Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery

Ritwik Gupta[1,2] Bryce Goodman[3,5] Nirav Patel[3,5] Richard Hosfelt[1,2] Sandra Sajeev[1,2]
Eric Heim[1,2] Jigar Doshi[6] Keane Lucas[4,5] Howie Choset[1] Matthew Gaston[1,2]
[1]Carnegie Mellon University [2]Software Engineering Institute [3]Defense Innovation Unit
[4]Joint Artificial Intelligence Center [5]Department of Defense [6]CrowdAI, Inc.

Abstract

We present a preliminary report for xBD, a new large-scale dataset for the advancement of change detection and building damage assessment for humanitarian assistance and disaster recovery research. Logistics, resource planning, and damage estimation are difficult tasks after a disaster, and putting first responders into post-disaster situations is dangerous and costly. Using passive methods, such as analysis on satellite imagery, to perform damage assessment saves manpower, lowers risk, and expedites an otherwise dangerous process. xBD provides pre- and post-event multi-band satellite imagery from a variety of disaster events with building polygons, classification labels for damage types, ordinal labels of damage level, and corresponding satellite metadata. Furthermore, the dataset contains bounding boxes and labels for environmental factors such as fire, water, and smoke. xBD will be the largest building damage assessment dataset to date, containing $\sim$700,000 building annotations across over 5,000 km^2 of imagery from 15 countries.

1. Introduction

With recent, abrupt changes in weather patterns around the world, natural disasters have become more unpredictable and have wider impacts than ever before [1]. Improvements in the fields of machine learning and computer vision have led to increased interest in creating smart, autonomous solutions to problems that many first responders face worldwide. The increasing availability of satellite imagery from sources such as USGS, NOAA, and ESA allows researchers to create models for a variety of humanitarian assistance and disaster recovery (HADR) tasks. Training accurate and robust models necessitates large, diverse datasets. Unfortunately, datasets for these use cases are hard to obtain. Although large-scale disasters bring catastrophic damage, they are relatively infrequent, so the availability of relevant satellite imagery is low. Furthermore, building design differs depending on where a structure is located in the world. As a result, damage of the same severity can look different from place

Figure 1: From top left (clockwise): California wildfire; Sulawesi tsunami; India monsoon; Lombok earthquake. Imagery from DigitalGlobe.

to place, and data must exist to reflect this phenomenon. Last, guidance for assessing building damage from satellite imagery for a wide variety of disasters is lacking in available literature.

In order to fully support machine learning for building damage assessment, datasets of appropriate scope, scale, size, and standard must be available. For this reason, we introduce xBD, a satellite imagery dataset for building damage assessment. xBD addresses the limitations enumerated above by collecting data across 8 disaster types, 15 countries, and thousands of square kilometers of imagery. Furthermore, we introduce a Joint Damage Scale that provides guidance and an assessment scale to label building damage in satellite imagery. xBD is used to introduce the xView 2.0 challenge

problem and address operational concerns of the HADR and emergency response community.

2. Related Work

Assessing damage from satellite imagery, especially for disparate disaster types, is a complicated task. Intuitively, it is easy to imagine that different disasters—for example, floods versus fires—would impact buildings in drastically different ways. In order to train robust building damage assessment models, datasets that can provide imagery from multiple types and severities of disasters are critical.

2.1. Existing data sources

Existent satellite imagery datasets that contain damaged buildings cover only single disaster types, with different labeling criteria for damaged structures [14, 4, 28, 12].

Furthermore, there are limited datasets that provide labels for forces of nature (fire, water, wind, etc.) that caused damage to a building visible in satellite imagery. Related datasets [9, 15] provide derivative location information of where these forces may occur, but they do not contain imagery of the damaged structures themselves. From an operational perspective, it is much more useful to be able to classify a building as, for example, "damage level 5 caused by fire and wind" as opposed to simply "damage level 5." For any given disaster, currently it is too difficult and time consuming to assess which factors damaged a given building to any level of granularity. It is simply assumed that a tornado caused wind damage, a wildfire caused fire-related damage, and more. However, this ignores secondary issues that arise because of natural disasters, such as fires that start after earthquakes, wind damage associated with flooding disasters, and more. A lack of datasets for this problem directly impedes any potential for model development.

Similarly, there are few public datasets of satellite imagery with associated bounding boxes and labels for environmental factors such as smoke, water, lava, and fire. Existence of these environmental factors in the satellite imagery provides insights that can aid HADR image analysts in disambiguating the causal factors and severity of building damage.

