Damage Estimation and Localization from Sparse Aerial Imagery

1st René García Franceschini
Institute for Data, Systems and Society
Massachusetts Institute of Technology
Cambridge, MA
ragarcia@mit.edu

2nd Jeffrey Liu
MIT Lincoln Laboratory
Lexington, MA
Jeffrey.Liu@ll.mit.edu

3rd Saurabh Amin
Dept. of Civil and Environmental Engineering
Massachusetts Institute of Technology
Cambridge, MA
amins@mit.edu

***Abstract*—Aerial images provide important situational awareness for responding to natural disasters such as hurricanes. They are well-suited for providing information for damage estimation and localization (DEL); *i.e.*, characterizing the type and spatial extent of damage following a disaster. Despite recent advances in sensing and unmanned aerial systems technology, much of post-disaster aerial imagery is still taken by handheld DSLR cameras from small, manned, fixed-wing aircraft. However, these handheld cameras lack IMU information, and images are taken opportunistically post-event by operators. As such, DEL from such imagery is still a highly manual and time-consuming process. We propose an approach to both detect damage in aerial images and localize it in world coordinates, with specific focus on detecting and localizing flooding. The approach is based on using structure from motion to relate image coordinates to world coordinates via a projective transformation, using class activation mapping to detect the extent of damage in an image, and applying the projective transformation to localize damage in world coordinates. We evaluate the performance of our approach on post-event data from the 2016 Louisiana floods, and find that our approach achieves a precision of 88%. Given this high precision using limited data, we argue that this approach is currently viable for fast and effective DEL from handheld aerial imagery for disaster response.**

***Index Terms*—Weakly-supervised learning, Class activation mapping, Structure from motion, GIS, Disaster response.**

I. Introduction

Natural disasters, such as hurricanes and floods, can cause major loss of life and property; the intensity, scope, and the frequency of such disasters may be further exacerbated by global climate change [1]. Timely information about the distribution and nature of damage following a disaster can help provide important context and information for emergency managers' decision-making [2]. Increasingly, satellite and aerial imagery are being incorporated into post-disaster needs assessment [3]. However, while techniques exist for extracting information from orthorectified satellite and aerial imagery [4]–[8], methods for more general aerial imagery (such as oblique imagery from handheld cameras) have received far less attention. This is a critical limitation because post-disaster aerial imagery taken from handheld DSLR cameras from small, manned, fixed-wing aircraft remains popular due to the relatively low cost, high availability, conformity with existing regulations, and existing training programs associated with the practice [9], [10]. These handheld cameras lack IMU information, and images are taken opportunistically post-event by human operators, resulting in sparsely-sampled images taken at oblique angles. In this paper, we pose the question: **how can we use aerial imagery from an arbitrary camera setup in order to rapidly and effectively aid in post-disaster situational awareness?**

We focus on a specific component of post-disaster needs assessment, which we refer to as Damage Estimation and Localization (DEL). We broadly define *damage* as an identifiable destruction of an infrastructure component or utility resulting from a specific event (in our case, a natural disaster). We then define *estimation* as the detection of an instance of damage in an image. Finally, *localization* is the act of assigning world coordinates to the estimated instance of damage. Our main contribution is a practically implementable approach that uses sparse, oblique aerial disaster imagery from handheld cameras to carry out DEL, with a specific focus on DEL for images of flooding. To our knowledge, our approach is the only one that does estimation without relying on training data that includes bounding boxes or segmented images; and localization without inertial measurement unit (IMU) information or a known geotransform. We show that this method achieves a precision of 88% when compared against official estimates from the 2016 Louisiana floods. Furthermore, our approach can readily incorporate the types of data that other approaches rely on should these datasets become available in the future. We believe this approach is an important contribution to the disaster relief sector for two reasons. First, it provides disaster relief responders the means to do fast and effective DEL without

This material is based upon work supported by the United States Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

This work was also made possible in part due to funding from NSF D-ISN project award # 2039771.

