Disaster mapping from satellites: damage detection with crowdsourced point labels

Danil Kuzin
Department of Computer Science
Department of Physics
Lancaster University, Lancaster, UK
d.kuzin@lancaster.ac.uk

Olga Isupova
Department of Computer Science
University of Bath, Bath, UK
oi260@bath.ac.uk

Brooke D. Simmons
Department of Physics
Lancaster University, Lancaster, UK
b.simmons@lancaster.ac.uk

Steven Reece
Department of Engineering Science
University of Oxford, Oxford, UK
reece@robots.ox.ac.uk

Abstract

High-resolution satellite imagery available immediately after disaster events is crucial for response planning as it facilitates broad situational awareness of critical infrastructure status such as building damage, flooding, and obstructions to access routes. Damage mapping at this scale would require hundreds of expert person-hours. However, a combination of crowdsourcing and recent advances in deep learning reduces the effort needed to just a few hours in real time. Asking volunteers to place point marks, as opposed to shapes of actual damaged areas, significantly decreases the required analysis time for response during the disaster. However, different volunteers may be inconsistent in their marking. This work presents methods for aggregating potentially inconsistent damage marks to train a neural network damage detector.

1 Introduction

Maps of damage and the health of infrastructure are key urgent needs of responders and decision makers in the immediate aftermath of a crisis. Satellite imagery is useful for datasets mapping as it can cover large areas rapidly and uniformly at high enough resolution to identify major damage. Machine learning approaches promise a rapid assessment of arbitrarily large amounts of satellite data via assessment of changes between pre- and post-event images, but disaster mapping in particular has many challenges, including weather conditions, low resolution of images, diversity of environments, variable temporal intervals between collection of images, difference in projections due to movement of satellites. These factors pose major obstacles to out of the box machine learning algorithms and existence of universally applicable datasets.

Typical approaches to damaged building detection from satellite imagery consist of two stages: localization of building footprints and classification of the damage severity for each of these footprints based on pre- and post-event imagery. The current state-of-the-art approaches for both of these stages use neural networks of various architectures with different approaches to the combination of pre- and post-event images. The localization of building footprints can be viewed as an object detection problem of predicting the bounding box for each building; or as an instance segmentation problem of predicting the pixel mask for each building. Both object detection and segmentation approaches have been used for building footprint localization including SSD [Li et al., 2019], R-CNN [Weber and Kané, 2020], and LinkNet [Golovanov et al., 2018]. Damage detection approaches combine

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

pre- and post-event imagery either on the neural network input layer by stacking together both optical images into a 6-channel image [Xu et al., 2019], or by processing each image via a separate feature extraction neural network [Weber and Kané, 2020] and concatenating features before performing detection. These were succeeded by cross region transfer learning [Xu et al., 2019] and simultaneous segmentation and classification with dilated residual networks [Gupta and Shah, 2020].

All the above models were trained on large, carefully-prepared, high-resolution datasets. Some of these datasets are available online and include building footprint datasets and damaged building datasets. However, careful data preparation is not always possible during time-critical disaster responses. Therefore, we focus our attention on mapping damage with free resources and in a timely manner, the crucial factors for a disaster response operation. This involves images captured under non-ideal conditions, at varying resolutions, and labeled with volunteer *point marks*. Point marks are single dot-like labels that we ask our volunteers to place on satellite imagery. They are much faster to collect than bounding boxes or segmented building footprints, and the speed is highly desirable during a live response deployment. These marks must then be aggregated to identify a consensus for each building as different volunteers may have different opinions about the same object. Our work is two-fold: we first create a training dataset of damage severity on building footprints using only the point marks from volunteers, and then demonstrate how this dataset can be used to train a neural network to map damages on new unseen data. The second step allows us to cover more areas to provide information for disaster responders without the need to rely on volunteers.

1.1 Application Context

In 2017 hurricanes Irma and Maria impacted multiple islands in the Caribbean. Disaster imagery was made available from multiple sources including NASA's Landsat 8, ESA's Sentinel-2, Planet's Dove (`planet.com`), RapidEye and SkySat constellations, and Maxar's (formerly DigitalGlobe) satellite constellation (`maxar.com`). We focus here on data from live deployments by the Planetary Response Network (PRN, `planetaryresponsenetwork.org`), a collaboration between Zooniverse crowdsourcing platform (`zooniverse.org`), response and resilience organizations, and machine learning researchers. We present a data processing pipeline for detecting damaged buildings, which we developed primarily using imagery and labels from PRN deployments in the Caribbean.