Last, many available datasets present multi-view imagery for change detection, land classification, and other tasks [7, 6, 2, 18]. Imagery from these datasets can span multiple visits over one site and provide a rich resource for any sort of time-series imagery analysis. xBD emulates these datasets and provides pre- and post-disaster imagery over consistent sites for additional environmental context.

2.2. Assessment scales

There is a large corpus of literature that addresses how to assess damage after disasters such as fires, earthquakes, and hurricanes [20, 27, 24, 13]. However, since each scale in this corpus is specifically scoped to a limited set of disaster types and is generally meant for in-person damage assessment, the scales cannot be simultaneously used to assess building damage by multiple disasters in satellite imagery. There are some notable attempts to assess multiple types of disasters. The HAZUS software from FEMA [11] is used to assess earthquakes, floods, and hurricanes. More generally, the FEMA Damage Assessment Operations Manual [10] describes damage qualitatively, but mostly relies on in-person assessments and provides specific criteria only for residential dwellings. As a result, there does not exist a standard scale for assessing damage across damage types from aerial imagery.

3. Dataset Details

In this section we describe the xBD dataset in detail, including the image collection considerations, damage scale, annotation process, and quality control. We also show statistics about the dataset that describe the diversity, scale, and depth of the imagery.

The dataset consists of the following: building polygons, which are granular representations of a building footprint; ordinal regression labels for damage, which rate how damaged a building is on an increasing integer scale; multi-class labels, which relate all the environmental factors that caused the damage seen in the imagery; and environmental factor bounding boxes and labels, which are a rough approximation of the area covered by smoke, water, fire, and other environmental factors.

3.1. Motivating processes for image collection

In collaboration with imagery analysts from the California Air National Guard, we identified the process by which human analysts currently label building damage from satellite imagery and the implicit decisions they make in order to provide insights to first responders. The process is as follows: when a disaster occurs, analysts receive aerial and satellite imagery of the impacted regions from state, federal, and commercial sources. Analysts make an initial overall assessment of what sub-regions look the most damaged and further analyze, identify, and count the number of structures damaged. The scale used to assess the damage as well as the types of structures that will be assessed depend on the type of disaster, requesting agency, and use case for the assessment.

This process informed a set of criteria that guided the specific data we targeted for inclusion in the dataset, as well as weaknesses of the current damage assessment processes. Each disaster is treated in isolation. The process human analysts use is not repeatable or reproducible across different disaster types. This irreproducible data presents a major issue for use by machine learning algorithms; different disasters affect buildings in different ways, and building structures
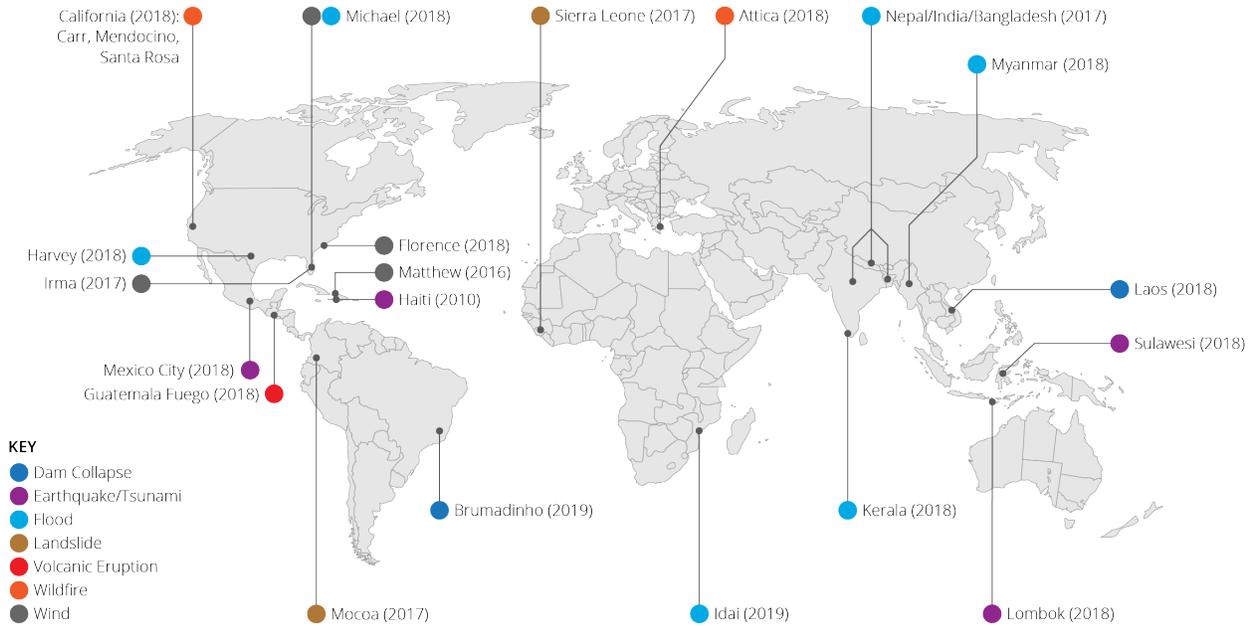
Figure 2: Disaster types and disasters represented in xBD over the world.

