Detecting Natural Disasters, Damage, and Incidents in the Wild

Ethan Weber[1](✉), Nuria Marzo[1], Dim P. Papadopoulos[1], Aritro Biswas[1], Agata Lapedriza[1,3], Ferda Ofli[2], Muhammad Imran[2], and Antonio Torralba[1]

[1] Massachusetts Institute of Technology, Cambridge, USA
{ejweber,nmarzo,dimpapa,abiswas,agata,torralba}@mit.edu
[2] Qatar Computing Research Institute, HBKU, Ar-Rayyan, Qatar
{fofli,mimran}@hbku.edu.qa
[3] Universitat Oberta de Catalunya, Barcelona, Spain

Abstract. Responding to natural disasters, such as earthquakes, floods, and wildfires, is a laborious task performed by on-the-ground emergency responders and analysts. Social media has emerged as a low-latency data source to quickly understand disaster situations. While most studies on social media are limited to text, images offer more information for understanding disaster and incident scenes. However, no large-scale image datasets for incident detection exists. In this work, we present the Incidents Dataset, which contains 446,684 images annotated by humans that cover 43 incidents across a variety of scenes. We employ a baseline classification model that mitigates false-positive errors and we perform image filtering experiments on millions of social media images from Flickr and Twitter. Through these experiments, we show how the Incidents Dataset can be used to detect images with incidents in the wild. Code, data, and models are available online at http://incidentsdataset.csail.mit.edu.

Keywords: Image classification · Visual recognition · Scene understanding · Image dataset · Social media · Disaster analysis · Incident detection

1 Introduction

Rapid detection of sudden onset disasters such as earthquakes, flash floods, and other emergencies such as road accidents is extremely important for response organizations. However, acquiring information in the occurrence of emergencies is labor-intensive and costly as it often requires manual data processing and expert assessment. To alleviate these manual efforts, there have been attempts to apply computer vision techniques on satellite imagery, synthetic aperture radar, and other remote sensing data [14,25,60,74]. Unfortunately, these approaches are still costly to deploy and they are not robust enough to obtain relevant data under time-critical situations. Moreover, satellite imagery is susceptible to noise such as clouds and smoke (i.e., common scenes during hurricanes and wildfires), and only provides an overhead view of the disaster-hit area.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12364, pp. 331–350, 2020.
https://doi.org/10.1007/978-3-030-58529-7_20

On the other hand, studies show that social media posts in the form of text messages, images, and videos are available moments after a disaster strikes and contain information pertinent to disaster response such as reports of damages to infrastructure, urgent needs of affected people, among others [13,35]. However, unlike other data sources (e.g., satellite), social media imagery remains unexplored, mainly because of two important challenges. First, image streams on social media are very noisy, and disasters are not an exception. Even after performing a text-based filter, a large percentage of images in social media are not relevant to specific disaster categories. Second, deep learning models, that are the standard techniques used for image classification, are data-hungry, and yet no large-scale ground-level image dataset exists today to build robust computational models.

In this work we address these challenges and investigate how to detect natural disasters, damage, and incidents in images. Concretely, our paper has the following three main contributions. First, we present the large-scale Incidents Dataset, which consists of 446,684 scene-centric images annotated by humans as positive for natural disasters (class-positives), types of damage or specific events that can require human attention or assistance, like traffic jams or car accidents. We use the term *incidents* to refer to the 43 categories covered by our dataset (Sect. 3). The dataset also contains an additional set of 697,464 images annotated by humans as negatives for specific incident categories (class-negatives). As discussed in Sect. 2, the Incidents Dataset is significantly larger, more complete, and much more diverse than any other dataset related to incident detection in scene-centric images. Second, using the full set of 1.1M images in our dataset, we train different deep learning models for incident classification and incident detection. In particular, we use a slightly modified binary cross-entropy loss function, which we refer to as class-negative loss, that exploits our class-negative images. Our experiments in Sect. 5 show the importance of using class-negatives in order to train a model that is robust enough to be deployed for incident detection in real scenarios, where the number of negatives is large. Third, we perform extended incident detection experiments on large-scale social media image collections, using millions of images from Flickr and Twitter. These experiments, presented in Sect. 6, show how our model, trained with the Incidents Dataset and the class-negative loss, can be effectively deployed in real situations to identify incidents in social media images. We hope that the release of the Incidents Dataset will spur more work in computer vision for humanitarian purposes, specifically natural disaster and incident analysis.

2 Related Work

Computer Vision for Social Good. Existing vision-based technologies are short of reaching out to diverse geographies and communities due to biases in the commonly used datasets. For instance, state-of-the-art object recognition models perform poorly on images of household items found in low-income countries [79]. To remedy this shortcoming, the community has made recent progress in areas

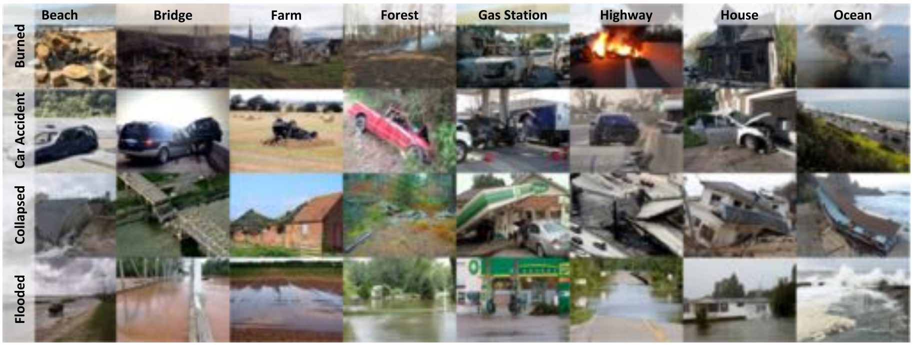


Fig. 1. Example images from the Incidents Dataset. Incidents (left) happen in many places (top), which we capture by having 43 incident and 49 place categories. Notice that a car accident can occur on a beach, farm, highway, etc. The place categories help by adding diversity to the dataset.

including agriculture [23,40,62,68], sustainable development [34,36,81], poverty mapping [59,77,80], human displacement [38,39], social welfare [10,26,27,51], health [2,50,82], urban analysis [4,11,41,52,53,85], and environment [42,70]. These studies, among many others, have shown the potential of computer vision to create impact for social good at a global scale.