Previous work has addressed the identification of general damage levels in individual satellite image segments. Mapping individual building damage was requested by Rescue Global, a humanitarian organisation with which we regularly deploy following disasters. Motivated by these needs, we focus here on identifying damage to individual structures. This need is common in the humanitarian assistance and disaster response (HADR) domain as it directly feeds ongoing situational awareness for stakeholders.

Automated building damage detection from satellite imagery is a significant current research problem: Shen et al. [2020] fuse pre- and post-event images for damage detection, Lee et al. [2020] use semi-supervised learning to reduce the amount of training data required for damage detection, Benson and Ecker [2020] assess generalization error of damage detection models on new datasets, Boin et al. [2020] address the problem of class imbalance for damaged building detection, Xu et al. [2019] discuss different architectures for building damage assessment.

In this context our work aims to reduce the required time for mapping of training data during the disasters in new conditions, where the existing models need to be fine-tuned. This is achieved by using crowdsourced marks for labelling. Crowdsourcing has proven to be useful during our live deployments before when we carefully aggregated crowd labels to remove any inconsistencies between volunteers.

2 The Crowd-labelled Dataset

Our aim is to detect objects of an arbitrary shape, specifically building footprints, from marks that only point to the objects. In this section we describe how we create the dataset from the pre- and post-event satellite images and marked points of damage by the crowd. First, we detect building footprints on both images, using a segmentation neural network, trained on a building footprints dataset. After that, we extract bounding boxes for segmented buildings and aggregate marks of the crowd for every footprint.

Table 1: Prediction quality of bounding boxes for building footprint detection. Values are given in the percentage form.

Timestamp	AP^{50}	F_1	Precision	Recall
Pre-event	38	57	52	63
Post-event	23	47	44	49

2.1 Building Footprint Detection

There are several approaches for point supervision that we can apply to convert point marks into damaged building footprints. Bearman et al. [2016] use the pre-trained objectness prior that is trained on other datasets; it assigns probabilities that pixels belong to objects and these probabilities are included in the loss function. Papadopoulos et al. [2017] use points in the centers of the objects and estimate object sizes based on them. Points on the borders of the objects can be used to define the proposal object masks: Maninis et al. [2018] use extreme points for the objects, generate Gaussian probabilistic masks around the points, bound by an exterior of marked points and then generate masks for the objects to use as additional layers in a network; Zhou et al. [2019] use CornerNet and HourglassNet to detect corner points and center points for objects of each class, then group extreme points associated with centers, aggregate edges for extreme points, and extract octagon masks. Benenson et al. [2019] use corrective marks to specify areas that do or do not belong to objects, and create binary disks around these marks.

Instead of using marks specifically to determine shapes of objects, we ask volunteers to provide marks anywhere on damaged buildings to speed up the labelling process. In our case building footprints can be used for limiting the area of interest around these point marks. There exist multiple datasets to train a method for building footprint localisation, such as SpaceNet (`spacenet.ai`), Open Cities AI Challenge (`drivendata.org/competitions/60/building-segmentation-disaster-resilience`), DeepGlobe2018 [Demir et al., 2018] and damaged building datasets such as xView [Lam et al., 2018] and xBD [Gupta et al., 2019].

We use the xView2 dataset to train the ensemble of UNet meta-architectures [Ronneberger et al., 2015], that was used in the building localization solution for the xView2 challenge (`https://github.com/DIUx-xView/xView2_first_place`). The neural networks are used separately for pre- and post-event images to detect building footprints.

Satellite images are often obtained with different off-nadir angles at different days and therefore they do not align well for precise per-pixel analysis. To mitigate this problem we associate buildings between pre- and post-event images using their bounding boxes instead of segmentation masks. Segmentation masks are thresholded with the optimised value, then contours are extracted from binarised masks, converted to polygons, and their boundaries are used as bounding boxes. This approach is less sensitive to imagery misalignments and pixel-level precision is not usually required for disaster responders.