vary from country to country, so determinism in the assessment is a necessary property to ensure machine learning algorithms can learn meaningful patterns. Therefore, to ensure a holistic view of building damage, we source imagery from a variety of disaster types across many countries and environments. Table 2 lists all disasters in xBD and their corresponding disaster types.

An additional consideration for the dataset is the inclusion of labeled environmental factors such as smoke, landslide, and water. Human analysts implicitly acknowledge the existence of these factors, which affects their assessment as they mentally categorize the type of damage to expect. xBD contains explicit labels to assist any downstream machine learning task in accomplishing the same categorization.

A clear understanding of what buildings do *not* look like is necessary when training models to look for buildings, damaged or otherwise. Explicit techniques to establish this understanding exist; one example is hard negative mining [26, 25]. In order to support such techniques, we included a large amount of satellite imagery in xBD that may contain objects such as vehicles or natural features but no buildings.

3.2. Joint Damage Scale

To address the lack of a scale for building damage assessment that covers the various types of damage in this satellite imagery dataset, we present the Joint Damage Scale (Table 3), based mainly on HAZUS [11], FEMA's Damage

Assessment Operations Manual [10], the Kelman scale [21], and the EMS-98 [16]. Furthermore, various literature from the GIS community [5, 30, 8], paired with expert insights from the California and Indiana Air National Guards and the US Air Force, help ground the scale in operational relevance. Assessing damage via satellite imagery is a proxy for the real task, which requires on-the-ground human analysts using these cited scales to assign a metric of damage based on the functionality of the building, not its looks. It is not simple to reconcile this highly objective task of assessing functionality in person with assessing functionality from top-down satellite imagery. As such, this scale is not meant as an authoritative damage assessment rating, but it does provide the first attempt to create a unified assessment scale for building damage in satellite imagery across multiple disaster types, structure categories, and geographical location. Imagery analysts who provide preliminary building damage assessment as well as related machine learning applications would find this scale to be applicable.

The Joint Damage Scale ranges from no damage (0) to destroyed (3). This granularity is based on satellite imagery resolution, available imagery features, and operational usefulness. The descriptions of each damage level have been generalized to handle the wide variety of disasters present in xBD. The scale allows room for analyst discretion, which can result in some amount of label noise. Although such nuance is not ideal from an objective assessment standpoint,

it allows analysts to gracefully handle tough cases that fall between classification boundaries. Figure 5 and Appendix Figure 6 show examples of what it means to be in one of these categories.

Due to the limitations presented by the modality of satellite imagery, such as resolution, azimuth, and smear [29], this scale presents a best-effort trade-off between operational relevance and technical correctness, and thus cannot accommodate the degree of precision that a scale meant for in-person, human assessment provides.

Disaster Level	Structure Description
0 (No Damage)	Undisturbed. No sign of water, structural or shingle damage, or burn marks.
1 (Minor Damage)	Building partially burnt, water surrounding structure, volcanic flow nearby, roof elements missing, or visible cracks.
2 (Major Damage)	Partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud.
3 (Destroyed)	Scorched, completely collapsed, partially/ completely covered with water/mud, or otherwise no longer present.

Figure 3: Joint Damage Scale descriptions on a four-level granularity scheme.

3.3. Polygons

Polygons for all buildings are provided in standard, well-known text representation conformant to the ISO/IEC 13249-3:2016 standard [19]. Polygons will be provided in terms of their latitude and longitude as well as their respective pixel location in a given swath of imagery. Structures that are below a set size in terms of square pixels will not be annotated in the dataset.