Incident Detection on Satellite Imagery. There are numerous studies that combine traditional machine learning with a limited amount of airborne or satellite imagery collected over disaster zones [14,24,25,63,74,78]. For a detailed survey, see [17,19,37,60]. Oftentimes, these studies are constrained to particular disaster events. Recently, deep learning-based techniques have been applied on larger collections of remote-sensed data to assess structural damage [5,20,30,31,46,83] incurred by floods [7,56,67], hurricanes [47,75], and fires [21,64], among others. Some studies have also applied transfer learning [71] and few-shot learning [57] to deal with unseen situations in emergent disasters.

Incident Detection on Social Media. More recently, social media has emerged as an alternative data source for rapid disaster response. Most studies have focused heavily on text messages for extracting crisis-related information [35,66]. On the contrary, there are only a few studies using social media images for disaster response [1,3,15,16,45,54,55,58,61]. For example, [54] classifies images into three damage categories whereas [55] regresses continuous values indicating the level of destruction. Recently, [3] presented a system with duplicate removal, relevancy filtering, and damage assessment for analyzing social media images. [45,61] investigated adversarial networks to cope with data scarcity during an emergent disaster event.

Incident Detection Datasets. Most of the aforementioned studies use small datasets covering just a few disaster categories, which limits the possibility of creating methods for automatic incident detection. In addition, the reported results are usually not comparable due to lack of public benchmark datasets, whether it be from social media or satellites [69]. One exception is the xBD dataset [32], which contains 23,000 images annotated for building damage but covers only six disasters types (earthquake, tsunami, flood, volcanic eruption, wildfire, and wind). On the other hand, [30] has many more images but their dataset is constructed for detecting damage as anomaly using pre- and post-disaster images. There are also datasets combining social media and satellite imagery for understanding flood scenes [8,9] but they have up to 11,000 images only. In summary, existing incident datasets are small, both in number of images and categories. In particular, incident datasets are far, in size, from the current large datasets on image classification, like ImageNet [18] or Places [84], which contain millions of labeled images. Unfortunately, neither ImageNet nor Places covers incident categories. Our dataset is significantly larger, more complete, and much more diverse than any other available dataset related to incident detection, enabling the training of robust models able to detect incidents in the wild.

3 Incidents Dataset

In this section, we present the Incidents Dataset collected to train models for automatic detection of disasters, damage, and incidents in scene-centric images.

Incidents Taxonomy. We create a fine-grained vocabulary of 233 categories, covering high-level categories such as general types of damage (e.g., destroyed, blocked, collapsed), natural disasters including weather-related (e.g., heat wave, snow storm, blizzard, hurricane), water-related (e.g., coastal flood, flash flood, storm surge), fire-related (e.g., fire, wildfire, fire whirl), as well as geological (e.g., earthquake, landslide, mudslide, mudflow, volcanic eruption) events, and transportation and industrial accidents (e.g., train accident, car accident, oil spill, nuclear explosion). We then manually prune this extensive vocabulary by discarding categories that are hard to recognize from images (e.g., heat wave, infestation, famine) or by combining categories that are visually similar (e.g., snow storm and blizzard, or mudslide and mudflow). As a result of this pruning step, we obtain a final set of 43 incident categories.

Image Downloading and Duplicate Removal. Images are download from Google Images using a set of queries. To generate the queries and promote diversity on the data, we combine the 43 incident categories with place categories. For the place categories, we select the 118 outdoor categories of Places dataset [84] and merge categories belonging to the same super-category (e.g., topiary garden, Japanese garden, vegetable garden are merged into garden). After this process we obtain 49 different place categories. By combining incident and place categories, we obtain a total of 43 incidents × 49 places = 2107 pairs. Each pair is extended with incidents and places synonyms to create queries such as “car accident in highway” and “car wreck in flyover”, or “blizzard in street” and “snow storm in alley.” We obtain 10,188 queries in total and we download all images returned from Google Images for each query, resulting in a large collection of

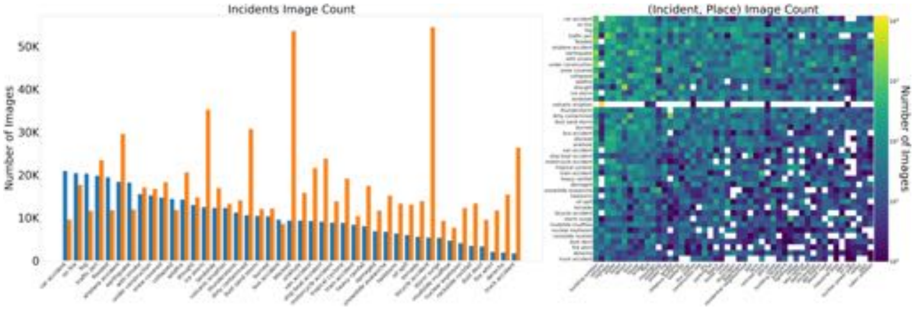


Fig. 2. Dataset composition. The number of positive and negative labeled incident images is shown on the left and the distribution of images for (incident, place) combinations is shown on the right. The dataset contains incidents in many different scenes. White cells indicate the *unlikely* (incident, place) combinations for which the Incidents Dataset does not contain any images (e.g., "car accident in volcano").

6,178,192 images. After that, we perform duplicate image removal as follows: we extract feature vectors from each image with a ResNet-18 [33] model trained on Places [84] and we cluster duplicate images with a radius-based Nearest Neighbor algorithm. This results in 3,487,339 unique images.

Image Labeling. Images obtained through Google Images are noisy, and they may not necessarily be relevant to the query they are downloaded for. Rather, the results may contain non-incident images with similar appearances (e.g., airplanes but not airplane accidents, fireplaces but not dangerous fires, bicycles and not bicycle accidents, etc.), images with other incidents, or completely random images. To clean the data we ask annotators to manually verify the images using the Amazon Mechanical Turk (MTurk) platform. Workers are shown a batch of images, and they have to answer whether each image belongs to a specific category or not. In particular, the interface used for image annotation is similar to [84]. Each image is annotated by one annotator. Each annotation batch contains 100 images, including 15 control images (10 positives and 5 negatives). Annotation batches are accepted when the accuracy in the control images is above 85%. Otherwise the annotations of the batch are discarded.

The images are annotated in several stages. First, we label 798,316 images from the initial 3,487,339 image collection, using the queries the images are downloaded from. For example, the images downloaded with the query "car accident in village" are labeled as positives or negatives for the class "car accident." This results in 193,648 class-positive incident images and 604,668 class-negatives. Class-negative images are those that we know do not show a specific incident class but they may contain another incident category. After the first annotation stage, we train a temporary incident recognition model, as described in Sect. 4, to determine which images to label next. We send images whose incident category confidence scores were greater than 0.5 to MTurk to get more class-positive and class-negative labels. This process is repeated until obtaining 446,684 positive incident images. Finally, these 446,684 images are sent to MTurk for annotation on place categories using the same interface. In this case, each image is assessed for the place category of its original query (e.g., an image downloaded with the query "wildfire in forest" is labeled as positive or negative for the "forest" category). Eventually, we obtain 167,999 images with positive place labels. The remaining images have negative place labels.

