Multi^{3}Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery

Jakub Fil* University of Kent
Marc Rußwurm TU Munich
Tim G. J. Rudner University of Oxford
Ramona Pelich LIST Luxembourg
Benjamin Bischke DFKI & TU Kaiserslautern
Veronika Kopačková Czech Geological Survey
Piotr Biliński University of Oxford & University of Warsaw

Abstract

We present a novel approach to performing rapid segmentation of flooded buildings by fusing multiresolution, multisensor, and multitemporal satellite imagery in a convolutional neural network. Our method significantly expedites the generation of satellite imagery-based flood maps, which are crucial for first responders and local authorities in the early stages of flood events. By incorporating multitemporal satellite imagery, our approach allows for a rapid and accurate post-disaster damage assessment, helping governments to better coordinate medium- and long-term financial assistance programs for affected areas. Our model consists of multiple streams of encoder-decoder architectures that extract temporal information from medium-resolution images and spatial information from high-resolution images before fusing the resulting representations into a single medium-resolution segmentation map of flooded buildings. We demonstrate that our model produces highly accurate segmentation of flooded buildings using only freely available medium-resolution imagery and can be improved through very high-resolution (VHR) data.

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. Floods can cause loss of life and substantial property damage, resulting in major economic ramifications for affected areas. Moreover, these effects disproportionately impact the most vulnerable members of society.

When a region is hit by heavy rainfall or a hurricane, an authorized representative of a national civil protection, rescue, or security organization 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. Once images have been obtained, satellite imagery specialists visually or semi-automatically interpret them to create flood maps to be delivered to disaster relief organizations. Due to the semi-automated nature of the map generation process, delivery of flood maps can take several hours after the imagery was provided. Further, the acquisition of images can be delayed by the satellite constellation due to weekly ground repeat cycles and local cloud cover.

In this paper, we propose *Multi^{3}Net*, a novel approach for rapid and accurate flood damage segmentation by fusing multiresolution and multisensor satellite imagery in a convolutional neural network.

*corresponding author: `jf330@kent.ac.uk`

Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

The network consists of multiple deep encoder-decoder streams, in which each individual stream is produces an output map based data from a single sensor. If data from multiple sensors is available, the streams are combined into a joint prediction map. We use this network for building footprint detection and segmentation of flooded buildings. Our method aims to reduce the amount of time needed to generate satellite imagery-based flood maps by fusing multiple satellite sensors. A segmentation map can be produced with data from a single satellite and subsequently improved when additional imagery becomes available. This way, it is possible to reduce the amount of time needed to generate satellite imagery-based flood maps, enabling first responders and local authorities to make swift and well-informed decisions when responding to flood events. Additionally, it allows for a speedy and accurate post-disaster damage assessment using multitemporal satellite imagery, helping governments to better coordinate medium- and long-term financial assistance programs for affected areas.

Related Work

Advances in computer vision and the rapid increase of high- and medium-resolution satellite imagery have given rise to a new area of research at the interface of machine learning and remote sensing, as summarized by (Zhang, Zhang, and Du, 2016; Zhu et al., 2017).

One popular task in this domain is the segmentation of buildings from remote sensing imagery which has led to competitions such as the DeepGlobe (Demir et al., 2018) and SpaceNet challenges (Van Etten, Lindenbaum, and Bacastow, 2018). U-Net-based approaches that replace the original VGG architecture (Simonyan and Zisserman, 2014) with, for example, ResNet encoders (He et al., 2016) have achieved the best results at the 2018 DeepGlobe challenge (Hamaguchi and Hikosaka, 2018). Recently developed computer vision models, such as Deeplab-V3 (Chen et al., 2017), PSP-net (Zhao et al., 2017), or DDSC (Bilinski and Prisacariu, 2018) augment these using an improved encoder architecture with a higher receptive field and additional context modules.

Segmentation of flooded buildings is similar in nature to building segmentation. However, it is more challenging than ordinary segmentation of building footprints, as the image scene includes additional, confounding features, i.e. damages caused by flooding. Adding a temporal dimension by using pre- and post-disaster imagery can help solve this challenge. Cooner, Shao, and Campbell (2016), for instance, insert a pair of pre- and post-disaster images into a feedforward neural network and into random forests, allowing them to identify damaged buildings after the 2010 Haiti earthquake.

Multi^{3}Net

The segmentation network used in this work is based on an encoder-decoder architecture. We use a modified version of ResNet (He et al., 2016) with dilated convolutions proposed by Yu, Koltun, and Funkhouser (2017) as a feature extractor that lets us 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 depicted in Figure 2. This addition to the network allows us to incorporate contextual image information into the encoded image representation. 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 sourced from different sensors with different resolutions that capture different properties of the Earth's surface across time. The network is trained end-to-end using back-propagation. Next, we will address each fusion type separately.

