinor Meredith, Alberto Ardid, and Tom Wilson, for interesting discussions around machine learning and building damage assessment. We would like to thank Sébastien Biass and one anonymous reviewer for their detailed and constructive reviews that considerably improved the manuscript and, Giovanni Macedonio for their editorial handling.

Deleted: and

11 Data availability

All trained models along with the code required to execute the damage assessment pipeline and instructions for usage are provided at: https://github.com/EllyTennant/UAVdamageAssessment

12 Funding

This research was supported by the Earth Observatory of Singapore via its funding from the National Research Foundation Singapore and the Singapore Ministry of Education under the Research Centres of Excellence initiative and comprises EOS contribution number 596. Additional support was provided by the AXA Research Fund as part of the Joint Research Initiative on Volcanic Risk in Asia.

13 References

An, G., Akiba, M., Omodaka, K., Nakazawa, T., & Yokota, H. (2021). Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-83503-7

Andaru, R. and Rau, J.Y. 2019. Lava dome changes detection at agung mountain during high level of volcanic activity using uav photogrammetry. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. International Society for Photogrammetry and Remote Sensing, pp. 173–179. doi: 10.5194/isprs-archives-XLII-2-W13-173-2019.

Anniballe, R., Noto, F., Scalia, T., Bignami, C., Stramondo, S., Chini, M. and Pierdicca, N. 2018. Earthquake damage mapping: An overall assessment of ground surveys and VHR image change detection after L'Aquila 2009 earthquake. *Remote Sensing of Environment* 210, pp. 166–178. doi: 10.1016/j.rse.2018.03.004.

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. In Neural Networks and Deep Learning. https://doi.org/10.1007/978-3-319-94463-0

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, *79*(1–2), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Biass, S., Bonadonna, C., & Houghton, B. F. 2019. A step-by-step evaluation of empirical methods to quantify eruption source parameters from tephra-fall deposits. Journal of Applied Volcanology, 8(1). https://doi.org/10.1186/s13617-018-0081-1

Biass, S., Jenkins, S., Lallemant, D., Lim, T.N., Williams, G. and Yun, S.H., 2021. Remote sensing of volcanic impacts. In Forecasting and Planning for Volcanic Hazards, Risks, and Disasters (pp. 473-491). Elsevier.

Biass, S., Reyes-Hardy, M. P., Gregg, C., di Maio, L. S., Dominguez, L., Frischknecht, C., Bonadonna, C., & Perez, N. 2024. The spatiotemporal evolution of compound impacts from lava flow and tephra fallout on buildings: lessons from the 2021 Tajogaite eruption (La Palma, Spain). Bulletin of Volcanology, 86(2). https://doi.org/10.1007/s00445-023-01700-w

Blong, R. 2003a. *A Review of Damage Intensity Scales*. Available at: http://www.es.mq.edu.au/NHRC/web/scales/scalesindes.htm.

Blong, R. 2003b. Building damage in Rabaul, Papua New Guinea, 1994. *Bulletin of Volcanology* 65(1), pp. 43–54. doi: 10.1007/s00445-002-0238-x.