978-1-6654-4337-1/21/$31.00 ©2021 IEEE
DOI 10.1109/ICMLA52953.2021.00028

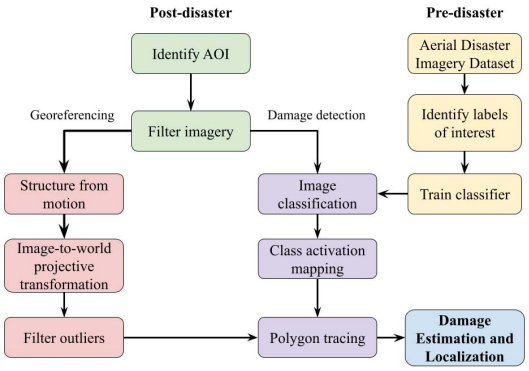


Fig. 1: Flowchart depicting our approach.

the need of sophisticated equipment. Second, it augments the utility of existing datasets (such as the one we use, described in Section III-A) by providing georeferenced damage annotations to a significant portion of the imagery.

Figure 1 provides a visual depiction of our approach. It consists of two stages: a pre-disaster and a post-disaster stage. In the pre-disaster stage, a neural network is trained to recognize the image-level damage labels within an aerial disaster imagery dataset. The post-disaster stage is comprised of two parallel 'pipelines', whose outputs are combined at the end. The first pipeline takes a collection of images from an area of interest and reconstructs the scene using structure from motion. The reconstructed point cloud then relates image coordinates to world coordinates via a projective transformation. The second pipeline takes individual images from the area of interest and produces polygons that cover the extent of the damage that is detected using class activation mapping. The projective transformation is then applied to the damage polygons to produce the final output. While our approach uses image-level binary damage indicators and class activation maps due to the limited availability of training data, the same estimated projective transformation could be applied to bounding boxes or segmentation masks generated from object detection or semantic segmentation algorithms, respectively.

II. Novelty of our Approach

We propose the approach detailed in Figure 1 for performing both components of DEL using the tools and datasets that currently exist for this context. Specifically for georeferencing, other common approaches for image registration turned out to be intractable for this application. While some authors have used visual feature-based methods such as SIFT to register satellite or top-down drone imagery to other known georeferenced images [11]–[14], methods such as SIFT have been shown to perform poorly under extreme changes in perspective and sensor specifications [13], [15]. This was the case when we attempted to use SIFT to georeference post-disaster aerial images and satellite images. Another approach georeferences images using Siamese neural networks in one of two ways: either training the network to match certain features (*e.g.* buildings) of a query image and a ground truth image [16] or by matching the entirety of a query image and a ground truth image [15], [17], [18]. While these approaches would be suitable for estimating the GPS tag of aerial images, estimating the full geotransform would require additional orientation information. Indeed, most available literature on registering oblique imagery relies on some variant of structure from motion or multiview stereo [19], which is the approach we take.

On the other hand, there is an abundance of literature and datasets on detecting damage after natural disasters. In particular, various deep learning approaches have detected damage with high accuracy in challenges such as xView2 [4]–[6]. However, such training data that estimates damage (either with bounding boxes or segmentation) consists of orthorectified satellite or aerial imagery instead of oblique aerial imagery. Previous work has also attempted to overcome lack of training data in either satellite or aerial images via transfer learning from satellite to aerial or vice versa [20]–[22]. Once again, all attempts at transfer learning that we are aware of were only applied to orthorectified imagery, and only only considered changes in resolution, not perspective. Moreover, while the remote sensing community has developed indexes to classify features such as water and vegetation [23] from multispectral sensors, this is not applicable to the imagery produced from the handheld cameras used in our context, which are only sensitive in the visible spectrum. Given the lack of tools and datasets developed for this space, we pursue a class activation mapping approach using image-level labels as a weakly supervised approach to detecting damage.

III. Methods

This section details the methods that support the key components of our approach. Section III-A first provides an overview of the dataset used in this analysis. Then, Sections III-B and III-C describe our estimation and localization pipelines, respectively.

A. Low Altitude Disaster Imagery dataset

We perform all our analyses using the Low Altitude Disaster Imagery (LADI) dataset [24]. LADI is a publicly available dataset consisting of images taken by the United States Civil Air Patrol (CAP) in the aftermath of natural disasters, and annotated by crowdsourced workers with hierarchical image-level labels representing five broad categories: Damage, Environment, Infrastructure, Vehicles, and Water. Within each category, there are a number of more specific annotation labels. We focus on the *flooding/water damage* label within in the "Damage" category.