We evaluate the prediction quality on the expertly labelled subset of images with bounding boxes, results are given in Table 1. Average precision (AP^{IoU}) metric details are given in common objects in context (COCO) challenges (`https://cocodataset.org/#detection-eval`), precision and recall in the context of object detection are defined as in PASCAL VOC challenges (`http://host.robots.ox.ac.uk/pascal/VOC/pubs/everingham15.pdf`). As it can be seen our ensemble method achieve reasonable results on both types of images with the higher performance on the pre-event images. The neural networks are trained on undamaged buildings, whereas the post-event images capture both undamaged and damaged buildings in contrast to the pre-event images. Moreover, images immediately after the disaster are usually of lower quality, as there is a less chance to obtain good projection with proper atmospheric conditions in the short period of time. Therefore, the performance is lower on them compared to the pre-event images, where the best ones are chosen from a sequence of images over the longer period of time.

2.2 Mark Aggregation

The crowd was asked to provide marks of three possible types that correspond to the severity of structural damage: *minor*, *significant*, and *catastrophic* damage. The details are given in Figure 2 in

Table 2: Classification accuracy of mark aggregation models. Values are F1 scores in the percentage form. Column names contain the support size for the metric. Average values for all labels are weighted by support.

Model	average / 135	empty / 80	minor / 12	significant / 30	catastrophic / 13
IBCC	92	92	40	58	65
MV	90	70	40	60	59

the appendix. Each mark represents a point on the map with given geospatial coordinates of latitude and longitude and label of the class.

The footprint detection step yields the buildings' bounding boxes. For each object, there is a set of corresponding mark labels from volunteers. Due to volunteers being imperfect, the marks are noisy, meaning that labels from different volunteers for the same objects may identify different levels of damage severity. In the label aggregation step we determine for each object the consensus label from the crowd labels. The standard input for such a crowdsourcing task is an object/volunteer matrix, where each cell contains a label from the corresponding volunteer for this particular object. To form this matrix in our case, in addition to the explicit labels from volunteers (i.e., *minor*, *significant*, and *catastrophic* damage in our case), we use *unseen* and *empty* labels. An *unseen* label indicates that this particular volunteer has not seen the image that contains this object. Almost every volunteer will have some unseen labels, as it is rare that a single volunteer labels all images over an entire dataset. *Empty* are objects where a volunteer has not marked damage, either by labeling the whole image as undamaged, or labeling the damage in other parts of the image only.

There are multiple approaches for crowdsourcing aggregation that can be used once the object/volunteer matrix has been constructed. We consider below the *majority voting* (MV) and *independent Bayesian classifier combination* (IBCC) models.

2.2.1 Majority Voting

The majority voting model treats all volunteer marks to be of equal importance. For each object the most common label is selected. When the data quality is low, a significant number of volunteers can miss true objects and there will be mostly empty labels. Therefore, we weighted the labels, with a lower value for the empty label. However, volunteers differ in skill and some volunteers may identify the damage better than others, but majority voting will treat their labels with the same importance. This can partially be solved by weighting the different volunteers according to their skill level. This insight is incorporated into the independent Bayesian classifier combination approach.

2.2.2 Independent Bayesian Classifier Combination

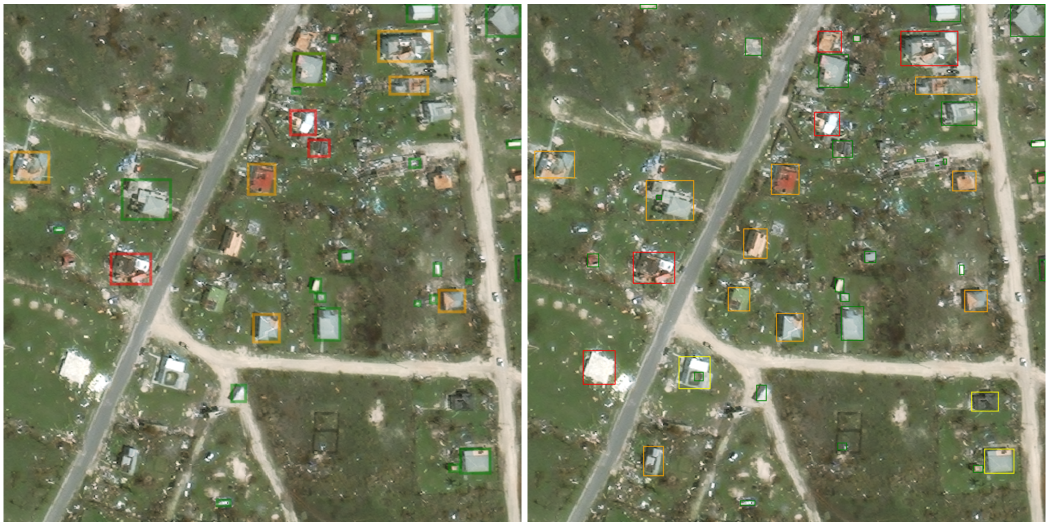
The Dawid-Skene model [Dawid and Skene, 1979] introduces weights for the quality of labels of each volunteer through a confusion matrix. Each row of the confusion matrix is the probability a volunteer assigns a label to an object conditional on the true class of the object. The individual confusion matrices for each volunteer are learnt from the data and the labels are subsequently aggregated.