Formatted: Font: Cambria, 11 pt

Bouchard, I., Rancourt, M.È., Aloise, D. and Kalaitzis, F. 2022. On Transfer Learning for Building Damage Assessment from Satellite Imagery in Emergency Contexts. *Remote Sensing* 14(11), pp. 1–29. doi: 10.3390/rs14112532.
Bruzzone, L. and Fernàndez Prieto, D. 2000. Automatic Analysis of the Difference Image for Unsupervised Change Detection. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING* 38(3), pp. 1171–1181.
Cheng, C.S., Behzadan, A.H. and Noshadravan, A. 2021. Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering* 36(6), pp. 695–710. doi: 10.1111/mice.12658.
Cole, P.D. et al. 2023. Explosive sequence of La Soufrière, St Vincent, April 2021: insights into drivers and consequences via eruptive products. Available at: https://doi.org/10.6084/m9.figshare.c.6474317.
Cotrufo, S., Sandu, C., Giulio Tonolo, F., & Boccardo, P. (2018). Building damage assessment scale tailored to remote sensing vertical imagery. European Journal of Remote Sensing, 51(1), 991–1005. https://doi.org/10.1080/22797254.2018.1527662
Cusicanqui, J., Kerle, N., & Nex, F. 2018. Usability of aerial video footage for 3-D scene reconstruction and structural damage assessment. Natural Hazards and Earth System Sciences, 18(6), 1583–1598. https://doi.org/10.5194/nhess-18-1583-2018
Deligne, N.I., Jenkins, S.F., Meredith, E.S., Williams, G.T., Leonard, G.S., Stewart, C., Wilson, T.M., Biass, S., Blake, D.M., Blong, R.J. and Bonadonna, C., 2022. From anecdotes to quantification: advances in characterizing volcanic eruption impacts on the built environment. Bulletin of Volcanology, 84(1), p.7.
Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255.
Duarte, D., Nex, F., Kerle, N. and Vosselman, G. 2020. Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Riva del Garda, Italy, pp. 4–7. Available at: https://research.utwente.nl/en/publications/satellite-image-classification-of-building-damages-using-airborne.
Dung Cao, Q. and Choe, Y. 2020. *Building Damage Annotation on Post-Hurricane Satellite Imagery Based on Convolutional Neural Networks*.
Gailler, L., Labazuy, P., Régis, E., Bontemps, M., Souriot, T., Bacques, G. and Carton, B. 2021. Validation of a new UAV magnetic prospecting tool for volcano monitoring and geohazard assessment. *Remote Sensing* 13(5), pp. 1–10. doi: 10.3390/rs13050894.
Galanis, M., Rao, K., Yao, X., Tsai, Y.L., Ventura, J. and Fricker, G.A. 2021. DamageMap: A post-wildfire damaged buildings classifier. *International Journal of Disaster Risk Reduction* 65. doi: 10.1016/j.ijdrr.2021.102540.
Ghosh, S. et al. 2011. Crowdsourcing for rapid damage assessment: The global earth observation catastrophe assessment network (GEO-CAN). *Earthquake Spectra* 27(SUPPL. 1). doi: 10.1193/1.3636416.
Girshick, R. (2015). Fast R-CNN. http://arxiv.org/abs/1504.08083
Gonzalez, D., Rueda-Plata, D., Acevedo, A. B., Duque, J. C., Ramos-Pollán, R., Betancourt, A., & García, S. (2020). Automatic detection of building typology using deep learning methods on street level images. Building and Environment, 177. https://doi.org/10.1016/j.buildenv.2020.106805
Gupta, R. and Shah, M. 2020. RescueNet: Joint building segmentation and damage assessment from satellite imagery. In: *Proceedings - International Conference on Pattern Recognition*. Institute of Electrical and Electronics Engineers Inc., pp. 4405–4411. doi: 10.1109/ICPR48806.2021.9412295.
Hayes, J., Wilson, T. M., Deligne, N. I., Cole, J., & Hughes, M. 2017. A model to assess tephra clean-up requirements in urban environments. Journal of Applied Volcanology, 6(1). https://doi.org/10.1186/s13617-016-0052-3
Hayes, J.L. et al. 2019. Timber-framed building damage from tephra fall and lahar: 2015 Calbuco eruption, Chile. *Journal of Volcanology and Geothermal Research* 374(October 2015), pp. 142–159. Available at: https://doi.org/10.1016/j.jvolgeores.2019.02.017.

