Supplementary Material for
Rapid Computer Vision-aided Disaster Response via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery

The full paper, dataset, and source code can be accessed at:
`https://github.com/FrontierDevelopmentLab/multi3net`.

Background

Earth Observation There is an increasing number of satellites monitoring the Earth's surface, each designed to capture distinct surface properties and to be used for a specific set of applications. Satellites with optical sensors acquire images in the visible and short-wavelength parts of the electromagnetic spectrum that contain information about chemical properties of the captured scene. Satellites with radar sensors, in contrast, use longer wavelengths than those with optical sensors, allowing them to capture physical properties of the Earth's surface (Soergel, 2010). Radar images are widely used in the fields of *Earth observation* and *remote sensing*, since radar image acquisitions are unaffected by cloud coverage or lack of light (Ulaby and Long, 2014). Examples of medium- and very high-resolution optical and medium-resolution radar images are shown in **??**.

Remote sensing-aided disaster response typically uses very high-resolution (VHR) optical and radar imagery. Very high-resolution optical imagery with a ground resolution of less than 1m is visually-interpretable and can be used to manually or automatically extract locations of obstacles or damaged objects. Satellite acquisitions of very high-resolution imagery need to be scheduled and become available only after a disaster event. In contrast, satellites with medium-resolution sensors of 10m–30m ground resolution monitor the Earth's surface with weekly image acquisitions for any location globally. Radar sensors are often used to map floods in sparsely built-up areas since smooth water surfaces reflect electromagnetic waves away from the sensor, whereas buildings reflect them back. As a result, conventional remote sensing flood mapping models perform poorly on images of urban or suburban areas.

We segment building footprints and flooded buildings and compare the results to state-of-the-art benchmarks. To assess model performance, we report the *Intersection over Union* (IoU) metric, which is defined as the number of overlapping pixels labeled as belonging to a certain class in both target image and prediction divided by the union of pixels representing the same class in target image and prediction. We use it to assess the predictions of building footprints and flooded buildings obtained from the model. We report this metric using the acronym 'bIoU'. Represented as a confusion matrix, bIoU $\equiv$ TP$/$(FP $+$ TP $+$ FN), where TP $\equiv$ True Positives, FP $\equiv$ False Positives, TN $\equiv$ True Negatives, and FN $\equiv$ False Negatives. Conversely, the IoU for the background class, in our case denoting 'not a flooded building', is given by TN$/$(TN$+$FP$+$FN). Additionally, we report the mean of (flooded) building and background IoU values, abbreviated as 'mIoU'. We also compute the pixel accuracy A, the percentage of correctly classified pixels, as $A \equiv$ (TP $+$ TN)$/$(TP $+$ FP $+$ TN $+$ FN).

Preprocessing In Section *Earth Observation*, we described the properties of short-wavelength optical and long-wavelength radar imagery. For Sentinel-2 optical data, we use *top-of-atmosphere* reflectances without applying further atmospheric corrections to minimize the amount of optical preprocessing need for our approach. For radar data, however, preprocessing of the raw data is necessary to obtain numerical values that can be used as network inputs. A single radar 'pixel' is expressed as a complex number z and composed of a real in-phase, $\mathrm{Re}(z)$, and an imaginary quadrature component of the reflected electromagnetic signal, $\mathrm{Im}(z)$. We use *single look complex* data to derive the radar intensity and coherence features. The intensity, defined as $I \equiv z^2 = \mathrm{Re}(z)^2 + \mathrm{Im}(z)^2$, contains information about the magnitude of the surface-reflected energy. The radar images are preprocessed according to Ulaby and Long (2014): (1) We perform *radiometric calibration* to compensate for the effects of the sensor's relative orientation to the illuminated scene and the distance between them. (2) We reduce the noise induced by electromagnetic interference, known as *speckle*, by applying a spatial averaging kernel, known as *multi-looking* in radar nomenclature. (3) We normalize the effects of the terrain elevation using a digital elevation model, a process known as *terrain correction*, where a coordinate is assigned to each pixel through *georeferencing*. (4) We average the intensity of all radar images over an extended temporal period, known as *temporal multi-looking*, to further reduce the effect of speckle on the image. (5) We calculate the *interferometric* *coherence* between images, $\mathbf{z}_t$, at times $t = 1, 2$,

$$\gamma = \frac{\mathbb{E}[\mathbf{z}_1 \mathbf{z}_2^*]}{\sqrt{\mathbb{E}[|\mathbf{z}_1|^2]\,\mathbb{E}[|\mathbf{z}_2|^2]}}, \tag{1}$$

where $\mathbf{z}_t^*$ is the complex conjugate of $\mathbf{z}_t$ and expectations are computed using a local *boxcar-function*. The coherence is a local similarity metric (Zebker and Villasenor, 1992) able to measure changes between pairs of radar images.