Multisensor Fusion Images obtained from different sensors are fed into dedicated information processing streams as described in the segmentation network architecture shown in Figure 1. We extract features separately from each satellite image and then combine the class predictions from each individual stream by first concatenating them and then applying additional convolutions. We conduct several experiments, fusing the feature maps in the encoder (similarly to FuseNet (Hazirbas et al., 2016)) and using different late fusion approaches such as sum fusion or element-wise multiplication. In our experiments, we found that a late-fusion approach, in which the output of each stream is fused using additional convolutional layers, achieved the best results. This finding is consistent with related work in computer vision on the fusion of RGB optical images and depth sensors (Couprie et al., 2013). In our setup, each stream produces a separate segmentation output map, each of which is fused by concatenating the tensors and applying two additional layers of 3×3 convolutions with PReLU

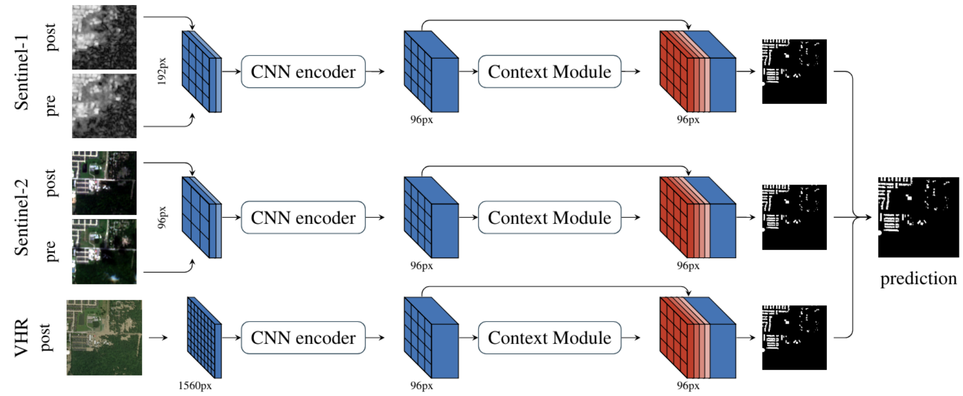

Figure 1: *Multi*^{3}*Net* architecture. Each satellite image is processed by a separate stream that extracts feature maps using a CNN-encoder and augments them with contextual features. Features are mapped to the same spatial resolution and model predictions are obtained by fusing predictions from each stream using additional convolutions.

activations and a 1×1 convolution. This way, the dimensions along the channels can be reduced until they are equal to the number of class labels.

Multiresolution Fusion In order to best incorporate the satellite images' different spatial resolutions, we consider two different approaches. If only Sentinel-1 and Sentinel-2 imagery is available, we transform the feature maps to a common resolution of 96px $\times$ 96px at 10m ground resolution, removing one upsampling layer in the Sentinel-2 subnetwork. If VHR optical imagery is available as well, we also remove the upsampling layer in the VHR subnetwork to match the feature maps of the two Sentinel imagery streams. In order to quantify changes in a satellite scene over time, we use pre- and post-disaster satellite imagery. We achieved the best results by concatenating both images to a single input tensor and processing them with the network described in Figure 1.

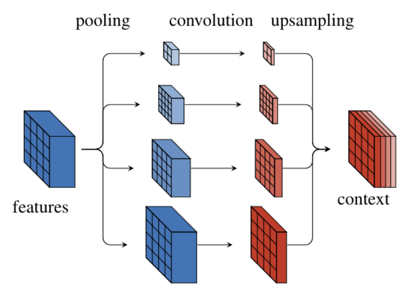

Figure 2: The context aggregation module extracts and combines image features at different image resolutions, similar to (Zhao et al., 2017).

Data

To avoid spatial autocorrelation, we chose two neighboring, non-overlapping districts of Houston, Texas, as training and test areas. We use medium-resolution satellite imagery with a pixel size of 5m–10m acquired before and after the disaster event along with VHR post-hurricane images with a ground pixel size of 0.5m. Medium-resolution satellite imagery is freely available for any location globally and acquired weekly through the European Space Agency's Copernicus Program. To obtain finer image details, such as exact building delineations, we use VHR post-event images obtained through the DigitalGlobe Open Data Program.

For radar data, we construct a three-band image from the intensity, multitemporal filtered intensity, and interferometric coherence of the radar image. We merge the intensity, multitemporal filtered intensity, and coherence images obtained pre- and post-disaster into single, three-band images, respectively. Details on the area of interest, creation of the input data, example images, and Earth observation terminology can be found in the supplementary material.

