Rapid Computer Vision-Aided Disaster Response via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery

Tim G. J. Rudner* University of Oxford
Marc Rußwurm TU Munich
Jakub Fil University of Kent
Ramona Pelich LIST Luxembourg
Benjamin Bischke DFKI & TU Kaiserslautern

Veronika Kopačková Czech Geological Survey
Piotr Biliński University of Oxford & University of Warsaw

Abstract

Natural disasters can cause loss of life and substantial property damage. Moreover, the economic ramifications of disaster damage disproportionately impact the most vulnerable members of society. In this paper, we propose *Multi*^{3}*Net*, a novel approach for rapid and accurate disaster damage segmentation by fusing multiresolution, multisensor, and multitemporal satellite imagery in a convolutional neural network. In our method, segmentation maps can be produced as soon as at least a single satellite image acquisition has been successful and subsequently be improved upon once additional imagery becomes available. This way, we are able to reduce the amount of time needed to generate satellite imagery-based disaster damage maps, enabling first responders and local authorities to make swift and well-informed decisions when responding to disaster events. We demonstrate the performance and usefulness of our approach for earthquake and flood events. To encourage future research into image fusion for disaster relief, we release the first open-source dataset of fully preprocessed and labeled multiresolution, multispectral, and multitemporal satellite images of disaster sites along with our source code at `https://github.com/FrontierDevelopmentLab/multi3net`.

Introduction

In 2017, Houston, Texas, the fourth largest city in the United States, was hit by tropical storm Harvey, the worst storm to pass through the city in over 50 years. Harvey flooded large parts of Houston, inundating over 154,170 homes and leading to more than 80 deaths. According to the National Hurricane Center, the storm caused over 125 billion USD in damage, making it the second costliest storm ever recorded in the United States. Natural disasters can cause loss of life and substantial property damage. Moreover, the economic ramifications of disaster damage disproportionately impact the most vulnerable members of society.

When a region is hit by a natural disaster, authorized representatives of national civil protection, rescue, and security organizations can activate the International Charter 'Space and Major Disasters'. Once the Charter has been activated, commercial Earth observation companies and national space organizations task their satellites to acquire imagery of the affected region. As soon as images have been obtained, satellite imagery specialists visually or semi-automatically interpret them to create flood maps to be delivered to disaster relief organizations. However, Due to the semi-automated nature of the map generation process, delivery of flood maps to first responders can take several hours after the imagery was provided.

First workshop on AI for Social Good (NeurIPS 2018), Montréal, Canada.

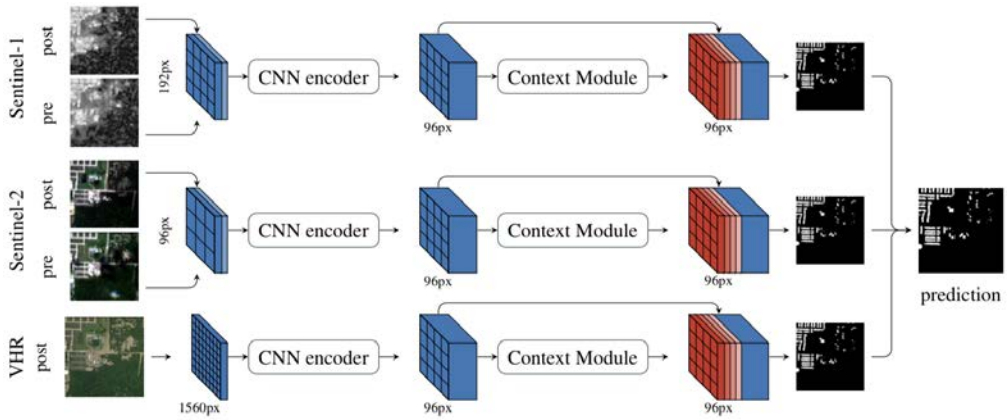


Figure 1: Overview of Multi^{3}Net's multi-stream architecture. Each satellite image is processed by a separate stream that extracts feature maps using a CNN-encoder and then augments them with contextual features. Features are mapped to the same spatial resolution, and the final prediction is obtained by fusing the predictions of individual streams using additional convolutions.

