We sincerely thank reviewer 2 for their comments and for the time taken to review the manuscript. We have addressed all comments and provide our response below. Reviewer comments are in black, our responses are in green, and excerpts from the revised manuscript are in blue.

1. The manuscript is too focused on methodology. I understand that it represents the heart of the work, but the text seems too unbalanced in relation to the results and the chosen volcanic application (the 2021 eruption of the La Soufrière volcano, St Vincent and the Grenadines). I suggest lightening Sections 2 and 3, moving even more details into the supplementary material and simplifying the main text. This would definitely make reading faster and more fluent.

    - We have significantly reworked the manuscript reducing the focus on methodology by removing some of the details in Section 2 to the supplementary material as suggested.

        - The following description of the UAV labelling has been moved to the supplementary:
          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., 2024). 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.

        - All of the description of the sieve network has been moved to the supplementary:
          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. Bounding boxes produced by the detector are passed to the sieve network to filter out detections that are false positives. A false positive occurs 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).

          The dataset used for training and evaluating the sieve network consists of randomly cropped background samples from full sized images in the training and validation sets. Samples were cropped from each of the datasets, and samples containing buildings were removed until 100 no-building samples were achieved for each dataset. These samples were supplemented with an additional 10% targeted image samples on the observation that trained detectors

were mistakenly detecting cars and boats. For the building dataset we stochastically sampled the equivalent number (n=990 train, 660 validation) from the building images. Experiments for the sieve network were conducted using two different CNN architectures (ResNet50 and GoogleNet), and by undertaking a grid search to find the best hyperparameter combination (learning rate, batch size, and L2 regularisation). A total of five experiments were conducted, each consisting of three replicates.

  - The reference in the main manuscript to the sieve network now reads:
    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 the supplementary material.

- The following text regarding details of the cross validation has been moved to the supplementary:
  - The full image set consists of images collected by three different parties across 13 different locations on the island. To test the robustness of our models to location, we trained on nine out of the ten locations present in the combined training and validation sets and evaluated each model's performance on the remaining location. To test the robustness to the dataset, we trained models and evaluated the performance for each of the three locations that contain images from more than one dataset (e.g., Chateaubelair-GOV, Chateaubelair-UWI-TV, Chateaubelair-SRC) separately.

  - The caption of Figure 5 contains sufficient information to understand the process of cross validation:

    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.

- We have shortened the description of the model evaluation metrics which now reads:
  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 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.

- We have moved details of the faster RCNN detector to the supplementary, the new text is now significantly reduced and reads:

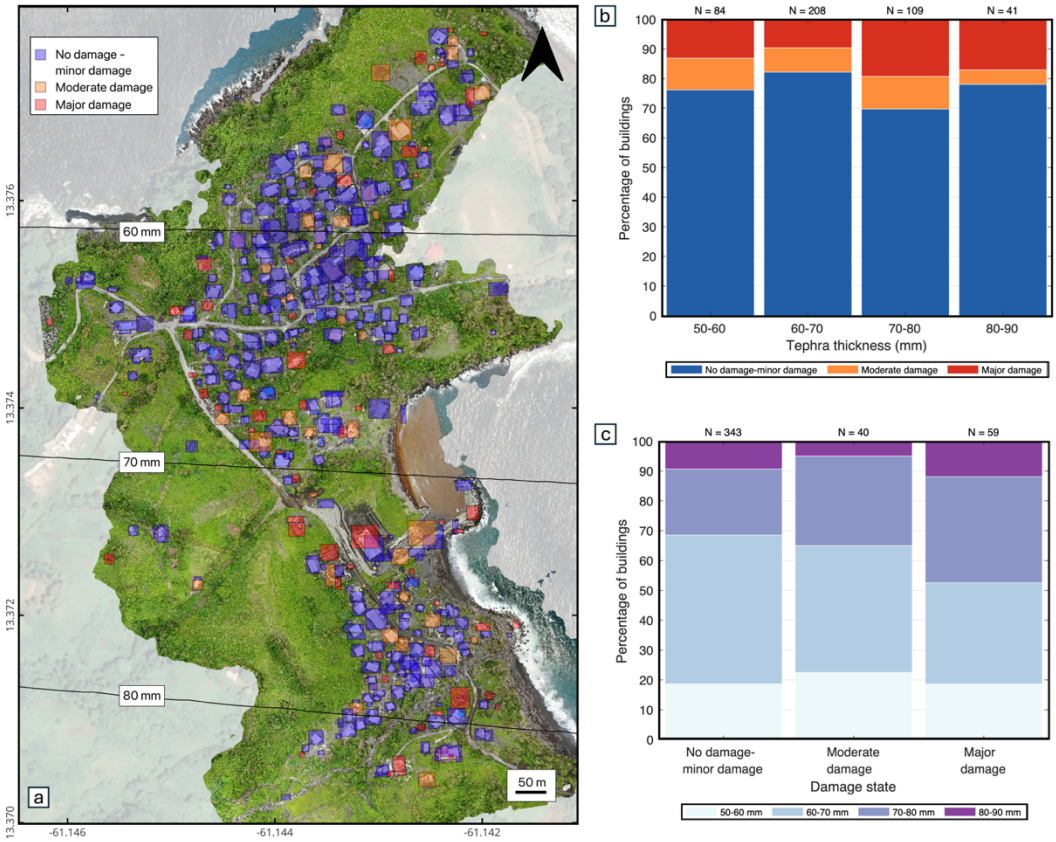
  For building localisation, we used the cutting edge two-stage object detector Faster R-CNN (Ren et al. 2017). When applied to a test image containing the relevant objects, Faster R-CNN outputs the positions within the image (X, Y, width, and height in pixels) of bounding boxes containing the object, 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 (see supplementary material for details).