While the current LADI dataset provides image-level annotations, it does not provide any bounding box or segmentation information. This limits our ability to train an object detection or image segmentation classifier to localize classes within the images. At the time of writing, we were unable to find any other publicly available datasets of post-disaster aerial

imagery from low altitude, oblique perspectives which provide bounding box or segmentation annotations. While we are aware that the Volan2018 dataset does provide bounding box information for disaster imagery for various related classes [25], to our current knowledge, it is not publicly available.

B. Damage estimation within an image

This section describes our damage estimation pipeline. We initially pose this as a classification problem of detecting a given type of damage within an image; from this classifier, we extract the class activation map to estimate the location of damage within the image.

1) Classification with ResNet: Our first step is to detect whether an image contains flooding anywhere in the image. We pose this problem as a classification problem, where for an image X_i there is a label $Y_i^{true} \in \{0, 1\}$ that corresponds to whether an image contains flooding or not. Our goal is to predict Y_i^{true}. In the construction of the LADI dataset, images were shown to a variable number of workers (generally between 3-5); each worker was asked to identify which, if any, labels for a given category (e.g. "Damage") applied to that image [24]. Responses from all workers were recorded. For the sake of simplicity, we propose three different empirical labelling schemes for determining their ground truth labels to account for differing annotations between workers:

A) $B_{i,j} > 1$,
B) $B_{i,j} > 2$,
C) $B_{i,j} > 1$ and $B_{i,j}/w_i > \underset{\forall i}{\text{median}}\{B_{i,j}/w_i\}$,

where $B_{i,j}$ is the number of workers that labelled image i as class j and w_i is the number of workers that labelled image i at all. We trained and tested using different combinations of these labelling scheme to determine a balance between filtering out noise and preserving a sufficiently representative dataset.

To perform the image labeling task, we use a ResNet-50 backbone architecture for this paper, due to its consistent performance in a variety of image classification tasks [26]. We split the dataset 80%/10%/10% for the training, validation and testing sets, respectively. Images were scaled such that the shorter dimension was 224 pixels long, and then cropped into a 224×224 tile. Random rotations and horizontal flips were applied during training for data augmentation. We initialized the ResNet with pre-trained ImageNet weights [27], and changed the output layer dimension from 1000 to 1. We then train the ResNet with a batch size of 8, a learning rate of 0.001, a momentum of 0.9, using stochastic gradient descent as the optimization algorithm with the Binary Cross-Entropy (BCE) loss:

$$L = \sum_{i=1}^{m} Y_i^{true} \log \sigma(Y_i^{pred}) + (1 - Y_i^{true}) \log \sigma(1 - Y_i^{pred}), \tag{1}$$

where $Y_i^{pred} \in \mathbf{R}$ is the output of the neural network, m is the number of images in the batch and $\sigma(\cdot)$ represents the Sigmoid function.

(a) Original image (b) CAM mask (c) Polygon tracing

Fig. 2: Stages of our polygon tracing approach.

2) Class activation mapping and polygon tracing: After training, we utilize the class activation mapping (CAM) approach from Zhou et al [28] to localize the extent of the detected class within the image. CAM is a technique for weakly-supervised object detection (*i.e.*, where bounding boxes are not explicitly trained on). It leverages the average pooling layer at the end of the ResNet architecture to detect areas within an image that are important for classifying a particular class. While newer variations of CAM exist, such as Grad-cam and Score-CAM [29], [30], we decided to proceed with the original implementation from Zhou et al. Using the terminology in [28], the output on the final fully connected layer is given by:

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y), \tag{2}$$

where w_k^c is the weight corresponding to class c for the k-th unit within conv5 (the final block within ResNet-50) and $f_k(x, y)$ is the activation of the same k-th unit at location (x, y) (such that $\sum_{x,y} f_k(x, y)$ is the output of the global pooling layer). Let $M_c(x, y) = \sum_k w_k^c f_k(x, y)$, so that:

$$S_c = \sum_{x,y} M_c(x, y). \tag{3}$$

Here, $M_c(x, y)$ can be viewed as a measure of importance of a spatial coordinate (x, y) for the class c = *flooding/water damage*, and hence referred to the class activation map. In order to determine the boundaries of the flooding instances, we threshold on M_c:

$$M_c^{\text{mask}}(x, y) = \begin{cases} 1 & \text{if } M_c(x, y) \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The last step in the estimation pipeline is to convert the masked image into a set of polygons using [31]. This enables us to easily transform the boundaries of flooding across coordinate systems. Fig. 2 shows different stages of this procedure.

C. Damage localization

Next, we transform these polygons, which are in image coordinates, into world coordinates; this forms our localization pipeline. In this section, we describe the process of using structure from motion to estimate the projective transformation relating image and world coordinates, and applying the transformation to the flooding polygons.

1) Reconstruction using structure from motion: Structure from motion is a technique that, using images from a camera moving through an environment, can produce a point cloud of the environment [32]. By taking advantage of the GPS tags from the image metadata or from outside sensors, structure from motion has been used to create inexpensive, georeferenced elevation models from drone and aircraft imagery [33]. We use this technique as an intermediate step to obtaining the projective transformation that relates image coordinates to world coordinates. We base our implementation off the well-known OpenSfM library, an open source library for structure from motion, following its default configuration. For more information, please refer to the OpenSfM documentation [34]. In our implementation, we chose a specific area of interest and select all images within the LADI dataset whose GPS tags were within a 5km buffer around the area of interest.

Since fixed wing aircrafts have a relatively large turning radius compared to rotary-wing aircraft, sequential images collected from fixed-wing platforms tend to be approximately collinear. This means that some reconstructions potentially have an additional degree of freedom from rotating about the line that goes through the GPS coordinates. Therefore, it is necessary to estimate the direction of the up-vector (*i.e.*, the vector opposite to the direction of gravity) and enforce it in the reconstruction. Previous implementations of structure from motion in urban environments have suggested estimating vanishing points to estimate the up-vector [35]. This can be difficult if there are few straight features (such as roads) or high amounts of vegetation, which is common in rural areas.

To address the issue of estimating the up-vector, we propose an approach which assumes the ground is approximately flat. We first fit a plane through the reconstructed features using RANSAC [36]. There is a pair of possible antiparallel unit normal vectors to this plane, one of which is the up-vector. Because of the aerial nature of the data, the location of the images must be above the ground plane. Therefore, we choose the vector that has a positive projection onto the image location in East, North Up (ENU) coordinates and denote it v_{up}. Finally, we rotate the reconstruction so that v_{up} indeed points upwards. Specifically, we rotate it by R_z such that $R_z v_{up} = \hat{z}$ when it is initialized, and the up-vector is enforced during bundle adjustment.

We initially incorporated a digital elevation model (DEM) of the local topology to realign the reconstruction. However, in the case discussed in this paper, the inclusion of the DEM did not yield any performance gains, since the region that we considered was relatively flat. Therefore, we do not report these results.

2) Image-to-world projective transformation: The final step in our georeferencing pipeline is estimating the transformation from image coordinates to world coordinates, and applying this transformation to the detected damage polygons. As discussed previously, the images are of mostly flat surfaces, meaning both sets of coordinates can be related by a projective transformation that can be estimated with at least four correspondences [32], and outliers can be filtered through RANSAC [36]. Of all of the images that were reconstructed using OpenSfM, we retained those where at least 20% of matches between image coordinates and world coordinates were inliers.

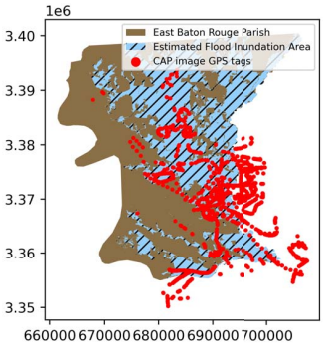


Fig. 3: Map of East Baton Rouge parish and image GPS tags.