The Bayesian version of the Dawid-Skene model, called independent Bayesian classifier combination [Kim and Ghahramani, 2012], places a Dirichlet prior over class probabilities and also over each row of the confusion matrix for each volunteer. These priors can express the skill level of the volunteer when known. Approximations for the posterior object class probabilities and the posterior distributions over the confusion matrices are calculated efficiently using variational Bayes [Simpson et al., 2013, Isupova et al., 2018].

To compare the accuracy of IBCC and MV we used the expertly annotated data as ground truth. The results are given in Table 2. It is not possible to choose the empty label weight such that MV aggregates several noisy points into the correct empty label, as the optimal weight varies in different cases. Therefore, it leads to overestimation of damaged buildings by MV. This is the main reason why IBCC shows better performance according to the metrics. Examples of the results of the different models are shown in Figure 3 in the appendix.

Table 3: Bounding boxes with damage severity prediction quality on the test data.

AP	AP50	AP75	APs	APm	APl
18.488	31.169	21.876	16.170	27.217	0.000



(a) Predictions

(b) Ground truth

Figure 1: Damage detection by the neural network. Ground truth is generated with the approach described in Section 2. Colors of bounding boxes represent labels: green — empty, i.e., undamaged, yellow — minor, orange — significant, and red — catastrophic.

3 Damage Detection

In this section we demonstrate the viability of our approach for building a training dataset for damage detection from point crowdsourced marks by training a neural net on this dataset. In this case we train using post-event images only.

For the neural network architecture we use Faster-RCNN [Ren et al., 2015] with ResNet-50+FPN backbone implemented in detectron2 library (`github.com/facebookresearch/detectron2`) pre-trained on ImageNet. We achieve the results on the COCO metrics for the bounding boxes presented in Table 3. For the reference, this model achieves AP of 37.9 on the COCO2017 dataset (`github.com/facebookresearch/detectron2/blob/main/MODEL_ZOO.md`).

Figure 1 shows example predictions on the test data, additional examples are provided in the appendix. Compared to the COCO dataset, in our case the objects are generally smaller (for example, we have no objects in *large* category of bounding boxes) and the image quality is reduced due to the environment. Moreover, the size of our training dataset is much smaller than the COCO dataset size and training labels in the COCO dataset are carefully tuned. Nevertheless, as it can be seen in Figure 1 the trained neural network is able to correctly identify approximate areas of damage. This is the information actually required by the disaster responders as they use damage mapping to plan their operation and logistic. Potential errors in terms of individual buildings are not very crucial for this purpose.

4 Conclusions

We propose a crowdsourced point-based labelling strategy that reduces the time to create structural damage datasets following disasters. Our method can be used to classify the damage severity of individual buildings in the disaster zone during the live response operations. We rely on only point marks placed anywhere on damaged structural buildings. This allows us to employ non-expert volunteers and to speed up the labelling process. The approach is robust to labelling inaccuracies of different volunteers and to common misalignments of pre- and post-event satellite imagery. We have demonstrated that these datasets can be used to train a neural network object detection damage mapper with the sound accuracy. This mapper can then be used to label new areas affected by the disaster during the same response deployment.

5 Acknowledgements

DK and BDS acknowledge support from UKRI via the Medical Research Council [FLF grant number MR/T044136/1] and the Biotechnology and Biological Sciences Research Council [grant number BB/T018941/1]. DK and BDS also acknowledge Lancaster University IAA funding support. SR acknowledges support from the Lloyd's Register Foundation through the Alan Turing Institute's Data Centric Engineering programme and for EPSRC funding as a Researcher in Residence with the Satellite Applications Catapult. DK, BDS and SR acknowledge support from the Science and Technology Facilities Council [grant number ST/S00307X/1].