He, K., Zhang, X., Ren, S., & Sun, J. 2015. Deep Residual Learning for Image Recognition. http://arxiv.org/abs/1512.03385
Iannelli, G., & Dell'Acqua, F. (2017). Extensive Exposure Mapping in Urban Areas through Deep Analysis of Street-Level Pictures for Floor Count Determination. Urban Science, 1(2), 16. https://doi.org/10.3390/urbansci1020016
Ishii, M., Goto, T., Sugiyama, T., Saji, H. and Abe, K. 2002. Detection of Earthquake Damaged Areas from Aerial Photographs by Using Color and Edge Information. pp. 23–25.
Jenkins, S., & Spence, R. 2009. Vulnerability curves for buildings and agriculture The MIAVITA project is financed by the European Commission under the 7th Framework Programme for Research and Technological Development, Area "Environment", Activity 6.1 "Climate Change, Pollution and Risks."
Jenkins, S.F., McSporran, A., Wilson, T.M., Stewart, C.S., Leonard, G.A., Cevuard, S., Garaebiti, E., In preparation. Tephra fall impacts to buildings: The 2017-2018 Manaro Voui eruption, Vanuatu. Journal of Volcanology and Geothermal Research
Jenkins, S., Komorowski, J.C., Baxter, P.J., Spence, R., Picquout, A., Lavigne, F. and Surono. 2013. The Merapi 2010 eruption: An interdisciplinary impact assessment methodology for studying pyroclastic density current dynamics. *Journal of Volcanology and Geothermal Research* 261, pp. 316–329. Available at: http://dx.doi.org/10.1016/j.jvolgeores.2013.02.012.
Jenkins, S. F., Spence, R. J. S., Fonseca, J. F. B. D., Solidum, R. U., & Wilson, T. M. 2014. Volcanic risk assessment: Quantifying physical vulnerability in the built environment. Journal of Volcanology and Geothermal Research, 276, pp 105–120. https://doi.org/10.1016/j.jvolgeores.2014.03.002
Jenkins, S.F., Phillips, J.C., Price, R., Feloy, K., Baxter, P.J., Hadmoko, D.S. and de Bélizal, E. 2015. Developing building-damage scales for lahars: application to Merapi volcano, Indonesia. *Bulletin of Volcanology* 77(9). doi: 10.1007/s00445-015-0961-8.
Johnson, J.M. and Khoshgoftaar, T.M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data* 6(1). doi: 10.1186/s40537-019-0192-5.
Joseph, E.P. et al. 2022. Responding to eruptive transitions during the 2020–2021 eruption of La Soufrière volcano, St. Vincent. *Nature Communications* 13(1). doi: 10.1038/s41467-022-31901-4.
Jung, J., Kim, D. J., Lavalle, M., & Yun, S. H. (2016). Coherent Change Detection Using InSAR Temporal Decorrelation Model: A Case Study for Volcanic Ash Detection. IEEE Transactions on Geoscience and Remote Sensing, 54(10), 5765–5775. https://doi.org/10.1109/TGRS.2016.2572166
Karnik,V. Schenkov, Z, Schenk, V. 1984. Vulnerability and the MSK scale. Engineering Geology, 20 (1984) pp161-168
Kerle, N., Nex, F., Gerke, M., Duarte, D. and Vetrivel, A. 2019. UAV-based structural damage mapping: A review. *ISPRS International Journal of Geo-Information* 9(1), pp. 1–23. doi: 10.3390/ijgi9010014.
Khajwal, A.B., Cheng, C.S. and Noshadravan, A. 2023. Post-disaster damage classification based on deep multi-view image fusion. *Computer-Aided Civil and Infrastructure Engineering* 38(4), pp. 528–544. doi: 10.1111/mice.12890.
Lerner, G.A. et al. 2021. The hazards of unconfined pyroclastic density currents : a new synthesis and classification according to their deposits , dynamics , and thermal and impact This manuscript is a non-peer reviewed preprint submitted to Journal of Volcanology and Geothermal . pp. 1–48.
López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., & García-Martín, Á. (2019). Semantic-Aware Scene Recognition. https://doi.org/10.1016/j.patcog.2020.107256
Li, S., Tang, H., He, S., Shu, Y., Mao, T., Li, J. and Xu, Z. 2015. Unsupervised Detection of Earthquake-Triggered Roof-Holes from UAV Images Using Joint Color and Shape Features. *IEEE Geoscience and Remote Sensing Letters* 12(9), pp. 1823–1827. doi: 10.1109/LGRS.2015.2429894.
Li, Y., Hu, W., Dong, H. and Zhang, X. 2019a. Building damage detection from post-event aerial imagery using single shot multibox detector. *Applied Sciences (Switzerland)* 9(6). doi: 10.3390/app9061128.