Of the retained images, we found that some images produced extremely large image footprints (*i.e.*, the projection of the image edges onto the ground). Upon inspection, we saw that these were images that were so oblique that the horizon was visible. Because these images require more complex transformations, we decided to disregard these images for our implementation. We considered two criteria for eliminating such images. First, we only eliminated images whose total area were greater than some value γ_1. Second, we did not consider images where the ratio of the longest side to the shortest side of the minimum area rectangle that covered the entire footprint were greater than γ_2. The projective transformation is applied to all polygons generated by the procedure in Section III-B to obtain our flooding estimate. We report the results for a variety of combinations of γ_1 and γ_2 parameters to illustrate the effectiveness of our approach.

IV. Evaluation and Results

In this section, we evaluate the performance of our approach at DEL using images from the 2016 Louisiana floods. Figure 3 shows the administrative boundary of the East Baton Rouge parish in Louisiana, the parish's estimated flood inundation area [37], and the coordinates of all CAP image with GPS locations within 5 km of the administrative boundary. In total, the flooding event covered 536 km^2 (44% of the total area of the parish). Our analysis includes 1615 CAP images that were taken in August 2016 immediately after the flooding event.

A. Classification results

Table I shows the testing accuracy, precision and recall values for the three ResNet50 classifiers that were trained (one for each ground truth training labelling scheme defined in Section III-B1). In order to properly compare the three models, each of the three classifiers was also evaluated against the remaining two labelling schemes. Regardless of the labelling, the actual images that comprised the testing set (as well as the training and validation sets) were the same for all three

Train label	Test label	Accuracy (%)	Precision (%)	Recall (%)
A	A	77	65	79
	B	71	38	91
	C	70	42	83
B	A	77	75	53
	B	83	54	68
	C	80	55	64
C	A	79	73	65
	B	80	48	80
	C	83	53	69

TABLE I: Accuracy, precision and recall values for the three ResNet models. *Train label* refers to the set of labels used in training, while *Test label* refers to the set of labels against which each model was evaluated.

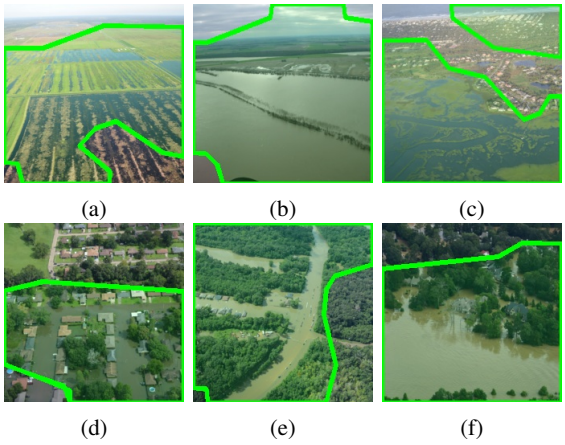


Fig. 4: Sample of CAP images identified as flooding by ResNet model A and their associated damage polygons.

schemes. For the purposes of this paper, we will refer to each of the models according to their training labelling scheme.

Unsurprisingly, each of three models had the highest accuracy when compared against the labelling scheme they were trained on. With the other two metrics, though, there are noticeable trends. In terms of precision, model B had the highest precision when evaluated against any labelling scheme, followed by C and finally A. With recall, the opposite holds: A has the highest recall across the board, followed by C and then B. These trends are not difficult to justify, since B necessarily has a higher standard for classification as flooding than A. For flooding, C is a compromise between the most lenient labelling scheme (A) and the strictest one (C). In this particular application, we consider false positives to be less serious than false negatives; that is, we would rather think that someone was in danger from flooding when they are not (false positive) than think that they are not in danger when they are (false negative). As such, we proceed using model A for the remainder of the section.