In this paper, we propose *Multi^{3}Net*, a novel approach for rapid and accurate disaster damage segmentation by fusing multiresolution, multisensor, and multitemporal satellite imagery in a convolutional neural network. The network consists of multiple deep encoder-decoder streams, each of which produces an output map based on data from a single sensor. If data from multiple sensors is available, the streams are combined into a joint prediction map.

Our method aims to reduce the amount of time needed to generate satellite imagery-based flood maps by fusing images from multiple satellite sensors. Segmentation maps can be produced as soon as at least a single satellite image acquisition has been successful and subsequently be improved upon once additional imagery becomes available. This way, the amount of time needed to generate satellite imagery-based flood maps can be reduced significantly, helping first responders and local authorities make swift and well-informed decisions when responding to flood events. Additionally, by incorporating multitemporal satellite imagery, our method allows for a speedy and accurate post-disaster damage assessment, helping governments better coordinate medium- and long-term financial assistance programs for affected areas.

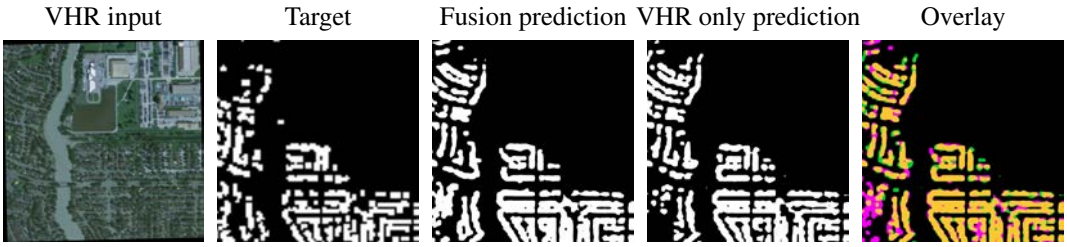
Related Work

Mapping disaster damage using high-resolution imagery has long been an area of research in the field of remote sensing (Barnes, Fritz, and Yoo, 2007a; Yamazaki, 2001), where methods are typically tailored to specific disaster types, such as floods (Scarsi et al., 2014; Goldberg et al., 2018), hurricanes (Cao and Choe, 2018; Ramlal, Davis, and De Bellott, 2018), or earthquakes (Brunner, Lemoine, and Bruzzone, 2010; Cooner, Shao, and Campbell, 2016). Damage caused by hurricanes and earthquakes is often idetified using high-resolution optical or radar imagery (Barnes, Fritz, and Yoo, 2007b), whereas floods (in non-urban areas) are usually identified using low-spatial resolution long-wavelength radar satellite images (Scarsi et al., 2014). Identifying flooding in urban areas, however, is more challenging for conventional remote sensing approaches (Soergel, 2010).

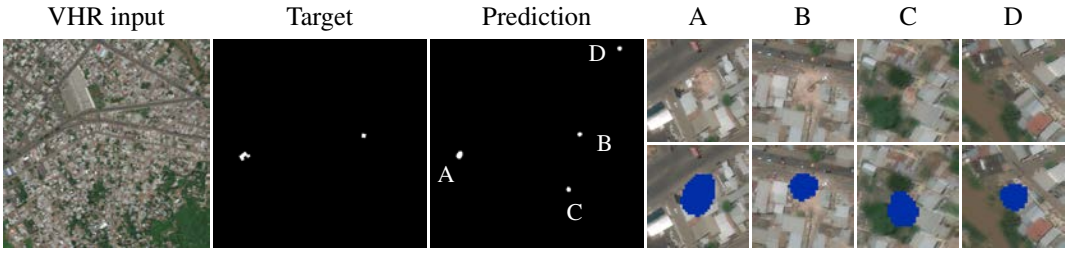
Recent advances in computer vision and the rapid increase of commercially and publicly available medium- and high-resolution satellite imagery have given rise to a new area of research at the interface of machine learning and remote sensing, as summarized by Zhu et al. (2017) and Zhang, Zhang, and Du (2016). Single-stream convolutional neural network approaches have demonstrated the benefits of deep feature learning in end-to-end architectures (Sun et al., 2017; Narazaki et al., 2018). For the segmentation of building footprints from satllite images, U-Net-based approaches that replace the original VGG architecture (Simonyan and Zisserman, 2014) with, for example, ResNet encoders (He et al., 2016) achieved the best results in the 2018 DeepGlobe challenge (Hamaguchi and Hikosaka, 2018). Recently developed computer vision models, such as DeepLab-v3 (Chen et al., 2017), PSPNet (Zhao et al., 2017), or DDSC (Bilinski and Prisacariu, 2018), however, use improved encoder architectures with a higher receptive field and additional context modules.