Bounding boxes for environmental factors are rough and potentially overlap annotated buildings as well. These boxes are provided for a relative understanding of the environment.

3.4. Multi-class labels

We provide multi-class labels for environmental factors that caused damage to structures. These labels are "fire", "water", "wind", "land shift", and "other" in the form of a many-hot vector (i.e., a vector in the form of $[0, 1, 0, 0, 1, 0, ...]$, where a 1 is present at the index of the correct classes). These labels are high level and meant to provide a coarse level of discrimination to maintain a low level of noise in this set of annotations.

3.5. Quality control

Quality control is an essential part of creating an accurate dataset with many labels and often complicated polygons. A three-stage quality control process was adapted

Figure 4: Polygons over a small part of the coast of Florida with damage labels. Imagery from DigitalGlobe.

from xView 1.0 [22] with significant alterations to account for the initial round of automated labeling. The first stage consists of global verification where all annotators verify the orientation, positioning, and fidelity of all automatic and human-extracted polygons. After all polygons are verified, all human-annotated labels are verified.

The second stage uses a targeted, experienced set of annotators to check for continuity of polygons, granularity of multi-class labels, empty chips, and any disagreeing labels.

The final, third stage relies on expert annotators that consist of the authors, HADR subject matter experts, and remote-sensing/satellite imagery experts who randomly sample 4% of all individual tasks. If any task is found to be incorrect, the entire batch—which itself consists of many tasks—is re-submitted for annotation and sent back for expert verification.

3.6. Targeted dataset statistics

All imagery for xBD is sourced from DigitalGlobe.[1] DigitalGlobe provides high-resolution imagery at ~0.5m ground sample distance, which provides ample resolution for this labeling task. Furthermore, we are able to obtain pre- and post-disaster imagery in multi-band (3, 4, or 8) formats, which give xBD greater representational capacity. Each image is also accompanied by metadata such as "sun azimuth" and "off-nadir", which enable data processing pipelines to account for various skews in the imagery itself. The imagery consists of 22 different disasters and is gathered from 15 countries (Figure 2) at various times of the year. Disasters were picked on the basis of what impact people often and at a high severity. Furthermore, disasters had to have a range on their levels of severity, otherwise their impact on buildings would not be differentiable enough to provide good labels.

Overall, xBD will provide approximately 700,000 building polygons with ordinal damage and multi-class damage

---

[1] https://www.digitalglobe.com/

causal labels, as well as approximately 1,000 environmental factor bounding boxes with corresponding class labels. The distributions over the labels are unknown at this time, but figures and exact numbers will be provided once the entire data collection process has ended.

3.7. Unique dataset features

As briefly covered in Section 2, xBD presents many attributes that are not available in other datasets that could be used for this task.

xBD contains imagery from the majority of common disaster types labeled with a common, expert-verified damage assessment scale. By including vastly different modes of damage in the dataset, we provide a much more robust view of damage than related datasets.

The dataset also has large geographical diversity. In particular, imagery contains buildings in both highly dense and hyper-sparse settings. This range presents difficulties for proper localization methods. Furthermore, since imagery is sourced from around the world, buildings are organized and designed in different ways, providing a representational complexity not present in existent literature.

4. Challenge

xBD is being released in conjunction with the second iteration of the global xView challenge. For the purposes of the challenge, the dataset is used in a specific fashion that balances research interest and the operational concerns of various relief-providing agencies, but this use does not preclude it from being used for other tasks.

4.1. Challenge statement

xBD provides building polygons, ordinal regression labels for building damage, and multi-class labels for environmental factors that caused the damage. Given training data, the challenge is to create models and methods that can extract building polygons and assess the building damage level of polygons on an ordinal scale. Furthermore, the models and methods must assign an additional multi-class label to each polygon that indicates which natural force caused the damage to the building.

4.2. Performance metric

A metric for scoring the xView 2.0 challenge needs to smoothly combine scores for the tasks of regression, classification, and localization. An ideal metric would weight these tasks from high to low in their listed order. xView 1.0 had to make explicit design trade-offs which resulted in certain classes of objects being scored higher for accurately localizing them. In the case of xView 2.0, every disaster type is treated as equally important, but due to the prevalence of models that can localize building polygons, the weighting for

localization is not high. The majority of the difficulty relies on the fine-grained regression and classification required to accurately assess building damage.