Results and Discussion

To perform segmentation of flooded buildings, we use multi-temporal data from Sentinel-1 and Sentinel-2 along with post-event VHR imagery in *Multi*^{3}*Net*. We will assess our model vis-à-vis other approaches using pixel accuracy and the intersection over union (IoU) metric.

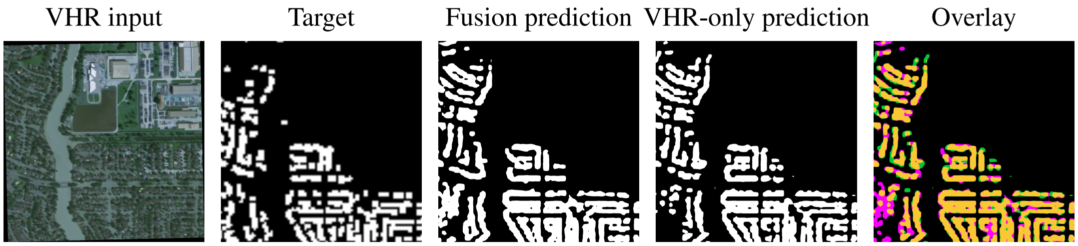

Figure 3: Comparison of results for the segmentation of flooded buildings by fusion-based and VHR-only models. In the overlay image, predictions added by the fusion are marked in magenta, predictions that were removed are green, and predictions that overlapped in both are yellow.

Table 1 shows that fusing images from all resolutions and sensors across time yielded the best performance (75.3% mIoU), and that fusing only globally available medium-resolution Sentinel-1 and Sentinel-2 images also performed well, reaching a mean IoU score of 59.7%. Figure 3 presents flood damage segmentation results for the VHR-only and full-fusion models. The overlay image shows the differences between the two predictions. Fusing images from multiple resolutions and sensors across time eliminates false positives, and delineates the shape of detected structures more accurately. The buildings in the bottom left corner, highlighted in magenta, were only detected using multisensor input.

Additionally, we compared our model to state-of-the-art building footprint segmentation models on the Austin partition of the INRIA aerial labels dataset (Maggiori et al., 2017a) and found that our model performed best (73.4% bIoU) at this task (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 1: Mean IoU (mIoU), building IoU (bIoU), and pixel accuracy for flooded building segmentation using Multi^{3}Net.

Model	bIoU	Accuracy
Maggiori et al. (2017b)	61.2%	94.2%
Ohleyer (2018)	65.6%	94.1%
Multi^{3}Net	**73.4%**	**95.7%**

Table 2: Building IoU (bIoU) and pixel accuracy for building footprint segmentation using VHR imagery of Austin in the INRIA aerial labels dataset.

Conclusion

Satellite imagery can be a valuable asset for disaster response. Many existing approaches in remote sensing, however, are only tailored towards singular objectives, such as segmentation of flooded buildings in sparsely populated areas using radar imagery. Computer vision can help make the most of Earth observation data.

In this work, we introduced a novel end-to-end trainable neural network architecture for fusion of multiresolution, multisensor, and multitemporal satellite images, showed that it outperforms state-of-the-art approaches on building footprint and flooded building segmentation tasks, and demonstrated that publicly available medium-resolution imagery alone can be used for effective segmentation of flooded buildings. Our approach is applicable to different types of flood events, and could be used to predict damage caused by other types of disasters. It substantially reduces the amount of time needed to produce flood maps for first responders compared to current methods. In future work, we plan to use our method to perform segmentation of buildings damaged by earthquakes and hurricanes, for both of which labeled satellite imagery is available. We hope that this work will 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 imagery of disaster sites along with our source code[1].

[1] `https://github.com/FrontierDevelopmentLab/multi3net`.

Acknowledgements

This research was conducted at the Frontier Development Lab (FDL), Europe. The authors gratefully acknowledge support from the European Space Agency, NVIDIA Corporation, Satellite Applications Catapult, and Kellogg College, University of Oxford.

References

Bilinski, P., and Prisacariu, V. 2018. Dense decoder shortcut connections for single-pass semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40(4):834–848.

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.

Couprie, C.; Farabet, C.; Najman, L.; and LeCun, Y. 2013. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*.

Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; and Raskar, R. 2018. Deepglobe 2018: A challenge to parse the earth through satellite images. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Haklay, M., and Weber, P. 2008. Openstreetmap: User-generated street maps. *Ieee Pervas Comput* 7(4):12–18.

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*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. 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.

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

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.