Dataset Statistics. The Incidents Dataset contains 1,144,148 labeled images in total. Of these, 446,684 are class-positive incident images, 167,999 of which have also positive place labels. Figure 1 shows some sample images from our dataset. Figure 2 shows the number of images per incident, place, and combined (incident, place). Although the common practice when collecting datasets is just to keep images with positive labels, we will show in Sect. 4 and Sect. 5 that class-negative images are particularly valuable for incident detection in the wild because they can be used as hard negatives for training.

4 Incident Model

In this section, we present our model for recognizing and detecting different incident types in scene-centric images.

Multi-task Architecture. The images in our Incidents Dataset are accompanied with an incident and a place label (see Sect. 3). We choose to build a single model that jointly recognizes incident and place categories following a standard multi-task learning paradigm [12,65,73]. This architecture offers efficiency as it can jointly recognize incidents and places, and we also did not observe any difference in the performance when training a model for a single task. In our experiments, we employ a Convolutional Neural Network (CNN) architecture with two task-specific output layers. Specifically, our network is composed of a sequence of shared convolutional layers followed by two parallel branches of fully-connected layers, corresponding to incident and place recognition tasks.

Training with a Cross-Entropy Loss. The standard and most successful strategy for training an image classification model (either for incidents or places) is to employ a cross-entropy loss on top of a softmax activation function for both outputs of the network. Note that this is the standard procedure for single-label classification of objects [18], scenes [84] or actions [73].

In our real-world scenario of detecting incidents in social media images, many of the test scene-centric images do not belong to any of the incidents categories and they should be classified as images with "no incident." This can be handled by adding an extra neuron in the output layer that should fire on "no incident" images. Notice that this requires training the model with additional absolute negative images, i.e., images that do not show any incident.

Training with a Class-Negative Loss. Even during an incident, the number of images depicting the incident is only a small proportion of all the images shared in social media. For this reason, our task of finding incidents in social media imagery is more closely related to that of detection [28,49,72] than classification. In particular, our model must find positive examples out of a pool of many challenging negatives (e.g. a chimney with smoke or a fireplace are not disaster situations, yet they share visual features similar to our "with smoke" and "on fire" incident categories). To handle this problem and mitigate false positive detections, either the training process can be improved [22,43] or the predictions can be adjusted at test time [44,48]. For our task, we choose to modify our training process to incorporate class-negatives.

In particular, similar to [22], we modify a binary cross entropy (BCE) loss to use partial labels for single-label predictions. Our partial labels consist of both the class-positive and class-negative labels obtained during the image annotation process (Sect. 3). Notice that class-negative images are, in fact, hard negatives for the corresponding classes because of the way they were selected during labeling: they are either false-positive results returned from the image search engine or false-positive predictions with high confidence scores using our model. More formally, we modify BCE by introducing a weight vector to mask the loss where we've obtained partial labels. This is given by the equation:

$$\text{Loss} = \sum_{x_i, y_i, w_i \in X, Y, W} [w_i[y_i \log(A(x_i)) + (1 - y_i) \log(1 - A(x_i))]] \tag{1}$$

where A is the activation function (typically a sigmoid), X the prediction, Y the target, and W the weight vector. $X, Y, W \in \mathbb{R}^N$, and N is the number of classes.

For a training image with a class-positive label, we set $y_i = 1$ and $W = 1^N$ because we can conclude all information is known (i.e., due to our single-label assumption, the image is considered as negative for all the other classes). For a class-negative training image of the class i, we set $y_i = 0$ and $w_i = 1$. We do not set $W = 1^N$ in this case because we do not have ground truth positive or negative labels for the rest of the classes (different incidents may or may not appear in the image). Hence, for any unknown class j, i.e., $j \neq i$, we set $w_j = 0$.

The final loss $\mathcal{L}$ is given by the sum of the incidents loss $\mathcal{L}_d$ and the place loss $\mathcal{L}_p$, where both $\mathcal{L}_d$ and $\mathcal{L}_p$ are given by Eq. (1).

5 Experiments on the Incidents Dataset

Data. We split the images of the Incidents Dataset into training (90%), validation (5%), and a test (5%). As a reference, the training set contains 1,029,726 images, with 401,875 class-positive and 683,572 class-negative incident labels, and 151,665 class-positive and 265,415 class-negative place labels. Note that an image may have more than one class-negative label. Since the number of class-positive place labels is much lower than the number of class-positive incident labels, we augment the training set with 42,318 images from the Places dataset [84]. However, while training, we do not back-propagate the incidents loss on the additional Places images (which have no incident) since we already

Table 1. Ablation study. Performance comparison of the proposed model under different settings on both test sets. The best mAP is achieved by the model that uses CN loss with additional Places images as well as class negatives.

Architecture	Training with			Test set		Augmented test set	
	Loss	Class negatives	Additional places Images	Incident mAP	Place mAP	Incident mAP	Place mAP
ResNet-18	CE		✓	62.04	**47.85**	60.60	53.60
ResNet-18	CN		✓	61.15	46.61	59.88	53.41
ResNet-18	CN	✓		66.59	46.59	65.39	52.82
ResNet-18	CN	✓	✓	66.35	46.71	65.76	62.04
ResNet-50	CN	✓	✓	**67.65**	**47.56**	**67.19**	**63.20**

have class-negatives for the incidents that are better negative examples than these images from Places with no incidents.

The *test set* contains 57,215 images, and we also construct an *augmented test set* that is enriched with 2,365 extra images from Places that we assume contain no incidents. Unlike other image classification datasets, our *test set* contains class-negative images, which are important to evaluate the ability of a model to detect incidents in test images.

Incident Classification. We first evaluate the ability of our model to classify an image to the correct incident category, using just the images from the test set that belong to an incident category. Note that this experiment is similar to a within-the-dataset classification task where every test image belongs to a target category. We use a ResNet-18 [33] as backbone and train the model using the class-negative loss. We evaluate the incidents classification accuracy only on the part of test set that has positive incident labels. The top-1 accuracy is 77.3%, while the top-5 accuracy is 95.9%. As a reference, the performance of the same architecture trained on the same images but with a cross-entropy loss, which is a more standard choice for this classification task, is only slightly better, with 78.9% top-1 and 96.3% top-5 accuracy.