4.3. Challenge restrictions

To better accommodate operational use cases, the data used for inference in the challenge will be purposefully lowered in resolution in a stochastic fashion. Since high-resolution imagery cannot be guaranteed for all parts of the world, for all disasters, around the clock, any models created must work across dynamic resolution limitations.

As with xView 1.0, inference for any submitted model will be limited to CPU only with an upper limit on computation time per image.

5. Potential Use Cases

We provide a brief, non-exhaustive list of compelling use cases for xBD that are both operationally useful and academically interesting.

Obstructed road segmentation. The dataset includes images of many roads that are broken, covered with a variety of debris, flooded, and otherwise obstructed. Unsupervised models already excel at road segmentation in satellite imagery [3, 23, 17], but limited literature currently addresses segmentation and identification of roads with obstructions. Models that can detect obstructed roads would provide advanced routing capabilities to first responders and disaster planners who need to know how to navigate through a disaster environment.

Routing across obstructed roads. As an extension to the previous use case, xBD could also be used as a dataset for automatic route planning across obstructed or no-longer-usable roads. Not much literature exists that targets this problem, partially due to a lack of datasets providing obstructed roadways.

Force of nature identification. xBD provides bounding boxes and labels for many environmental factors—such as fire, water, and lava—that cause damage. It is essential for first responders to know where any of these factors may be present. It is possible to create robust detection models for this task with the data provided in xBD.

6. Conclusion

xBD presents the largest satellite imagery dataset for building damage assessment with over 700,000 labeled building instances covering over 5,000 km^2 of imagery.

Furthermore, xBD contains labeled data about environmental factors and hazards that can be used to aid building damage assessment or other research tasks not considered by the xView 2.0 challenge.

By combining insights from disaster recovery and emergency response experts from many US Government agencies,

Figure 5: (From left to right) "No damage", "minor", and "major" wind damage. "Destroyed" example not available yet.

we are able to generate a high-quality dataset for research purposes while maintaining operational relevance.

References

[1] M. K. V. Aalst. The Impacts of Climate Change on the Risk of Natural Disasters. *Disasters*, 30(1):5–18, 2006. 1

[2] M. Bosch, Z. Kurtz, S. Hagstrom, and M. Brown. A Multiple View Stereo Benchmark for Satellite Imagery. In *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, Oct. 2016. 2

[3] A. Buslaev, S. Seferbekov, V. Iglovikov, and A. Shvets. Fully Convolutional Network for Automatic Road Extraction from Satellite Imagery. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 197–1973. IEEE, June 2018. 5

[4] S. A. Chen, A. Escay, C. Haberland, T. Schneider, V. Staneva, and Y. Choe. Benchmark Dataset for Automatic Damaged Building Detection from Post-Hurricane Remotely Sensed Imagery. *arXiv:1812.05581 [cs]*, Dec. 2018. 2

[5] L. Chiroiu and G. André. Damage Assessment Using High Resolution Satellite Imagery: Application to 2001 Bhuj, India, Earthquake. *Proceedings of the 7th National Conference on Earthquake Engineering: Urban Earthquake Risk*, page 13, Nov. 2001. 3

[6] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6172–6180. IEEE, June 2018. 2

[7] R. C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban Change Detection for Multispectral Earth Observation Using Convolutional Neural Networks. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 2115–2118. IEEE, July 2018. 2

[8] F. Dell'Acqua and D. A. Polli. Post-Event Only VHR Radar Satellite Data for Automated Damage Assessment: A Study on COSMO/SkyMed and the 2010 Haiti Earthquake. *Photogrammetric Engineering & Remote Sensing*, 77(10):1037–1043, Oct. 2011. 3

[9] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–17209. IEEE, June 2018. 2

[10] Federal Emergency Management Agency. Damage Assessment Operations Manual: A Guide to Assessing Damage and Impact. Technical report, Federal Emergency Management Agency, Apr. 2016. 2, 3

[11] Federal Emergency Management Agency. Hazus Hurricane Model User Guidance. Technical report, FEMA, Apr. 2018. 2, 3

[12] R. Foulser-Piggott, R. Spence, K. Saito, D. M. Brown, and R. Eguchi. The Use of Remote Sensing for Post-Earthquake Damage Assessment: Lessons from Recent Events, and Fu-

ture Prospects. *Proceedings of the Fifthteenth World Conference on Earthquake Engineering*, page 10, 2012. 2