B. Class activation mapping results

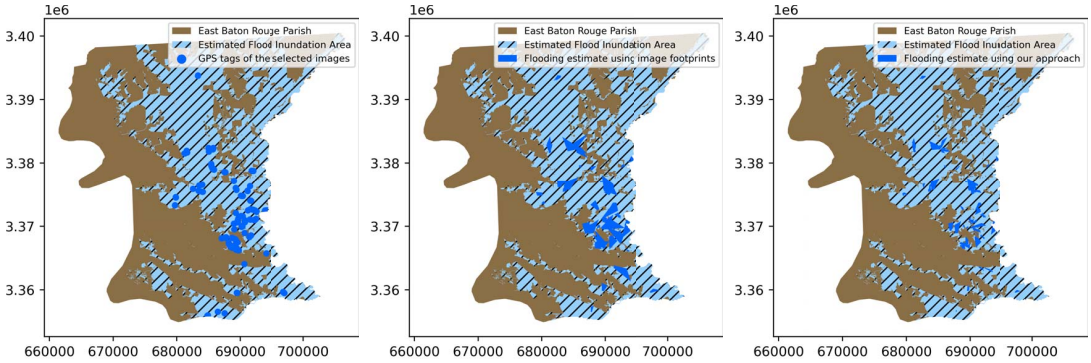
To get a better sense of how class activation mapping performs at identifying regions of *flooding/water damage* within an image, we show a few sample images with an overlay of the detected flooding outline. Figure 4 shows a sample of LADI images that were classified as having *flooding/water damage*, along with the estimated extent of flooding. We can see that for the most part, this the CAM provides a decent coarse estimate of the extent of water in the image. Flooding is slightly more complicated. While Figures 4a and 4b very clearly show flooding events, the top portion of Figure 4c seems to simply be picking up the shoreline. As an important note, we noticed that many images that show large bodies of water also tend to include the horizon in the flooding polygon (*e.g.* Figure 4b). This might be because flooding typically covers a large portion of area, and therefore images that include the horizon might be more likely to also include flooding. This underscores the importance of filtering images with large footprints after georeferencing. Figures 4d, 4e and 4f are images that were identified as flooding from the Louisiana 2016 floods. Even in this case where many images have large portions of flooding, our approach is still able to trace the extent of the water. While we would ideally want to provide a full performance evaluation, the lack of segmented images for this context makes this infeasible without significant effort put into manual labelling.

C. DEL results

Of the 1615 CAP images that were considered, 809 were successfully reconstructed by OpenSfM. At the same time, of the 1615 images 996 were identified as having flooding. Finally, 559 images completed the georeferencing pipeline *and* were identified as flooding. Additional images were the filtered based on the criteria described in Section III-C2.

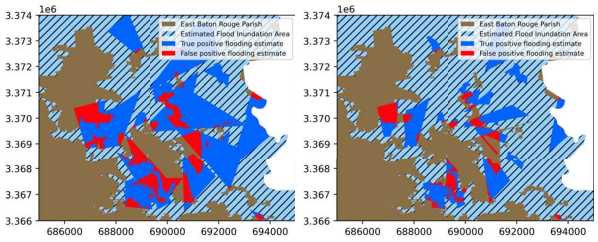
We used these images to estimate flooding in three different methods. Firstly, we use the GPS tag of these images as a baseline, where we calculate the precision as the proportion of the flood images that lie in the FEMA estimates. Secondly, we estimate the flooding using the entire footprints of images classified as containing *"flooding/water"* (SfM + binary classification). Finally, we consider our approach of both (SfM + CAM) as the flood estimate. We only consider the flooding within the East Baton Rouge administrative boundary, since we do not have data on flood extent outside of the boundary.

Figure 5 shows the different flooding estimates overlaid against the official estimates, as well as the precision values of each method. Note that the precision we report is compared against the flooding extent estimates, and not with ground truth mask of the images (which are not available in the dataset). In Figure 6, precision is the total area colored dark blue divided by the sum of the dark blue and red areas. The reported precision includes both the contributions of class activation mapping and structure from motion. These estimates were made with $\gamma_1 = 4$ and $\gamma_2 = 5$ km^2, so that ultimately 243 images were used. These results show a clear improvement going from using the GPS locations of the images to using the georeferenced footprint. This suggests that using the GPS tags of the images on their own is insufficient, since a large



(a) GPS tags (Precision: 52%) (b) Image footprints (Precision: 80%) (c) Our approach (Precision: 88%)

Fig. 5: Flooding estimates and precision values using: (a) GPS tags, (b) image footprints (SfM + binary classification), and (c) our approach (SfM + CAM).