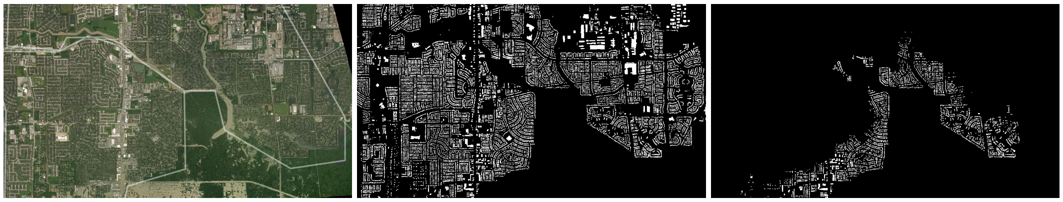
Data Addendum

Area of Interest We chose two neighboring, non-overlapping districts of Houston, Texas as training and test areas. Houston was flooded in the wake of Hurricane Harvey, a category 4 hurricane that formed over the Atlantic on August 17, 2017, and made landfall along the coast of the state of Texas on August 25, 2017. The hurricane dissipated on September 2, 2017. In the early hours of August 28, extreme rainfalls caused an 'uncontrolled overflow' of Houston's Addicks Reservoir and flooded the neighborhoods of 'Bear Creek Village', 'Charlestown Colony', 'Concord Bridge', and 'Twin Lakes'.

Ground Truth We chose this area of interest because accurate building footprints for the affected areas are publicly available through OpenStreetMap. Flooded buildings have been manually labeled through crowdsourcing as part of the DigitalGlobe Open Data Program (DigitalGlobe, 2018). When preprocessing the data, we combine the building footprints obtained from OpenStreetMap with point-wise annotations from DigitalGlobe to produce the ground truth map shown in Figure 3c. The geometry collections of buildings (shown in Figure 3b) and flooded buildings (shown in Figure 3c) are then rasterized to create 2m or 10m pixel grids, depending on the satellite imagery available. Figure 3a shows a very high-resolution image of the area of interest overlaid with boundaries for the East and West partitions used for training and testing, respectively.

Data Preprocessing For radar data, we construct a three-band image consisting of the intensity, multitemporal filtered intensity, and interferometric coherence. We compute the intensity of two radar images obtained from Sentinel-1 sensors in stripmap mode with a ground resolution of 5m for August 23 and September 4, 2017. Additionally, we calculate the interferometric coherence for an image pair without flood-related changes acquired on June 6 and August 23, 2017, as well as for an image pair with flood-induced scene changes acquired on August 23 and September 4, 2017, using Equation (1). As the third band of the radar input, we compute the multitemporal intensity by averaging all Sentinel-1 radar images from 2016 and 2017. This way, speckle noise affecting the radar image can be reduced. We merge the intensity, multitemporal filtered intensity, and coherence images obtained both pre- and post-disaster into separate three-band images. The multi-band images are then fed into the respective network streams.

Sentinel-2 measures the surface reflectances in 13 spectral bands with 10m, 20m, and 60m ground resolutions. We apply bilinear interpolations to the 20m band images to obtain an image representation with 10m ground resolution. Finally, we extract rectangular tiles of size 960m×960m from the set of satellite images to use as input samples for the network. This tile extraction process is repeated every 100m in the four cardinal directions to produce overlapping tiles for training and testing, respectively. The large tile overlap can be interpreted as an offline data augmentation step.



(a) VHR imagery with dataset boundaries (b) OpenStreetMap building footprints (c) Annotated flooded buildings

Figure 3: Images illustrating the size and extent of training and testing datasets (Figure 3a), available rasterized ground truth annotations as OpenStreetMap building footprints (Figure 3b), and expert-annotated labels of flooded buildings (Figure 3c).

Method Addendum

Network Training & Evaluation We initialize the encoder with the weights of a ResNet34 model (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009). When there are more than three input channels in the first convolution (due to the 10 spectral bands of the Sentinel-2 satellite images), we initialize additional channels with the average over the first convolutional filters of the RGB channels. Multi^{3}Net was trained using the *Adam* optimization algorithm (Kingma and Ba, 2014) with a learning rate of 10^{-2}. The network parameters are optimized using a cross entropy loss

$$H(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_i \mathbf{y}_i \log(\hat{\mathbf{y}}_i), \qquad (2)$$

between ground truth $\mathbf{y}$ and predictions $\hat{\mathbf{y}}$. We anneal the learning rate according to the poly policy (power $= 0.9$) introduced in Chen et al. (2018) and stop training once the loss converges. For each batch, we randomly sample 8 tiles of size 960m×960m (corresponding to 96px×96px optical and 192px×192px radar images) from the dataset. We augment the training dataset by randomly rotating and flipping the image vertically and horizontally in order to create additional samples. To segment flooded buildings with Multi^{3}Net, we first pre-train the network on building footprints. We then use the resulting weights for network initialization and train Multi^{3}Net on the footprints of flooded buildings.