Incident Detection. We consider here a more realistic scenario of detecting incidents in images, evaluating the performance of the model on the whole test set that also includes images with negative labels. We measure this performance using the average precision (AP) metric, and we report the mean over all categories (mAP) for both the incidents and the places.

The obtained results are shown in Table 1, that presents, in fact, an ablation study exploring the use of different model architectures (ResNet-18 and ResNet-50), losses (cross-entropy and class-negative), and training data. Each model is pre-trained on the Places365 dataset [84] for the task of scene classification and then fine-tuned with the corresponding Incidents training data until convergence (at least 10 epochs). We used the Adam optimizer with an initial learning rate of 1e−4 and a batch size of 256 images, with shuffling. For each model, we report the incident and the place mAP on both the *test set* and the *augmented test set*.

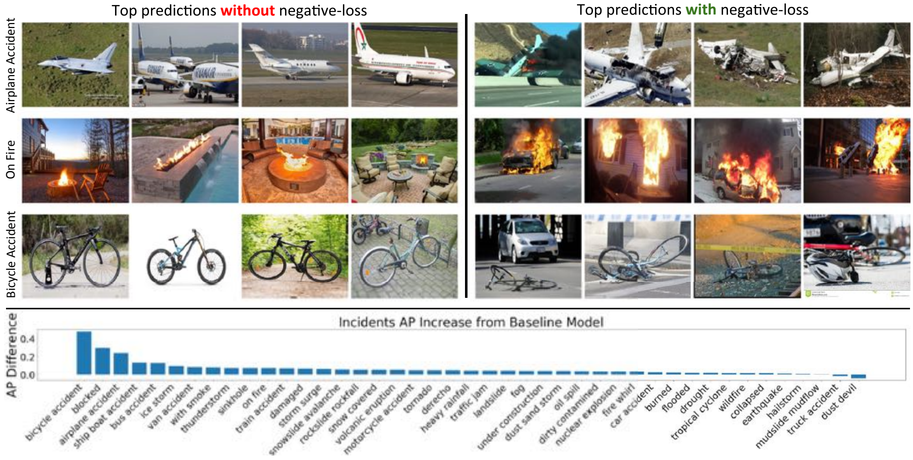


Fig. 3. Using the class-negative loss. Top confidence images for "airplane accident", "on fire", and "bicycle accident" categories when training without (left) and with (right) the class-negative loss. (Bottom) We report incident AP increments achieved by our model over the baseline model.

We observe that the incident mAP significantly improves by 4.3% (on the test set) to 5.2% (on the augmented test set) when we move from the cross-entropy (CE) loss to the class-negative loss (CN) using the class negatives (first and fourth row of Table 1). Figure 3 shows some top-ranked images for three incident categories by these two models. We can observe that, without using the class negatives during training, the model is not able to distinguish the difference between a fireplace and a house on fire or detect when a bicycle is broken because of an accident. The bottom of Fig. 3 shows the change in AP per incident category achieved by the CN model over the CE model. Notice that for nearly all incident categories the AP is much higher with CN model.

As a reference, the performance of the CN loss without using any class negatives, which corresponds to the standard BCE loss, is only less than 1% worse than the CE (first and second row of Table 1). Using additional Places images during training does not affect the incident detection but it vastly improves the place detection, especially in the case of the augmented test set, where mAP increases by 9.2% (third and fourth row of Table 1). Switching from a ResNet-18 to a deeper ResNet-50 architecture gives an extra final boost of incident mAP performance by 1.3% (fifth row of Table 1).

To further demonstrate the improved performance of our model trained with the CN loss (final), we compare it against the model trained with a CE loss (baseline) on 208 hand-selected hard-negative images used for MTurk quality control and not seen during training. Our final model recognizes 176 images correctly as true negatives with confidence score below 0.5 (85% accuracy) while the baseline model predicts the majority of them as false positives (30% accuracy).

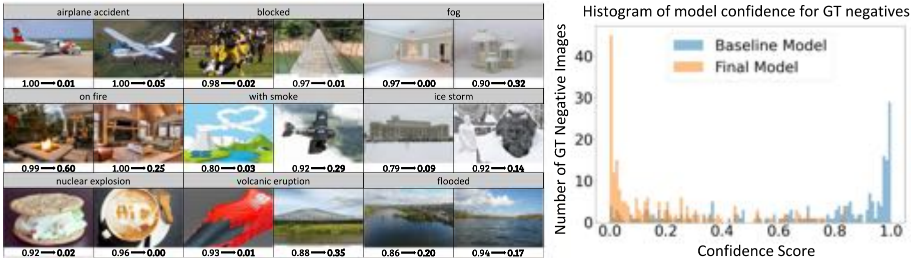


Fig. 4. GT negative test. (Left) Sample images withheld for quality control on MTurk are GT negatives not seen by the model during training. We report the changes in confidence scores between the baseline and final model below each image. (Right) We visualize the distribution of confidence scores obtained by both models for all 208 GT negative images. Our final model is more conservative when predicting incident confidence scores for hard-negative examples.

In Fig. 4, we investigate the changes in the confidence scores between the baseline and final models. Figure 4 (left) displays some qualitative examples of false positives for different incident categories. We observe that the confidence scores significantly decrease when using the final model. More concretely, the final model does not associate airplane features blindly to airplane accident, does not confuse rivers with flood scenes, or does not mistake clouds as smoke. Figure 4 (right) shows the distribution of confidence scores obtained by the baseline and final models. Notice that a perfect detector should assign 0 score to all of these images. Overall, this analysis shows, consistently with the other experiments explained in this section, how our final model is more robust against difficult cases, which is very important for filtering disaster images in the wild.

6 Detecting Incidents in Social Media Images

In this section, we examine how our incident detection model, trained with class-negative loss, performs in three different real-world scenarios using millions of images collected from two popular social media platforms: Flickr and Twitter.