(a) Image footprint (b) Our approach

Fig. 6: Close-up of flooding estimates for image footprint and our approach, showing true and false positive regions.

number of images containing flooding were taken over areas that were not flooded, and vice versa.

SfM + binary classification

γ_1 (unitless) \ γ_2 (km^2)	**1**	**2.5**	**5**	**10**
2	86	85	85	87
3	85	84	85	86
4	85	84	80	72
5	85	84	77	68

SfM + CAM

γ_1 (unitless) \ γ_2 (km^2)	**1**	**2.5**	**5**	**10**
2	90	89	89	89
3	90	88	87	87
4	90	88	88	87
5	90	88	87	87

TABLE II: Precision for both approaches in percent for various values of γ_1 and γ_2.

Furthermore, we see that our approach provides an additional improvement in precision compared to using the full image footprint from SfM and binary classification alone. Especially in areas at the edges of the flooding extent, our method provides a more precise outline of the official estimates than using the full image footprints. This can be seen most clearly in Figure 6. To characterize the improvement that our approach provides over using the image footprints, we compute in Table II the precision values for various combinations of γ_1 and γ_2. We see that for all chosen combinations of thresholds, our approach has higher precision than the approach using only the image footprints. While the precision of the footprint approach degrades from 86% to 68% as the values of γ_1 and γ_2 increase, the performance of our approach (SfM + CAM) only decreases from 90% to 87%. Thus, our approach is not only more robust to the choices of these parameters, but also manages to perform consistently specifically on data that causes performance degradation in the footprint approach. In this way, our approach allows us filter less data without losing precision, and incorporate more of the highly-oblique source images that would otherwise affect performance in the footprint (SfM + binary) approach.

V. Discussion and Conclusion

In this paper, we focused on the problem of Damage Estimation and Localization (DEL) in the context of the geospatial mapping of damage from aerial images taken with handheld cameras in small, manned, fixed-wing aircraft. While this mode of collecting imagery is cost-efficient and compliant with existing regulations, the raw image data has several attributes which make them difficult to work with. In particular, such images are often highly oblique, sparsely and irregularly sampled, and do not contain IMU information. We proposed an approach to performing DEL from such images by combining structure from motion and class activation mapping to georeference images and detect damage within them. When compared against official flooding estimates from the 2016 Louisiana floods, our approach achieved a precision of 88%, which outperforms the naive approaches of using the image GPS locations or image footprint. While our paper focuses on DEL for flooding images due to the availability of such images, we believe that this approach is generalizable towards other types of imagery. Future work should focus on

implementing this approach on other types of damage, such as debris.

Our approach is not without limitations. First, there are some images, such as very oblique images which include the horizon that cannot be georeferenced using a projective transform. Second, there are a significant number of images that do not have any overlap with other images in terms of pointing at the same scene. We expect that these images could be incorporated into the DEL by georeferencing them against other imagery with a known geotransform, like satellite data. Developing methods to accomplish this capability is an important next step. Nevertheless, our approach performs quite well for images which have a significant overlap with other images, and do not include the horizon. Practically, simple changes in operational procedures—such as requiring images to be taken in bursts, and avoiding the horizon—can limit the number of images which have to be discarded. Most importantly, our approach can quickly generate an estimate of damage distribution from aerial imagery that is already being collected, without the need for any new sensors.

REFERENCES

[1] M. K. V. Aalst, "The impacts of climate change on the risk of natural disasters," *Disasters*, vol. 30, no. 1, pp. 5–18, 2006.

[2] T. Cova, *GIS in Emergency Management*, 12 1999, pp. 845–858.

[3] M. Technologies, "DigitalGlobe's rapid response to the devastating Fort McMurray," *Maxar Blog*, May 2016. [Online]. Available: https://blog.maxar.com/for-a-better-world/2016/digitalglobes-rapid-response-to-the-devastating-fort-mcmurray-wildfire

[4] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston, "xBD: A dataset for assessing building damage from satellite imagery," *arXiv preprint arXiv:1911.09296*, 2019.

[5] 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 preprint arXiv:1812.05581*, 2018.