Multi^{3}Net

Multi^{3}Net uses an encoder-decoder architecture. In particular, we use a modified version of ResNet (He et al., 2016) with dilated convolutions as feature extractors (Yu, Koltun, and Funkhouser, 2017)



(a) Comparison of predictions for the segmentation of flooded buildings for fusion-based and VHR-only models. In the overlay image, predictions added by the fusion are marked in magenta, predictions that were removed by the fusion are marked in green, and predictions present in both are marked in yellow.



(b) Segmentation of collapsed buildings in the Ecuadorian town of Portoviejo after an earthquake in 2016.

Figure 2: Qualitative segmentation results for flooded and collapsed buildings, resctively.

that allows us to effectively downsample the multi-resolution input streams to a common spatial dimension. Motivated by the recent success of multi-scale features (Zhao et al., 2017; Chen et al., 2017), we enrich the feature maps with an additional context aggregation module as described in (Zhao et al., 2017). The decoder component of the network uses three blocks of bilinear upsampling functions with a factor of $\times 2$, followed by a 3×3 convolution, and a PReLU activation function to learn a mapping from latent space to label space. This way, *Multi*^{3}*Net* is able to fuse images obtained at multiple points in time from multiple sensors with different resolutions and capture different properties of the Earth's surface across time. The network is trained end-to-end using backpropagation.

Multisensor Fusion We used a late fusion approach where each image type is fed into a dedicated information processing stream as shown in the segmentation network architecture depicted in Figure 1. We first extract features separately from each satellite image. Next, we combine the class predictions from each individual stream by first concatenating them and then applying additional convolutions. We compared the performance of several network architectures, fusing the feature maps in the encoder (as was done in FuseNet (Hazirbas et al., 2016)) and using different late-fusion approaches, such as sum fusion or element-wise multiplication, and found that a late-fusion approach, in which the output of each stream is fused using additional convolutional layers, achieved the best performance. In this setup, the segmentation maps from the different streams are fused by concatenating the segmentation map tensors and applying two additional layers of 3×3 convolutions with PReLU activations and a 1×1 convolution.

Multiresolution Fusion In order to best incorporate the satellite images' different spatial resolutions, we follow two different approaches. When only medium-resolution images are available, we transform the feature maps into a common resolution of 96px $\times$ 96px at a 10m ground resolution by removing one upsampling layer in the Sentinel-2 encoder network. Whenever very high-resolution (VHR) optical imagery is available as well, we also remove the upsampling layer in the very high-resolution subnetwork to match the feature maps of the two Sentinel imagery streams.

Multitemporal Fusion To detect changes in an image scene over time, we use pre- and post-disaster images. We achieved the best segmentation results by concatenating pre- and post-disaster images into a single input tensor and processing them with the network described in Figure 1.

Results and Discussion

To train our model, we use medium-resolution satellite imagery with a ground resolution of 5m–10m, acquired before and after disaster events, along with very high-resolution post-event images with

Table 1: Quantitative results from two experiments reporting building intersection over union (bIoU), mean IoU (mIoU), and pixel accuracy. Table 1a compares our method to state-of-the-art approaches for segmentation of building footprints. Table 1b compares different fusion inputs for segmentation of flooded buildings using *Multi*^{3}*Net*.

Model	**bIoU**	**Accuracy**
Maggiori et al. (2017b)	61.2%	94.2%
Ohleyer (2018)	65.6%	94.1%
This work	**73.4%**	**95.7%**

(a) Segmentation of building footprints using VHR imagery of Austin in the INRIA Aerial Labels Dataset.