6.1 Incident Detection from Flickr Images

The goal of this experiment is to illustrate how our model can be used to detect specific incident categories in the wild. For this purpose, we use 40 million geo-tagged Flickr images obtained from the YFCC100M dataset [76]. Since the images have precise geo-coordinates from EXIF data, we can use our incident detection model to filter for specific incidents and compare distance to ground-truth locations. We evaluate only earthquake and volcanic eruption incidents in this experiment as we could find reasonable ground-truth data to compare the results. Specifically, we downloaded the GPS coordinates, i.e., latitude and

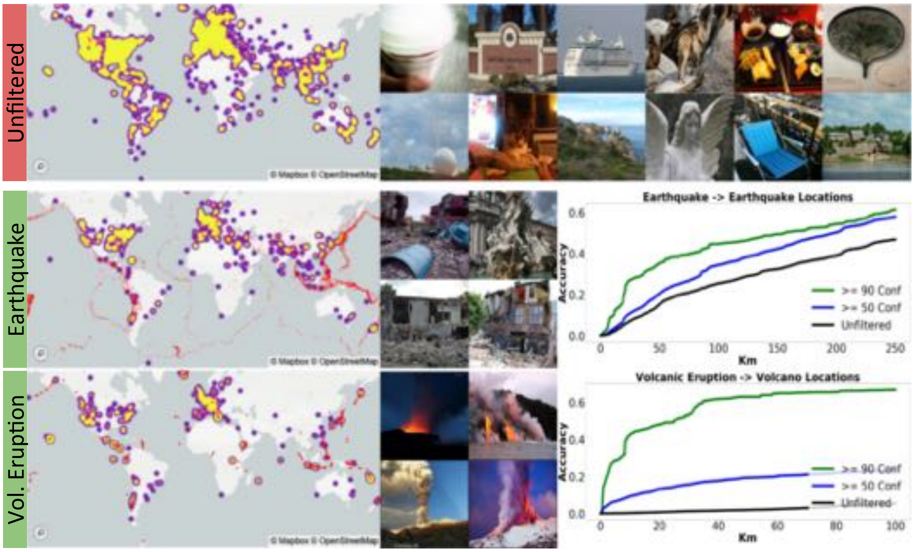


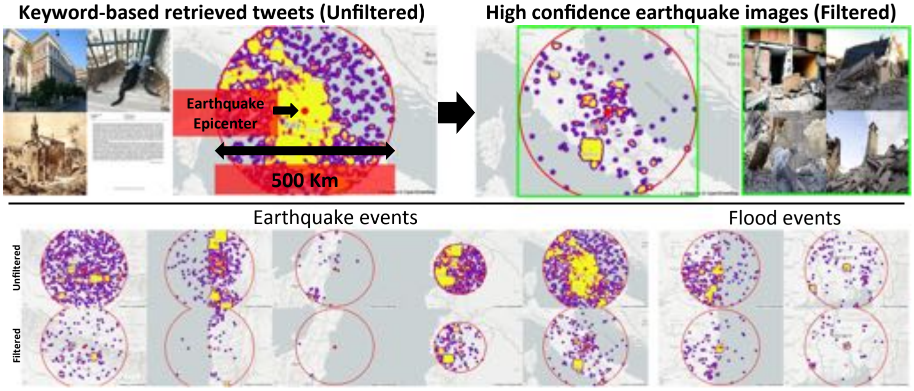
Fig. 5. Filtering Flickr images. (Top) Left: map visualization of Flickr image locations (complete unfiltered set). Right: random Flickr images. (Middle) Earthquake filtering. Left: map visualization of the location of images filtered by the earthquake category (earthquake epicenters are displayed as red dots). Middle: examples of images filtered with the earthquake category. Right: Accuracy@XKm, defined as the percent of images within X kilometers from an epicenter. When filtering with a confidence threshold above 0.9 (green), images are much closer to epicenters than in the unfiltered case (black). (Bottom) Volcanic eruptions and volcanoes, with the same structure as the earthquake experiment. (Color figure online)

longitude, of volcanoes from the National Oceanic and Atmospheric Administration (NOAA) website[1] and a public compilation of earthquake epicenters[2]. We employ an Accuracy@XKm metric [29] to determine whether the predicted incident is correct or not. More concretely, we compute the percentage of images within X Km from the closest earthquake epicenter or volcano, respectively. We randomly sample images and report metrics for (i) unfiltered images, (ii) images with model confidence above 0.5, and (iii) images with model confidence above 0.9. Figure 5 shows that detected earthquake and volcanic eruption incidents appear much closer to expected locations when compared to random images.

6.2 Incident Detection from Twitter Images

In this experiment we aim to detect earthquakes and floods in noisy Twitter data posted during actual disaster events. We collected data from five earthquake and two flood events using event-specific hashtags and keywords. In total, 901,127

[1] https://www.noaa.gov/.
[2] https://raw.githubusercontent.com/plotly/datasets/master/earthquakes-23k.csv.


Keyword-based retrieved tweets (Unfiltered)
Earthquake Epicenter
500 Km
High confidence earthquake images (Filtered)
Earthquake events
Flood events
Unfiltered
Filtered

	Nepal 2015	Chile 2015	Japan 2016	Ecuador 2016	Italy 2016	Chennai 2015	Bangladesh 2017
Unfiltered AP	23.0	10.5	30.9	25.8	7.4	27.2	48.3
Filtered AP	**77.2**	**55.7**	**80.4**	**85.0**	**68.9**	**89.2**	**87.5**

Fig. 6. Twitter image filtering. (Top) Experiment outline for the earthquake filtering (we follow the same outline for floods): we consider all tweets within a 250 Km radius of the epicenter of a specific event and then we filter the images for the earthquake category. Left part shows image examples and location of the unfiltered images, while right part shows locations and examples of filtered images. (Middle) Locations of unfiltered (top) and filtered images (bottom) are shown for each one of the seven events (five earthquakes and two floods), respectively. (Bottom) Ground-truth labels obtained from MTurk for each event are used to compute the AP for unfiltered (top row) and filtered (bottom row) images. Notice that our model significantly outperforms the unfiltered baseline.

images were downloaded. Twitter GPS coordinates are not nearly as precise as the Flickr ones, so we consider only the 39,494 geo-located images within 250 Km from either (i) the earthquake epicenter or (ii) the flooded city center.

For all seven events shown in Fig. 6, we use MTurk to obtain ground-truth human labels (i.e., earthquake or not, and flood or not) for images within the considered radius. Then, we compare the quality of the initial set of the keyword-based retrieved Twitter images (unfiltered) to the quality of images retained by our model (filtered). We report the average precision (AP) per event for both earthquakes and floods. When considering all earthquake events and flood events, we obtain a average AP of 73.9% and 89.1% compared to the baseline AP of 11.9% and 28.2%, respectively. The baseline AP is the AP averaged over multiple trials of randomly shuffling the images, and it is given as a reference.