To train our models, we divided the area of interest into two partitions (i.e. non-overlapping subsets) covering two different neighborhoods, as shown in Figure 3a. We randomly divided the East partition into a training and a validation set at a 4:1 split. The model hyperparameters were optimized on the validation set. All model evaluations presented in this work were performed on the spatially separate test dataset.

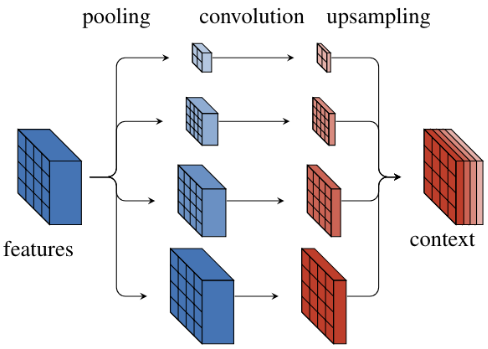


Figure 4: The context aggregation module used in our model which extracts and combines image features at different image resolutions, similarly to (Zhao et al., 2017).

Results Addendum

Building Footprint Segmentation—Single Sensors We tested our model on the auxiliary task of building footprint segmentation. The wide applicability of this task has led to the creation of several benchmark datasets, such as the DeepGlobe (Demir et al., 2018), SpaceNet (Van Etten, Lindenbaum, and Bacastow, 2018), and INRIA aerial labels datasets (Maggiori et al., 2017a), all containing very high-resolution RGB satellite imagery. Table 1a shows the performance of our model on the Austin partition of the INRIA aerial labels dataset. Maggiori et al. (2017b) use a fully convolutional network (Long, Shelhamer, and Darrell, 2015) to extract features that were concatenated and classified by a second multilayer perceptron stream. Ohleyer (2018) employ a Mask-RCNN (He et al., 2017) instance segmentation network for building footprint segmentation.

Using only very high-resolution imagery, Multi^{3}Net performed better than current state-of-the-art models, reaching a bIoU 7.8% higher than Ohleyer (2018). Comparing the performance of our model for different single-sensor inputs, we found that predictions based on very high-resolution images achieved the highest building IoU score, followed by predictions based on Sentinel-2 medium-resolution optical images, suggesting that optical bands contain more relevant information for this prediction task than radar images.

Building Footprint Segmentation—Image Fusion Fusing multiresolution and multisensor satellite imagery further improved the predictive performance. The results presented in Table 2 show that the highest accuracy was achieved when all data sources were fused.

Fusing Sentinel-1 and Sentinel-2 data produced highly accurate predictions (76.1% mIoU), only surpassed by predictions obtained by fusing Sentinel-1, Sentinel-2, and very high-resolution imagery (79.9%).

Data	mIoU	bIoU	Accuracy
S-1	69.3%	63.7%	82.6%
S-2	73.1%	66.7%	85.4%
VHR	78.9%	74.3%	88.8%
S-1 + S-2	76.1%	70.5%	87.3%
S-1 + S-2 + VHR	**79.9%**	**75.2%**	**89.5%**

Table 2: Results for the segmentation of building footprints using different input data in Multi^{3}Net.

Segmentation of Flooded Buildings—Comparison of Resolutions and Fusion Methods

We tested the performance of Multi^{3}Net with only single sensor medium-resolution inputs to its performance when fusing optical and radar medium-resolution images and found that fusing medium-resolution images from different sensors improved the mIoU score significantly, increasing it from 50.2% and 52.6%, respectively, to 59.7% (see Table 2).

Data	mIoU	bIoU	Accuracy
S-1	50.2%	17.1%	80.6%
S-2	52.6%	12.7%	81.2%
VHR	74.2%	56.0%	93.1%
S-1 + S-2	59.7%	34.1%	86.4%
S-1 + S-2 + VHR	**75.3%**	**57.5%**	**93.7%**

Table 3: Results for the segmentation of flooded buildings using different input data in Multi^{3}Net.

We also compared the performance of Multi^{3}Net to the performance of a baseline U-Net data fusion architecture, which has been successful at recent satellite image segmentation competitions, and found that our model outperformed the U-Net baseline on building footprint segmentation for all input types (see Table 4). We also compared the performance between Multi^{3}Net and a baseline U-Net fusion architecture on the segmentation of flooded buildings and found that our method performed significantly better, reaching a building IoU (bIoU) score of 75.3% compared to a bIoU score of 44.2% for the U-Net baseline.

Model	Data	mIoU	bIoU	Accuracy
Multi^{3}Net	Sentinel-1 + Sentinel-2	76.1%	70.5%	87.3%
	VHR	78.9%	74.3%	88.8%
	Sentinel-1 + Sentinel-2 + VHR	**79.9%**	**75.2%**	**89.5%**
U-Net	Sentinel-1 + Sentinel-2	-	60%	88%
	VHR	-	38%	77%
	Sentinel-1 + Sentinel-2 + VHR	-	**73%**	**89%**

Table 4: Building footprint segmentation results for Multi^{3}Net and a U-Net baseline.