Li, D., Cong, A., & Guo, S. 2019b. Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification. Automation in Construction, 101, 199–208. https://doi.org/10.1016/j.autcon.2019.01.017
Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. 2014. Microsoft COCO: Common Objects in Context. http://arxiv.org/abs/1405.0312
Lucks, L., Bulatov, D., Thönnessen, U. and Böge, M. 2019. Superpixel-wise assessment of building damage from aerial images. In: *VISIGRAPP 2019 - Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SciTePress, pp. 211–220. doi: 10.5220/0007253802110220.
Noh, H. Y., Jaiswal, K. S., Engler, D., & Wald, D. J. (2020). An efficient Bayesian framework for updating PAGER loss estimates. Earthquake Spectra, 36(4), 1719–1742. https://doi.org/10.1177/8755293020944177
Meng, S., Soleimani-Babakamali, M. H., & Taciroglu, E. 2023. Automatic Roof Type Classification Through Machine Learning for Regional Wind Risk Assessment. http://arxiv.org/abs/2305.17315
Meredith, E.S., Jenkins, S.F., Hayes, J.L., Deligne, N.I., Lallemant, D., Patrick, M. and Neal, C. 2022. Damage assessment for the 2018 lower East Rift Zone lava flows of Kīlauea volcano, Hawai'i. *Bulletin of Volcanology* 84(7). doi: 10.1007/s00445-022-01568-2.
Moradi, M. and Shah-Hosseini, R. 2020. Earthquake Damage Assessment Based on Deep Learning Method Using VHR Images. *Environmental Sciences Proceedings* 5(1), p. 16. doi: 10.3390/iecg2020-08545.
Naito, S. et al. 2020. Building-damage detection method based on machine learning utilizing aerial photographs of the Kumamoto earthquake. *Earthquake Spectra* 36(3), pp. 1166–1187. doi: 10.1177/8755293019901309.
Nex, F., Duarte, D., Steenbeek, A. and Kerle, N. 2019. Towards real-time building damage mapping with low-cost UAV solutions. *Remote Sensing* 11(3), pp. 1–14. doi: 10.3390/rs11030287.
Novikov, G., Trekin, A., Potapov, G., Ignatiev, V. and Burnaev, E. 2018. Satellite imagery analysis for operational damage assessment in emergency situations. In: *Lecture Notes in Business Information Processing*. Springer Verlag, pp. 347–358. doi: 10.1007/978-3-319-93931-5_25.
Post Disaster Needs Assessment (PDNA). 2022. St Vincent and the Grenadines
Pomonis, A. A., Spence, R., & Baxter, P.1999. Risk assessment of residential buildings for an eruption of Furnas Volcano, Sao Miguel, the Azores. Journal of Volcanology and Geothermal Research, 92, pp 107-131.
Pi, Y., Nath, N.D. and Behzadan, A.H. 2020. Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. *Advanced Engineering Informatics* 43. doi: 10.1016/j.aei.2019.101009.
Ren, S., He, K., Girshick, R. and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6), pp. 1137–1149. doi: 10.1109/TPAMI.2016.2577031.
Román, A., Tovar-Sánchez, A., Roque-Atienza, D., Huertas, I.E., Caballero, I., Fraile-Nuez, E. and Navarro, G. 2022. Unmanned aerial vehicles (UAVs) as a tool for hazard assessment: The 2021 eruption of Cumbre Vieja volcano, La Palma Island (Spain). *Science of the Total Environment* 843. doi: 10.1016/j.scitotenv.2022.157092.
Shen, Y. et al. 2021. BDANet: Multiscale Convolutional Neural Network with Cross-directional Attention for Building Damage Assessment from Satellite Images. Available at: http://arxiv.org/abs/2105.07364.
Singh, D. K., & Hoskere, V. 2023. Post Disaster Damage Assessment Using Ultra-High-Resolution Aerial Imagery with Semi-Supervised Transformers. Sensors, 23(19). https://doi.org/10.3390/s23198235
Spence,R.J.S.,Pomonis,A.,Baxter,P.J.,Coburn,A.W.,White,M.,Dayrit,M.,and Field Epidemiology Training Program Team. 1996.Building Damage Caused by the Mount Pinatubo Eruption of 15 June 1991, in: FireandMud: Eruptions and Lahars of Mount Pinatubo,Philippines, edited by: Newhall,C.G. and Punongbayan, R. S., University of Washington Press, London, UK, 1055–1061

Deleted: ¶

Spence, R., Martínez-Cuevas, S. and Baker, H. 2021. Fragility estimation for global building classes using analysis of the Cambridge earthquake damage database (CEQID). *Bulletin of Earthquake Engineering* 19(14), pp. 5897–5916. doi: 10.1007/s10518-021-01178-x.