6.3 Temporal Monitoring of Incidents on Twitter

In this section we demonstrate how our model can be used on Twitter data stream to detect specific incidents in time. To test this, we downloaded 1,946,850 images from tweets containing natural disaster keywords (e.g., blizzard, tornado, hurricane, earthquake, active volcano, coastal flood, wildfire, landslide)

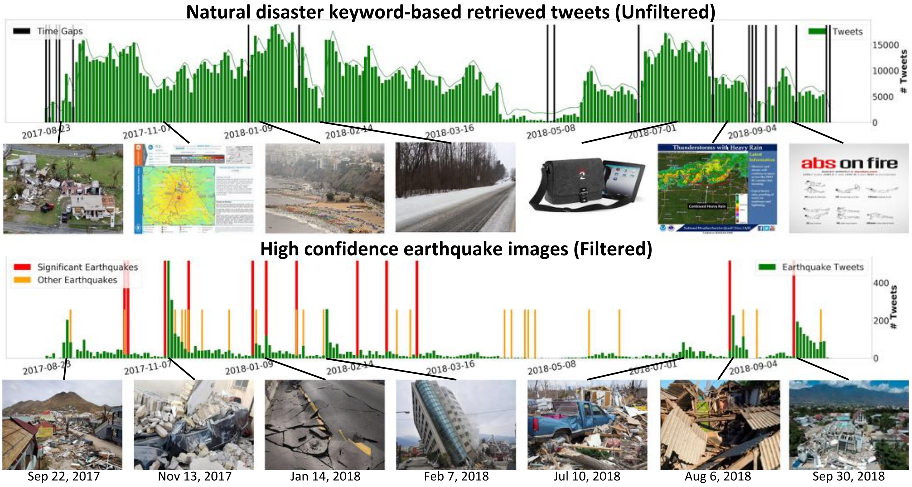
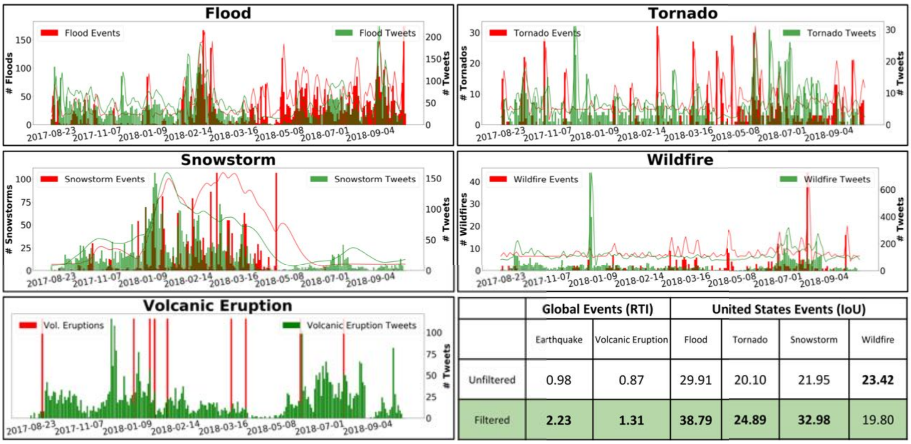


Fig. 7. Finding peaks in earthquake tweets. (Top) Histogram of tweets obtained from Twitter using natural disaster keywords from 2017–2018. Black lines indicate periods of time when our data collection server was inactive. (Bottom) Number of tweets with earthquake images per day after filtering with at least 0.5 confidence. For significant earthquakes (above 6.5 magnitude), we notice an increase in earthquake images immediately after the event. Furthermore, we notice a spike on July 20, 2018 not reported in the NOAA database. We manually checked the tweets and found images referring to a severe flood in Japan, indicating that the flood damage may resemble earthquake damage.

from Aug. 23, 2017 to Oct. 15, 2018. To quantify detection results, we obtained ground-truth event records from the “Significant Earthquake Database”, the “Significant Volcanic Eruption Database”, and the “Storm Events Database” of NOAA. The earthquake and volcanic eruptions ground-truth events are rare *global* events, while the storms (floods, tornadoes, snowstorms and wildfires) are much more frequent but reported only for the *United States*. We filter images with at least 0.5 confidence and compare against the databases (Fig. 7).

For earthquakes and volcanic eruptions, we report average Relative Tweet Increase (RTI) inspired by [6]. $\mathrm{RTI}_e = \sum_{d=e}^{e+w} N_d / \sum_{d=e}^{e-w} N_d$, where N_d is the number of relevant images posted on day d, e is the event day (e.g., day of earthquake), and w is an interval of days. We use $w = 7$ for our analysis to represent a week before and after an event. An RTI of 2 means that the average number of tweets in the week following an event is twice as high as the average number the week before. After filtering, the mean RTI ($\mathrm{mRTI} = \sum_{e \in E} RTI_e / |E|$) shows an average of 2.42 folds increase in tweets the week after an earthquake and 1.31 folds after a volcanic eruption (Fig. 8).

We notice that the mRTI would be even better if the ground truth databases were exhaustive. On Nov. 27, 2017 we detect the highest number of volcanic eruption images, but observe no significant eruption in the database. Looking



	Global Events (RTI)		United States Events (IoU)			
	Earthquake	Volcanic Eruption	Flood	Tornado	Snowstorm	Wildfire
Unfiltered	0.98	0.87	29.91	20.10	21.95	**23.42**
Filtered	**2.23**	**1.31**	**38.79**	**24.89**	**32.98**	19.80

Fig. 8. Temporal tweet filtering results. (Rows 1–2) For frequent events in the United States, we filter tweets for floods, tornadoes, snowstorms, and wildfires images and compare with ground truth frequency events obtained from NOAA. (Bottom Left) Filtered volcanic eruption images with ground truth events. (Bottom Right) Reported mRTI for global events and IoU for common US events.

into this, we found that Mount Agung erupted the same day, which caused the airport in Bali, Indonesia to close and left many tourists stranded[3].

For the more common events (e.g., tornadoes and snowstorms), we measure the correlation between tweet frequency and event frequency. We normalize both histograms, smooth with a low-pass filter, and report intersection over union (IoU) for United States incidents in Fig. 8. We notice an increase in IoU after filtering for flood, tornado, and snowstorm images. For wildfires, we notice a decrease in IoU and attribute this to the large spike in tweets in December 2017. Frequency correlation does not represent damage extent. In fact, a destructive wildfire occurred in California on Dec. 4, 2017 burning 281,893 acres[4].