[6] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. Murphy, "FloodNet: A high resolution aerial imagery dataset for post flood scene understanding," *arXiv preprint arXiv:2012.02951*, 2020.

[7] R. Gupta and M. Shah, "Rescuenet: Joint building segmentation and damage assessment from satellite imagery," *arXiv preprint arXiv:2004.07312*, 2020.

[8] A. J. Cooner, Y. Shao, and J. B. Campbell, "Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 haiti earthquake," *Remote Sensing*, vol. 8, no. 10, p. 868, 2016.

[9] D. Kozanas, "DHS/FEMA/PIA-055 FEMA Response Use of Unmanned Aircraft System (UAS) Derived Imagery," Department of Homeland Security, Tech. Rep., May 2020.

[10] Civil Air Patrol, "CAPabilities Briefing," Apr. 2020.

[11] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2. Ieee, 1999, pp. 1150–1157.

[12] J. Oh, C. K. Toth, and D. A. Grejner-Brzezinska, "Automatic georeferencing of aerial images using stereo high-resolution satellite images," *Photogrammetric Engineering & Remote Sensing*, vol. 77, no. 11, pp. 1157–1168, 2011.

[13] X. Zhuo, T. Koch, F. Kurz, F. Fraundorfer, and P. Reinartz, "Automatic uav image geo-registration by matching uav images to georeferenced image data," *Remote Sensing*, vol. 9, no. 4, p. 376, 2017.

[14] H. Goncalves, L. Corte-Real, and J. A. Goncalves, "Automatic image registration through image segmentation and sift," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 7, p. 2589–2600, Jul 2011.

[15] A. Shetty and G. X. Gao, "UAV pose estimation using cross-view geolocalization with satellite imagery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1827–1833.

[16] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616.

[17] D.-K. Kim and M. R. Walter, "Satellite image-based localization via learned embeddings," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2073–2080.

[18] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5624–5633.

[19] S. Verykokou and C. Ioannidis, "Oblique aerial images: a review focusing on georeferencing procedures," *International Journal of Remote Sensing*, vol. 39, no. 11, pp. 3452–3496, 2018.

[20] L. Cao, C. Wang, and J. Li, "Vehicle detection from highway satellite images via transfer learning," *Information sciences*, vol. 366, pp. 177–187, 2016.

[21] Y. Liang, S. T. Monteiro, and E. S. Saber, "Transfer learning for high resolution aerial image classification," in *2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 2016, pp. 1–8.

[22] S. Ji, S. Wei, and M. Lu, "A scale robust convolutional neural network for automatic building extraction from aerial and satellite imagery," *International journal of remote sensing*, vol. 40, no. 9, pp. 3308–3322, 2019.

[23] B.-C. Gao, "NDWI—a normalized difference water index for remote sensing of vegetation liquid water from space," *Remote sensing of environment*, vol. 58, no. 3, pp. 257–266, 1996.

[24] J. Liu, D. Strohschein, S. Samsi, and A. Weinert, "Large scale organization and inference of an imagery dataset for public safety," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2019, pp. 1–6.

[25] Y. Pi, N. D. Nath, and A. H. Behzadan, "Convolutional neural networks for object detection in aerial imagery for disaster response and recovery," *Advanced Engineering Informatics*, vol. 43, p. 101009, Jan. 2020.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[29] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[30] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[31] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985.

[32] A. M. Andrew, "Multiple view geometry in computer vision," *Kybernetes*, 2001.

[33] M. A. Fonstad, J. T. Dietrich, B. C. Courville, J. L. Jensen, and P. E. Carbonneau, "Topographic structure from motion: a new development in photogrammetric measurement," *Earth surface processes and Landforms*, vol. 38, no. 4, pp. 421–430, 2013.

[34] mapillary, "OpenSfM," Feb 2021. [Online]. Available: https://github.com/mapillary/OpenSfM

[35] C.-P. Wang, K. Wilson, and N. Snavely, "Accurate georegistration of point clouds using geographic data," in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 33–40.

[36] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[37] C. of Baton Rouge and P. of East Baton Rouge, "The great flood of 2016 story map," Aug 2016.