Data	**mIoU**	**bIoU**	**Accuracy**
S-1 + S-2	59.7%	34.1%	86.4%
VHR	74.2%	56.0%	93.1%
S-1 + S-2 + VHR	**75.3%**	**57.5%**	**93.7%**

(b) Segmentation of flooded buildings in Houston, TX, following Hurricane Harvey, 2017.

a ground resolution of 0.5m. Medium-resolution satellite imagery is publicly available for any location globally and acquired weekly by the European Space Agency's Sentinel-1 and Sentinel-2 satellite constellations. To obtain finer image details, such as building delineations, we use very high-resolution post-event images obtained through the DigitalGlobe Open Data Program. For radar data, we construct a three-band image consisting of the intensity, multitemporal filtered intensity, and interferometric coherence. Details about the data acquisition process and remote sensing terminology can be found in the supplementary material.

Building footprint segmentation We demonstrated the competitive performance of our model for the segmentation of building footprints. We assessed our model vis-à-vis other approaches using pixel accuracy and the intersection over union (IoU) metric. Our method outperformed state-of-the-art approaches for building footprint segmentation, reaching a building IoU of 73.4% (see Table 1a) on the Austin partition of the INRIA aerial labels dataset (Maggiori et al., 2017a).

Segmentation of disaster damage To segment footprints of flooded buildings, we used pre- and post-event images obtained by Sentinel-1 and Sentinel-2 along with post-event VHR imagery. Table 1b shows that fusing images from all sensors across time yielded the best results (75.3% mIoU). Fusing only medium-resolution Sentinel-1 and Sentinel-2 images without high-resolution imagery yielded a good segmentation accuracy (59.7% mIoU) as well. Figure 2a shows predictions for the segmentation of flooded buildings obtained from the very high-resolution-only and full-fusion models. The overlay image shows the differences between the two predictions. Fusing images obtained at multiple points in time from multiple sensors with different resolutions eliminates the majority of false positives and helps delineate the shape of detected structures more accurately.

We also used our method to segment collapsed buildings in the Ecuadorian town of Portoviejo following an earthquake in 2016. This task is much more challenging than segmenting flooded buildings due to the relative sparsity of collapsed buildings in our sample images. To achieve high predictive accuracy, we first pre-trained the network to perform standard building footprint segmentation before training the model on the footprints of collapsed buildings. This way, the model first learns to identify the set of 'buildings', before learning to segment the subset of collapsed buildings. We also modified the loss function to assign penalties ($\times 100$) for incorrectly identifying pixels that are labeled as belonging to the footprint of a collapsed building to discourage the network from over-predicting non-collapsed buildings (which make up over 90% of the pixels). Figure 2b shows that our model was able to correctly identify collapsed buildings (points A and B) as well as two buildings that were labeled as severely damaged (points C and D).

Conclusion

In disaster response, fast information extraction is crucial for first responders to coordinate disaster relief efforts, and satellite imagery can be a valuable asset for rapid mapping of affected areas. In this work, we introduced a novel end-to-end trainable convolutional neural network architecture for image segmentation via fusion of multiresolution, multisensor, and multitemporal satellite images. Our network outperformed state-of-the-art approaches on building footprint segmentation and achieved high accuracy in the segmentation of flooded buildings. We demonstrated that publicly and globally available medium-resolution imagery alone can be used for efficient segmentation of flooded buildings, making our method massively scalable. The source code as well as a dataset containing fully preprocessed and labeled multiresolution, multispectral, and multitemporal satellite imagery of disaster sites will be made publicly available.

References

Barnes, C. F.; Fritz, H.; and Yoo, J. 2007a. Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* 45(6):1631–1640.

Barnes, C. F.; Fritz, H. M.; and Yoo, J. 2007b. Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* 45:1631–1640.

Bilinski, P., and Prisacariu, V. 2018. Dense decoder shortcut connections for single-pass semantic segmentation. In *CVPR*.

Brunner, D.; Lemoine, G.; and Bruzzone, L. 2010. Earthquake damage assessment of buildings using vhr optical and sar imagery. *IEEE Transactions on Geoscience and Remote Sensing* 48:2403–2420.