[13] C. J. Friedland. *Residential Building Damage from Hurricane Storm Surge: Proposed Methodologies to Describe, Assess and Model Building Damage*. PhD thesis, Louisiana State University, 2009. 2

[14] A. Fujita, K. Sakurada, T. Imaizumi, R. Ito, S. Hikosaka, and R. Nakamura. Damage Detection from Aerial Images via Convolutional Neural Networks. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, pages 5–8. IEEE, May 2017. 2

[15] L. Giglio, J. T. Randerson, and G. R. van der Werf. Analysis of Daily, Monthly, and Annual Burned Area Using the Fourth-Generation Global Fire Emissions Database (GFED4). *Journal of Geophysical Research: Biogeosciences*, 118(1):317–328, 2013. 2

[16] G. Grünthal, R. Musson, J. Schwarz, and M. Stucchi. EMS-98 (European Macroseismic Scale). Technical report, European Seismological Commision, 1998. 3

[17] C. Henry, S. M. Azimi, and N. Merkle. Road Segmentation in SAR Satellite Images with Deep Fully-Convolutional Neural Networks. *IEEE Geoscience and Remote Sensing Letters*, 15(12):1867–1871, Dec. 2018. 5

[18] D. Ienco and R. Gaetano. TiSeLaC : Time Series Land Cover Classification Challenge. *TiSeLaC : Time Series Land Cover Classification Challenge*, Jan. 2007. 2

[19] Information technology – Database languages – SQL multimedia and application packages – Part 3: Spatial. Standard, International Organization for Standardization, Geneva, CH, Jan. 2016. 4

[20] M. Ji, L. Liu, and M. Buchroithner. Identifying Collapsed Buildings Using Post-Earthquake Satellite Imagery and Convolutional Neural Networks: A Case Study of the 2010 Haiti Earthquake. *Remote Sensing*, 10(11):1689, Nov. 2018. 2

[21] I. Kelman. *Physical Flood Vulnerability of Residential Properties in Coastal, Eastern England*. PhD thesis, University of Cambridge, U.K., 2002. 3

[22] 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:1802.07856 [cs]*, Feb. 2018. 4

[23] V. Mnih and G. E. Hinton. Learning to Detect Roads in High-Resolution Aerial Images. In *Computer Vision – ECCV 2010*, volume 6316, pages 210–223. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. 5

[24] S. Okada and N. Takai. Classifications of Structural Types and Damage Patterns of Buildings for Earthquake Field Investigation. *Journal of Structural and Construction Engineering (Transactions of AIJ)*, 64(524):65–72, 1999. 2

[25] A. Shrivastava, A. Gupta, and R. Girshick. Training Region-Based Object Detectors with Online Hard Example Mining. *arXiv:1604.03540 [cs]*, Apr. 2016. 3

[26] K.-K. Sung. *Learning and Example Selection for Object and Pattern Detection*. PhD thesis, Massachusetts Institute of Technology, Mar. 1996. 3

[27] A. Thieken, V. Ackermann, F. Elmer, H. Kreibich, B. Kuhlmann, U. Kunert, H. Maiwald, B. Merz, K. Piroth,

J. Schwarz, R. Schwarze, I. Seifert, J. Seifert, and L.-F.-U. Innsbruck. Methods for the Evaluation of Direct and Indirect Flood Losses. *Proceedings of the 4th International Symposium on Flood Defence*, page 10, May 2008. 2

[28] United States Geological Survey. Building Damage Assessment in the City of Mocoa. *United Nations Institute for Training and Research*, Apr. 2017. 2

[29] W. A. Wahballah, T. M. Bazan, and M. Ibrahim. Smear effect on high-resolution remote sensing satellite image quality. In *2018 IEEE Aerospace Conference*, pages 1–14, Mar. 2018. 4

[30] F. Yamazaki, K. Kouchi, M. Matsuoka, M. Kohiyama, and N. Muraoka. Damage Detection From High-Resolution Satellite Images for the 2003 Boumerdes, Algeria Earthquake. *Proceedings of 3rd International workshop Remote Sensing for Post-Disaster Response*, page 13, Sept. 2005. 3

Appendices

Figure 6: (Top) "Major" and "destroyed" examples for flooding. (Bottom) "No damage" and "destroyed" examples for fire.

Figure 7: The same chip in red/green/blue (left) and red/near-IR/blue (right).