References

A. Bearman, O. Russakovsky, V. Ferrari, and L. Fei-Fei. What's the point: Semantic segmentation with point supervision. In *Proccedings of ECCV*, pages 549–565, 2016.

R. Benenson, S. Popov, and V. Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of CVPR*, pages 11700–11709, 2019.

V. Benson and A. Ecker. Assessing out-of-domain generalization for robust building damage detection. In *AI+HADR Workshop at NeurIPS*, 2020.

J.-B. Boin, N. Roth, J. Doshi, P. Llueca, and N. Borensztein. Multi-class segmentation under severe class imbalance: A case study in roof damage assessment. In *AI+HADR Workshop at NeurIPS*, 2020.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska. DeepGlobe 2018: A challenge to parse the Earth through satellite images. In *CVPR Workshops*, pages 172–17209, 2018.

S. Golovanov, R. Kurbanov, A. Artamonov, A. Davydow, and S. I. Nikolenko. Building detection from satellite imagery using a composite loss function. In *CVPR Workshops*, pages 229–232, 2018.

R. Gupta and M. Shah. RescueNet: Joint building segmentation and damage assessment from satellite imagery. *arXiv preprint arXiv:2004.07312*, 2020.

R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston. Creating xBD: A dataset for assessing building damage from satellite imagery. In *CVPR Workshops*, 2019.

O. Isupova, Y. Li, D. Kuzin, S. J. Roberts, K. Willis, and S. Reece. BCCNet: Bayesian classifier combination neural network. In *NeurlPS Workshop on Machine Learning for the Developing World*, 2018.

H.-C. Kim and Z. Ghahramani. Bayesian classifier combination. In *AISTATS*, pages 619–627, 2012.

D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. Xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018.

J. Lee, J. Z. Xu, K. Sohn, W. Lu, D. Berthelot, I. Gur, P. Khaitan, et al. Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques. *arXiv preprint arXiv:2011.14004*, 2020.

Y. Li, W. Hu, H. Dong, and X. Zhang. Building damage detection from post-event aerial imagery using single shot multibox detector. *Applied Sciences*, 9(6):1128, 2019.

K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of CVPR*, pages 616–625, 2018.

D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari. Training object class detectors with click supervision. In *Proceedings of CVPR*, pages 6374–6383, 2017.

S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in NeurIPS*, 28:91–99, 2015.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

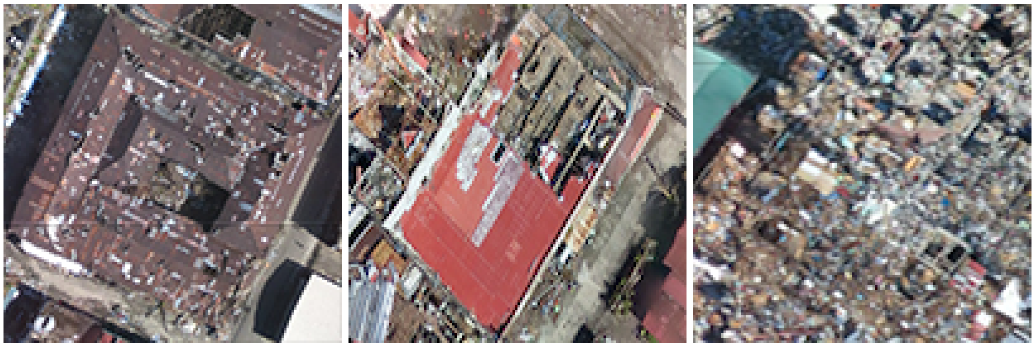
Y. Shen, S. Zhu, T. Yang, and C. Chen. Cross-directional feature fusion network for building damage assessment from satellite imagery. In *AI+HADR Workshop at NeurIPS*, 2020.

E. Simpson, S. Roberts, I. Psorakis, and A. Smith. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

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

J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva. Building damage detection in satellite imagery using convolutional neural networks. In *AI+HADR Workshop ar NeurIPS*, 2019.

X. Zhou, J. Zhuo, and P. Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of CVPR*, pages 850–859, 2019.



(a) Minor (b) Significant (c) Catastrophic

Figure 2: Severity of structural damage. *Minor* (a) (less than 20% of the building is damaged) where there is clearly damage, but the structure seems like it could still be used or inhabited; *Significant* (b) (between 20% and 60%) where the damage is severe, but the structure is still present and appears recognizable, even if it may be too damaged to be used or inhabited and *Catastrophic* (c) (over 60%) for which the structure is so badly damaged it is clearly unusable, and may not even exist anymore.