Spence, R.J.S., Kelman, I., Baxter, P.J., Zuccaro, G. and Petrazzuoli, S. 2005. *Natural Hazards and Earth System Sciences Residential building and occupant vulnerability to tephra fall*.

St Vincent and the Grenadines population and housing census, 2012

Szegedy, C., Vanhoucke, V., Ioffe, S., & Shlens, J. 2015. Rethinking the Inception Architecture for Computer Vision.

Valentijn, T., Margutti, J., van den Homberg, M., & Laaksonen, J. (2020). Multi-hazard and spatial transferability of a CNN for automated building damage assessment. *Remote Sensing*, *12*(17), 1–29. https://doi.org/10.3390/rs12172839

Vetrivel, A., Gerke, M., Kerle, N., Nex, F., & 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 Journal of Photogrammetry and Remote Sensing, 140, 45–59. https://doi.org/10.1016/j.isprsjprs.2017.03.001

Wang, Z., Zhang, F., Wu, C., & Xia, J.(2024. Rapid mapping of volcanic eruption building damage: A model based on prior knowledge and few-shot fine-tuning. International Journal of Applied Earth Observation and Geoinformation, 126. https://doi.org/10.1016/j.jag.2023.103622

Weber, E. and Kané, H. 2020. Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion. Available at: http://arxiv.org/abs/2004.05525.

Williams, G.T., Jenkins, S.F., Biass, S., Wibowo, H.E. and Harijoko, A. 2020. Remotely assessing tephra fall building damage and vulnerability: Kelud Volcano, Indonesia. *Journal of Applied Volcanology* 9(1), pp. 1–18. doi: 10.1186/s13617-020-00100-5.

Wilson, G., Wilson, T.M., Deligne, N.I. and Cole, J.W. 2014. Volcanic hazard impacts to critical infrastructure: A review. *Journal of Volcanology and Geothermal Research* 286, pp. 148–182. Available at: http://dx.doi.org/10.1016/j.jvolgeores.2014.08.030.

Xu, J.Z., Lu, W., Li, Z., Khaitan, P. and Zaytseva, V. 2019. Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks. (NeurIPS). Available at: http://arxiv.org/abs/1910.06444.

Yi, W., Sun, Y., & He, S. 2018. Data Augmentation Using Conditional GANs for Facial Emotion Recognition. Progress In Electromagnetics Research Symposium. Japan. 1-4 August.

Yorioka, D., Kang, H., Iwamura,K. 2020. Data Augmentation For Deep Learning Using Generative Adversarial Networks. IEEE 9th Global Conference on Consumer Electronics (GCCE)

Yun, S.H. et al. 2015. Rapid damage mapping for the 2015 Mw 7.8 Gorkha Earthquake Using synthetic aperture radar data from COSMO-SkyMed and ALOS-2 satellites. *Seismological Research Letters* 86(6), pp. 1549–1556. doi: 10.1785/0220150152.

Zhang, J.F., Xie, L.L. and Tao, X.X. 2003. Change Detection of Earthquake-damaged Buildings on Remote Sensing Image and its Application in Seismic Disaster Assessment. In: *International Geoscience and Remote Sensing Symposium (IGARSS)*. pp. 2436–2438. doi: 10.1109/igarss.2003.1294467.

Zou, Z., Shi, Z., Guo, Y. and Ye, J. 2019. Object Detection in 20 Years: A Survey. Available at: http://arxiv.org/abs/1905.05055.

Formatted: Font: Cambria, 11 pt

Deleted: ¶