7 Conclusion

In this paper, we explored how to automatically and systematically detect disasters, damage, and incidents in social media images in the wild. We presented the large-scale Incidents Dataset, which consists of 446,684 human-labeled scene-centric images that cover a diverse set of 43 incident categories (e.g., earthquake, wildfire, landslide, tornado, ice storm, car accident, nuclear explosion, etc.) in various scene contexts. Different from common practice, the Incidents Dataset includes an additional 697,464 class-negative images which can be used as hard negatives to train a robust model for detecting incidents in the wild. To that end,

[3] https://en.wikipedia.org/wiki/Mount_Agung.

[4] https://en.wikipedia.org/wiki/Thomas_Fire.

we also used a class-negative loss that capitalizes on this phenomenon. We then showed how the resulting model can be used in different settings for identifying incidents in large collections of social media images. We hope that these contributions will motivate further research on detecting incidents in images, and also promote the development of automatic tools that can be used by humanitarian organizations and emergency response agencies.

Acknowledgments. This work is supported by the CSAIL-QCRI collaboration project and RTI2018-095232-B-C22 grant from the Spanish Ministry of Science, Innovation and Universities.

References

1. Abavisani, M., Wu, L., Hu, S., Tetreault, J., Jaimes, A.: Multimodal categorization of crisis events in social media. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
2. Abdur Rehman, N., Saif, U., Chunara, R.: Deep landscape features for improving vector-borne disease prediction. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
3. Alam, F., Ofli, F., Imran, M.: Processing social media images by combining human and machine computing during crises. Int. J. Hum. Comput. Interact. **34**(4), 311–327 (2018)
4. Arietta, S.M., Efros, A.A., Ramamoorthi, R., Agrawala, M.: City forensics: using visual elements to predict non-visual city attributes. IEEE Trans. Visual Comput. Graphics **20**(12), 2624–2633 (2014)
5. Attari, N., Ofli, F., Awad, M., Lucas, J., Chawla, S.: Nazr-CNN: object detection and fine-grained classification in crowdsourced UAV images. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA) (2016)
6. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., Tesconi, M.: Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management. In: SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1749–1758. ACM (2014)
7. Ben-Haim, Z., et al.: Inundation modeling in data scarce regions. In: NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (2019)
8. Bischke, B., Helber, P., Schulze, C., Venkat, S., Dengel, A., Borth, D.: The multimedia satellite task at mediaeval 2017: emergency response for flooding events. In: Proceedings of the MediaEval 2017 Workshop, pp. 1–3 (2017)
9. Bischke, B., Helber, P., Zhao, Z., De Bruijn, J., Borth, D.: The multimedia satellite task at mediaeval 2018: emergency response for flooding events. In: Proceedings of the MediaEval 2018 Workshop, pp. 1–3 (2018)
10. Bonafilia, D., Gill, J., Basu, S., Yang, D.: Building high resolution maps for humanitarian aid and development with weakly- and semi-supervised learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
11. Can, G., Benkhedda, Y., Gatica-Perez, D.: Ambiance in social media venues: visual cue interpretation by machines and crowds. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2018)

12. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997). https://doi.org/10.1023/A:1007379606734
13. Castillo, C.: Big Crisis Data. Cambridge University Press, New York (2016)
14. Chehata, N., Orny, C., Boukir, S., Guyon, D., Wigneron, J.: Object-based change detection in wind storm-damaged forest using high-resolution multispectral images. Int. J. Remote Sens. **35**(13), 4758–4777 (2014)
15. Chen, T., Lu, D., Kan, M.Y., Cui, P.: Understanding and classifying image tweets. In: ACM International Conference on Multimedia, pp. 781–784 (2013)
16. Daly, S., Thom, J.: Mining and classifying image posts on social media to analyse fires. In: 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM), pp. 1–14 (2016)
17. Dell'Acqua, F., Gamba, P.: Remote sensing and earthquake damage assessment: experiences, limits, and perspectives. Proc. IEEE **100**(10), 2876–2890 (2012)
18. Deng, J., Dong, W., Socher, R., Li, L., Kai, L., Li, F.-F.: ImageNet: a large-scale hierarchical image database. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
19. Dong, L., Shan, J.: A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogrammetry Remote Sens. **84**, 85–99 (2013)
20. Doshi, J., Basu, S., Pang, G.: From satellite imagery to disaster insights. In: NeurIPS Workshop on Artificial Intelligence for Social Good (2018)
21. Doshi, J., et al.: FireNet: real-time segmentation of fire perimeter from aerial video. In: NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (2019)
22. Durand, T., Mehrasa, N., Mori, G.: Learning a deep convnet for multi-label classification with partial labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
23. Efremova, N., West, D., Zausaev, D.: AI-based evaluation of the SDGs: the case of crop detection with earth observation data. In: ICLR Workshop on Artificial Intelligence for Social Good (2019)
24. Fernandez Galarreta, J., Kerle, N., Gerke, M.: UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. Nat. Hazards Earth Syst. Sci. **15**(6), 1087–1101 (2015)
25. Gamba, P., Dell'Acqua, F., Trianni, G.: Rapid damage detection in the bam area using multitemporal SAR and exploiting ancillary data. IEEE Trans. Geosci. Remote Sens. **45**(6), 1582–1589 (2007)
26. Gebru, T., et al.: Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States. Proc. Nat. Acad. Sci. **114**(50), 13108–13113 (2017)
27. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Fei-Fei, L.: Fine-grained car detection for visual census estimation. In: The AAAI Conference on Artificial Intelligence (2017)
28. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
29. Gritta, M., Pilevar, M.T., Collier, N.: A pragmatic guide to geoparsing evaluation: toponyms, named entity recognition and pragmatics. Lang. Resour. Eval. (2019). https://doi.org/10.1007/s10579-019-09475-3
30. Gueguen, L., Hamid, R.: Large-scale damage detection using satellite imagery. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1321–1328 (2015)