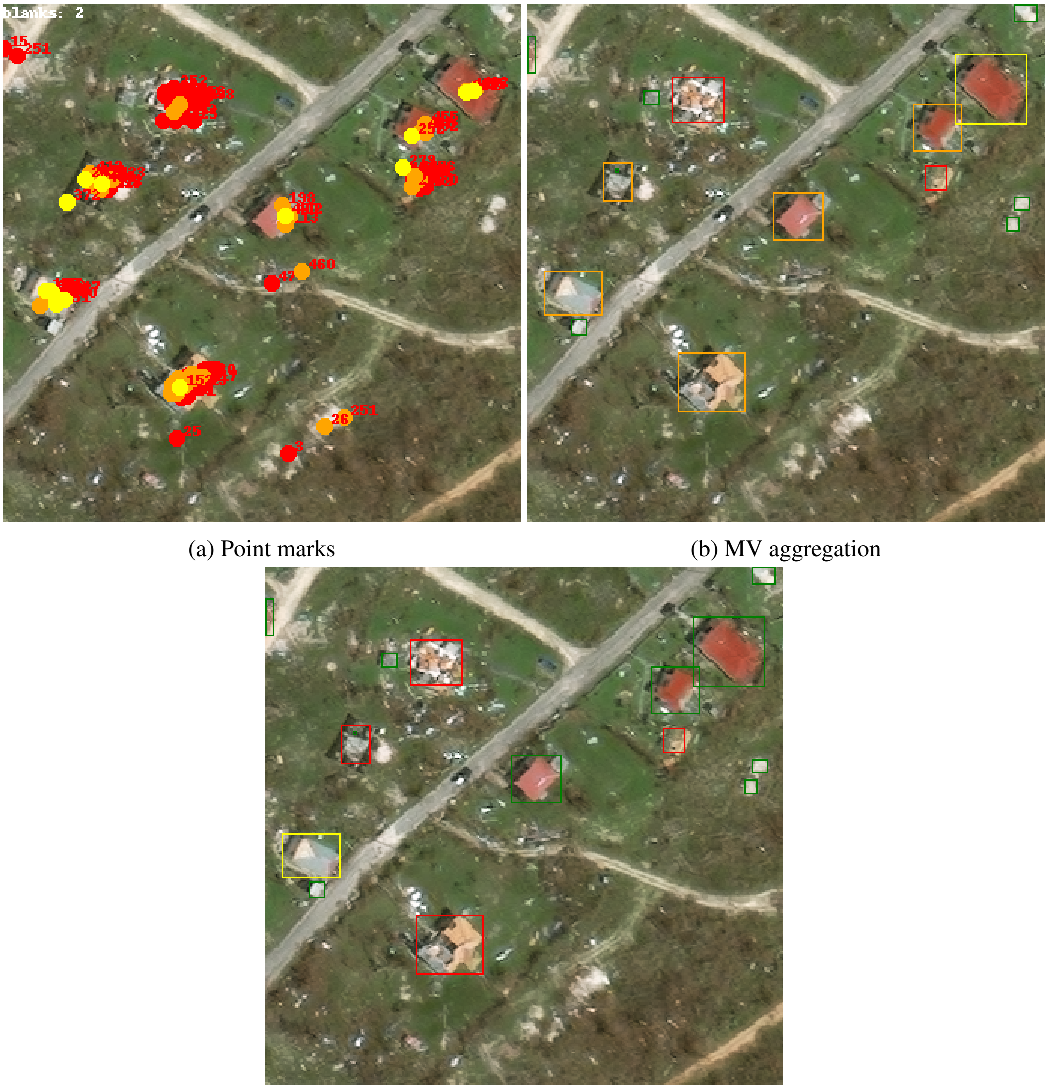
6 Appendix

In the appendix we provide additional figures to the main text.

Figure 2 provides examples for each of the damage severity classes that we use in this work.

Illustration of mark aggregation step with majority voting and independent Bayesian classifier combination (Section 2.2) is given in Figure 3.

In addition to Figure 1, we provide further examples of damage detection by the trained neural network (Section 3) on the test data.