  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 the supplementary material.

- In addition to the aforementioned reworking of the text and in response to comments from reviewer 1, we have elaborated on the case study. This further improves the balance between methodology and damage assessment results making the manuscript more appealing to a wider audience. To do this we interpolated between the tephra isopachs to extract tephra thicknesses for each building. The results of this are described in Section 4 (below), and discussed in the context of physical impacts to buildings in Section 5.3 (below). We have adapted Figure 7 to reflect this.



Section 4 now reads:

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 pipeline output georeferenced vector 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 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.

These results are discussed in Section 5.3:

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. 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 building typology and post-event impact data 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.

2. The location of Table 1 cannot be the Introduction. It provides a performance comparison of several models, including the one described in this manuscript, and should therefore be included in the Discussions. It is not logically correct to introduce F1, mean average precision and accuracy scores before even introducing the model. It is also not immediately clear what "P", "P & P", "C1" and "C2" mean.

    - We prefer to keep Table 1 in the introduction since we believe it adds important context to the points discussed in this section in particular by showing the types of hazards that have been considered, the datatypes used and the use of pre-disaster imagery in past studies. However in line with the reviewers comments and to ensure that this is the appropriate location we have made some adjustments to the table:
        - We have removed our results from this table meaning that C1 and C2 are no-longer referred to.
        - We have changed the header of column 4 from 'Pre and Post' to 'Pre-disaster imagery'. The contents of this column are now either 'Yes' or 'No' as opposed to P&P or P.
        - We have added the following text to the tables caption : *A detailed explanation of the scores used for evaluation is provided in Section 2.3.3.*

      With these adaptations we believe that the tables position within the introduction is now appropriate.

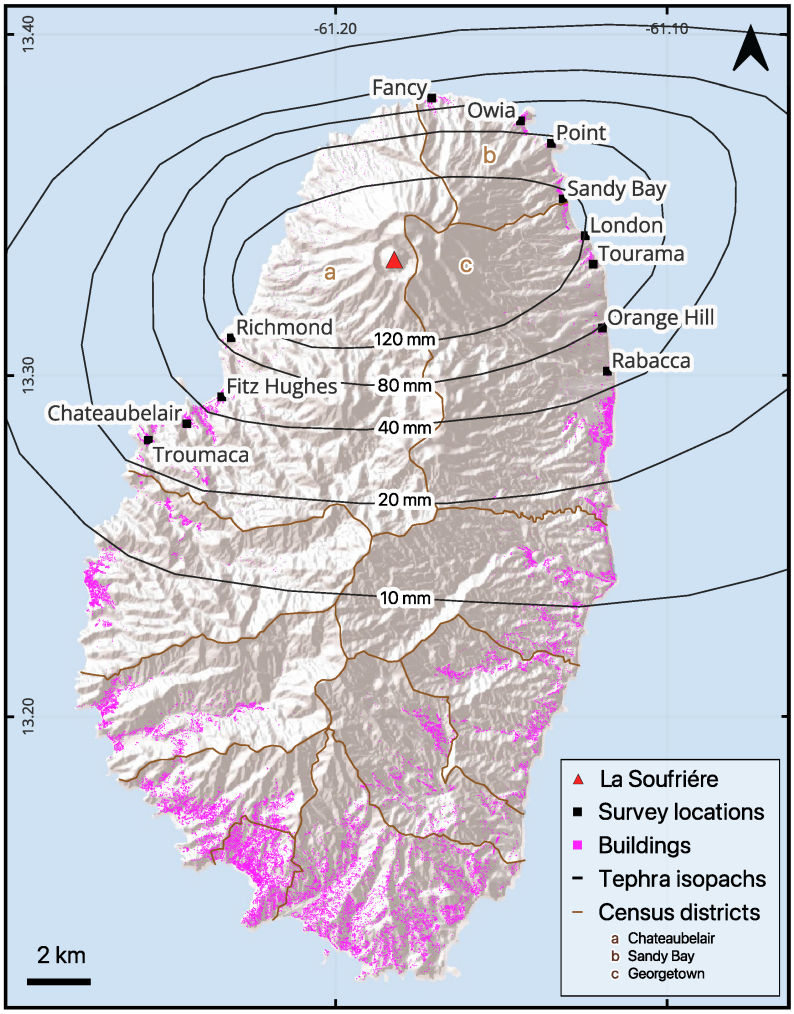
Table 1. A non-exhaustive list of works using deep learning on optical imagery for building damage assessment. Studies use different scores to evaluate performance: F1 scores are in italics, mean average precision scores are underlined, accuracy scores in