31. Gupta, R., et al.: Creating xBD: a dataset for assessing building damage from satellite imagery. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
32. Gupta, R., et al.: xBD: a dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296 (2019)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
34. Helber, P., et al.: Mapping informal settlements in developing countries with multi-resolution, multi-spectral data. In: ICLR Workshop on Artificial Intelligence for Social Good (2019)
35. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. ACM Comput. Surv. **47**(4), 67 (2015)
36. Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., Ermon, S.: Combining satellite imagery and machine learning to predict poverty. Science **353**(6301), 790–794 (2016)
37. Joyce, K.E., Belliss, S.E., Samsonov, S.V., McNeill, S.J., Glassey, P.J.: A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. Prog. Phys. Geogr. **33**(2), 183–207 (2009)
38. Kalliatakis, G., Ehsan, S., Fasli, M., D McDonald-Maier, K.: DisplaceNet: recognising displaced people from images by exploiting dominance level. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
39. Kalliatakis, G., Ehsan, S., Leonardis, A., Fasli, M., McDonald-Maier, K.D.: Exploring object-centric and scene-centric CNN features and their complementarity for human rights violations recognition in images. IEEE Access **7**, 10045–10056 (2019)
40. Kaneko, A., et al.: Deep learning for crop yield prediction in Africa. In: ICML Workshop on Artificial Intelligence for Social Good (2019)
41. Kataoka, H., Satoh, Y., Abe, K., Minoguchi, M., Nakamura, A.: Ten-million-order human database for world-wide fashion culture analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
42. Kellenberger, B., Marcos, D., Tuia, D.: When a few clicks make all the difference: improving weakly-supervised wildlife detection in UAV images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
43. Lee, K., Lee, H., Lee, K., Shin, J.: Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations (2018)
44. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 7167–7177. Curran Associates, Inc. (2018)
45. Li, X., Caragea, D., Caragea, C., Imran, M., Ofli, F.: Identifying disaster damage images using a domain adaptation approach. In: 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM) (2019)
46. Li, Y., Hu, W., Dong, H., Zhang, X.: Building damage detection from post-event aerial imagery using single shot multibox detector. Appl. Sci. **9**(6), 1128 (2019)
47. Li, Y., Ye, S., Bartoli, I.: Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. J. Appl. Remote Sens. **12**(4), 045008 (2018)
48. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)

49. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
50. McKinney, S.M., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
51. Nachmany, Y., Alemohammad, H.: Detecting roads from satellite imagery in the developing world. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
52. Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: Streetscore - predicting the perceived safety of one million streetscapes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 793–799 (2014)
53. Naik, N., Kominers, S.D., Raskar, R., Glaeser, E.L., Hidalgo, C.A.: Computer vision uncovers predictors of physical urban change. Proc. Nat. Acad. Sci. (2017). https://doi.org/10.1073/pnas.1619003114
54. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1–8 (2017)
55. Nia, K.R., Mori, G.: Building damage assessment using deep learning and ground-level image data. In: 14th Conference on Computer and Robot Vision (CRV), pp. 95–102. IEEE (2017)
56. Nogueira, K., et al.: Exploiting convnet diversity for flooding identification. IEEE Geosci. Remote Sens. Lett. **15**(9), 1446–1450 (2018)
57. Oh, J., Hebert, M., Jeon, H.G., Perez, X., Dai, C., Song, Y.: Explainable semantic mapping for first responders. In: NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (2019)
58. Peters, R., de Albuquerque, J.P.: Investigating images as indicators for relevant social media messages in disaster management. In: 12th International Conference on Information Systems for Crisis Response and Management (ISCRAM) (2015)
59. Piaggesi, S., et al.: Predicting city poverty using satellite imagery. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
60. Plank, S.: Rapid damage assessment by means of multi-temporal SAR–a comprehensive review and outlook to sentinel-1. Remote Sens. **6**(6), 4870–4906 (2014)
61. Pouyanfar, S., et al.: Unconstrained flood event detection using adversarial data augmentation. In: IEEE International Conference on Image Processing (ICIP), pp. 155–159 (2019)
62. Pryzant, R., Ermon, S., Lobell, D.: Monitoring ethiopian wheat fungus with satellite imagery and deep feature learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2017)
63. Radhika, S., Tamura, Y., Matsui, M.: Cyclone damage detection on building structures from pre-and post-satellite images using wavelet based pattern recognition. J. Wind Eng. Ind. Aerodyn. **136**, 23–33 (2015)
64. Radke, D., Hessler, A., Ellsworth, D.: FireCast: leveraging deep learning to predict wildfire spread. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 4575–4581 (2019)
65. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Learning multiple visual domains with residual adapters. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 506–516 (2017)
66. Reuter, C., Kaufhold, M.A.: Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. J. Contingencies Crisis Manage. **26**(1), 41–57 (2018)

67. Rudner, T.G.J., et al.: Multi^{3}Net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In: The AAAI Conference on Artificial Intelligence, pp. 702–709 (2019)
68. Rustowicz, R., Cheong, R., Wang, L., Ermon, S., Burke, M., Lobell, D.: Semantic segmentation of crop type in Africa: a novel dataset and analysis of deep learning methods. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
69. Said, N., et al.: Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools Appl. **78**, 31267–31302 (2019)
70. Schmidt, V., et al.: Visualizing the consequences of climate change using cycle-consistent adversarial networks. In: ICLR Workshop on Artificial Intelligence for Social Good (2019)
71. Seo, J., Lee, S., Kim, B., Jeon, T.: Revisiting classical bagging with modern transfer learning for on-the-fly disaster damage detector. In: NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (2019)
72. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
73. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
74. Skakun, S., Kussul, N., Shelestov, A., Kussul, O.: Flood hazard and flood risk assessment using a time series of satellite images: a case study in Namibia. Risk Anal. **34**(8), 1521–1537 (2014)
75. Sublime, J., Kalinicheva, E.: Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku Tsunami. Remote Sens. **11**(9), 1123 (2019)
76. Thomee, B., et al.: YFCC100M: the new data in multimedia research. Commun. ACM **59**(2), 64–73 (2016)
77. Tingzon, I., et al.: Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. In: ICML Workshop on Artificial Intelligence for Social Good (2019)
78. Turker, M., San, B.T.: Detection of collapsed buildings caused by the 1999 Izmit, Turkey earthquake through digital analysis of post-event aerial photographs. Int. J. Remote Sens. **25**(21), 4701–4714 (2004)
79. de Vries, T., Misra, I., Wang, C., van der Maaten, L.: Does object recognition work for everyone? In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2019)
80. Watmough, G.R., et al.: Socioecologically informed use of remote sensing data to predict rural household poverty. Proc. Nat. Acad. Sci. **116**(4), 1213–1218 (2019)
81. Workman, S., Zhai, M., Crandall, D.J., Jacobs, N.: A unified model for near and remote sensing. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
82. Wu, N., et al.: Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans. Med. Imaging **39**(1), 1184–1194 (2019)

83. Xu, J.Z., Lu, W., Li, Z., Khaitan, P., Zaytseva, V.: Building damage detection in satellite imagery using convolutional neural networks. In: NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (2019)
84. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 1452–1464 (2018)
85. Zhou, B., Liu, L., Oliva, A., Torralba, A.: Recognizing city identity via attribute analysis of geo-tagged images. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 519–534. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_34