Cao, Q. D., and Choe, Y. 2018. Deep learning based damage detection on post-hurricane satellite imagery. *CoRR* abs/1807.01688.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Cooner, A. J.; Shao, Y.; and Campbell, J. B. 2016. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 haiti earthquake. *Remote Sensing* 8:868.

Goldberg, M.; Li, S.; Goodman, S.; Lindsey, D.; Sjoberg, B.; and Sun, D. 2018. Contributions of operational satellites in monitoring the catastrophic floodwaters due to hurricane harvey. *Remote Sensing* 10(8):1256.

Hamaguchi, R., and Hikosaka, S. 2018. Building detection from satellite imagery using ensemble of size-specific detectors. In *CVPR Workshop*.

Hazirbas, C.; Ma, L.; Domokos, C.; and Cremers, D. 2016. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *ACCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Maggiori, E.; Tarabalka, Y.; Charpiat, G.; and Alliez, P. 2017a. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IGARSS*. IEEE.

Maggiori, E.; Tarabalka, Y.; Charpiat, G.; and Alliez, P. 2017b. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* 55(2):645–657.

Narazaki, Y.; Hoskere, V.; Hoang, T. A.; and Spencer Jr, B. F. 2018. Automated vision-based bridge component extraction using multiscale convolutional neural networks. *arXiv preprint arXiv:1805.06042*.

Ohleyer, S. 2018. Building segmentation on satellite images. https://project.inria.fr/aerialimagelabeling/files/2018/01/fp_ohleyer_compressed.pdf. Accessed: 2018-08-26.

Ramlal, B.; Davis, D.; and De Bellott, K. 2018. A rapid post-hurricane building damage assessment methodology using satellite imagery. *West Indian Journal of Engineering* 41(1).

Scarsi, A.; Emery, W. J.; Serpico, S. B.; and Pacifici, F. 2014. An automated flood detection framework for very high spatial resolution imagery. *IEEE Geoscience and Remote Sensing Symposium* 4954–4957.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soergel, U. 2010. *Radar Remote Sensing of Urban Areas*, volume 15. Springer.

Sun, G.; Hao, Y.; Rong, J.; Shi, S.; and Ren, J. 2017. Combined deep learning and multiscale segmentation for rapid high resolution damage mapping. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 1101–1105. IEEE.

Yamazaki, F. 2001. Applications of remote sensing and gis for damage assessment. *Structural Safety and Reliability* 1–12.

Yu, F.; Koltun, V.; and Funkhouser, T. A. 2017. Dilated residual networks. In *CVPR*.

Zhang, L.; Zhang, L.; and Du, B. 2016. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine* 4:22–40.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *CVPR*.

Zhu, X. X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; and Fraundorfer, F. 2017. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* 5(4):8–36.

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).

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, $\text{bIoU} \equiv \text{TP}/(\text{FP} + \text{TP} + \text{FN})$, where $\text{TP} \equiv$ True Positives, $\text{FP} \equiv$ False Positives, $\text{TN} \equiv$ True Negatives, and $\text{FN} \equiv$ False Negatives. Conversely, the IoU for the background class, in our case denoting 'not a flooded building', is given by $\text{TN}/(\text{TN} + \text{FP} + \text{FN})$. 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 $A \equiv (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN})$.

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, $\text{Re}(z)$, and an imaginary quadrature component of the reflected electromagnetic signal, $\text{Im}(z)$. We use *single look complex* data to derive the radar intensity and coherence features. The intensity, defined as $I \equiv z^2 = \text{Re}(z)^2 + \text{Im}(z)^2$, 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

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

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.

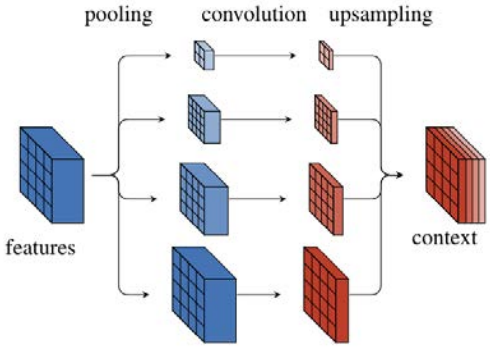


Figure 4: The context aggregation module used in our model which extracts and combines image features at different image resolutions, similarly to (Zhao et al., 2017).

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.