***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*
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*

3. In Figure 1, please include the location of Georgetown.

    - The original caption for Figure 1 stated that Georgetown refers to the district of Georgetown, this is located in the NE of the island and marked with the letter ‘c’. For added clarity we have put items a-c into the maps legend and removed these from the caption.
      New Figure 1:



4. The text is full of diagrams and tables. Perhaps it would be more attractive to a wider audience if the authors included some figures on the case study (for example, some images used for the model development, currently in the supplementary material).
    - In agreement with the local agency responsible for monitoring hazards at St Vincent (The University of the West Indies, Seismic Research Centre), to respect the privacy of the residents of St Vincent we did not include images of residential buildings. However, in response to reviewer 1s suggestion, we have added an additional figure into the methods which shows representative examples of the different damage states which we believe makes the study more appealing to a wider audience. Buildings shown in this figure were carefully selected for anonymity, with government or public buildings shown where possible.

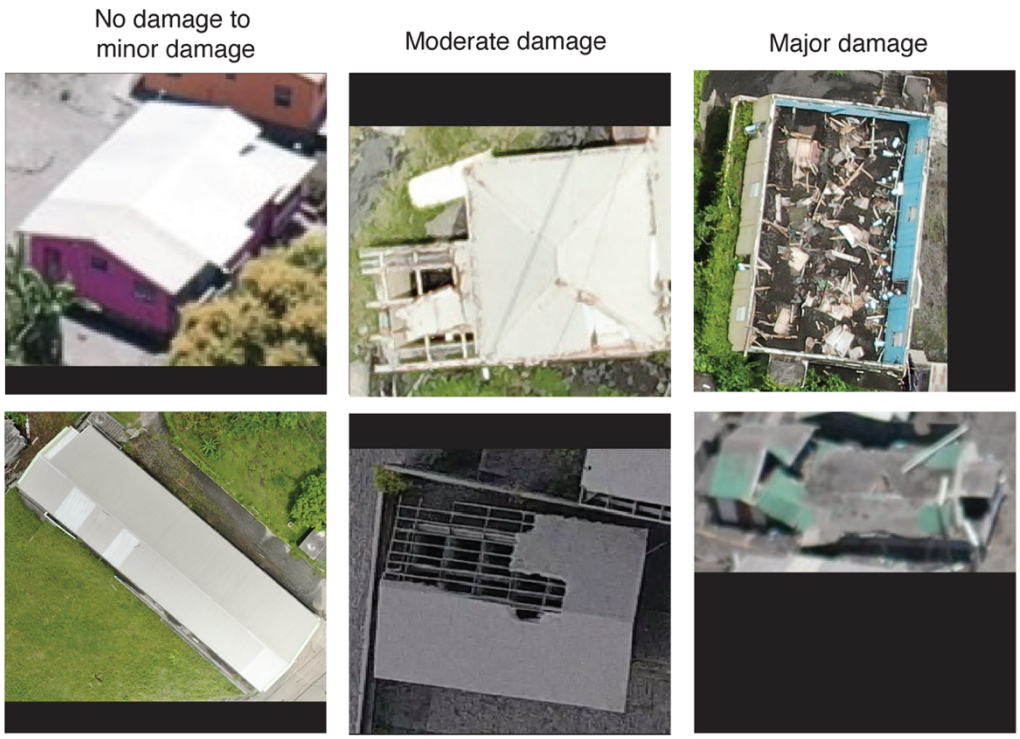


Figure 2. Example of the three damage states used in this work: No damage to minor damage, Moderate damage and, Major damage.