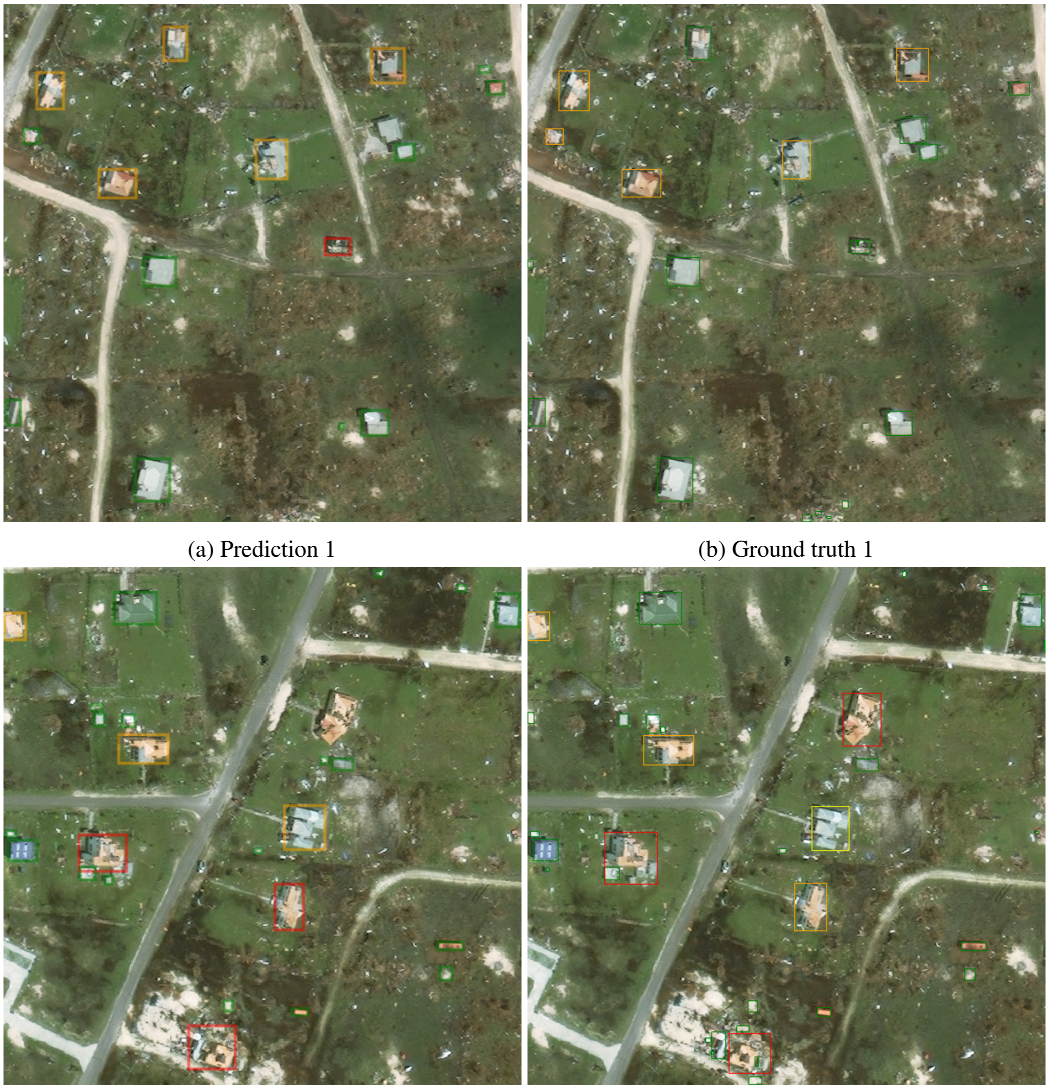


(a) Point marks

(b) MV aggregation

(c) IBCC aggregation

Figure 3: Comparison of crowdsourcing aggregation models. Marks from different volunteers (a) are aggregated using the MV (b) and IBCC (c) algorithms inside the detected building footprints. Colours indicate severity of damage: yellow — minor, orange — significant, red — catastrophic. On the image a the number of volunteers marked this image as empty of damage is depicted in the top left; also the image contains IDs of the volunteers that placed each damage mark.



(a) Prediction 1

(b) Ground truth 1

(c) Prediction 2

(d) Ground truth 2

Figure 4: Damage detection examples