Ulaby, F., and Long, D. G. 2014. *Microwave Radar and Radiometric Remote Sensing*.

Van Etten, A.; Lindenbaum, D.; and Bacastow, T. M. 2018. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*.

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

Zebker, H. A., and Villasenor, J. D. 1992. Decorrelation in interferometric radar echoes. *IEEE Trans. Geoscience and Remote Sensing* 30:950–959.

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
Multi³Net: Segmenting Flooded Buildings via Fusion of Multiresolution, Multisensor, and Multitemporal Satellite Imagery

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. Optical sensors acquire images in the visible and short-wavelength portions of the electromagnetic spectrum. These images contain multiple spectral signatures of depicted scenes that generally contain information about chemical properties. Radar sensors are based on longer wavelengths than optical sensors allowing them to capture physical properties of the Earth's surface (Soergel, 2010). They are widely used in the fields of *Earth observation* and *remote sensing* as radar image acquisition is unaffected by cloud coverage or daylight (Ulaby and Long, 2014). To illustrate the appearance of these types of images, we illustrate examples of optical and radar medium resolution images along with one very high resolution image in Figure 4.

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

Evaluation Metrics We perform building footprint and flooded building segmentation and evaluate the results against multiple state-of-the-art benchmarks. As a benchmark metric, we report the *Intersection over Union* (IoU). IoU 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 this metric to assess the predictions of building footprints and flooded buildings. We report it using the shorthand bIoU. Represented as a confusion matrix, we have $\text{bIoU} \equiv \text{TP}/(\text{FP}+\text{TP}+\text{FN})$, where $\text{TP} \equiv$ True Positives, $\text{FP} \equiv$ False Positives, 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 the quantity $\text{TN}/(\text{TN}+\text{FP}+\text{FN})$. Additionally, we also report the mean of IoU values for both classes—background and building (or flooded buildings), and denote it by the shorthand mIoU. We also compute the pixel accuracy, the percentage of correctly classified pixels, and denote it as $\text{A} \equiv (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN})$.

Preprocessing In Section *Earth Observation*, we addressed 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 required to reproduce our approach. For radar data, however, preprocessing of the raw data is necessary to obtain numerical values that can be introduced to the network. Radar 'pixels' are composed of the real in-phase $\text{Re}(z)$ and imaginary quadrature $\text{Im}(z)$ components of the reflected electromagnetic signal expressed as a complex number z. In this study, we have employed *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. To preprocess the radar images, we followed the following steps Ulaby and Long (2014): (1) We performed *Radiometric calibration* to compensate for the effects of the sensor's relative orientation to the illuminated scene and the distance between them. (2) We reduced the noise induced by electromagnetic interference, known as *speckle*, by applying a spatial averaging kernel, termed as *multi-looking* in the radar community. (3) We normalized 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 averaged the intensity of all radar images over an extended temporal period, which is known as *temporal multi-looking* to further reduce the effect of speckle on the image. (5) We calculate the *interferometric coherence*

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

between images at two times z_1 and z_2, where the expectation $\mathbb{E}[\,\cdot\,]$ is estimated using a local *boxcar-function*, and $\boldsymbol{z}^*$ denotes the complex conjugate of $\boldsymbol{z}$. Coherence is a local similarity metric (Zebker and Villasenor, 1992) able to measure changes between pairs of radar images.

Data Addendum

Area of Interest To evaluate our approach, we chose multiple neighboring districts of Houston, Texas as our area of interest. Houston was flooded in the wake of Hurricane Harvey, a category 4 hurricane that formed over the Atlantic on August 17th 2017 and made landfall on the coast of the state of Texas on August 25th, 2017. The hurricane dissipated on September 2nd, 2017. In the early hours of August 28th, extreme rainfalls caused an 'uncontrolled overflow' of Houston's Addicks Reservoir and flooded the neighborhoods '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 (Haklay and Weber, 2008). Flooded buildings have been manually labeled through crowd sourcing as part of the DigitalGlobe Open Data initiative [2]. When preprocessing the data, we combine the building footprints obtained from OpenStreetMap with the point-wise annotations from DigitalGlobe to produce ground truth maps such as the one shown in Figure 5. The resulting geometry collections of buildings, illustrated in Figure 5b, and flooded buildings, shown in Figure 5c, are then rasterized in 2m and 10m grids, depending on the available satellite data. Figure 5a shows our area of interest using a high-resolution image overlaid with boundaries for the east and west partitions that were used for training and validation, respectively.

Data Preprocssing For radar images, we compute three different radar-based images: 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 resolution of 5m for August 23, 2017 and September 4, 2017. Additionally, we calculate the interferometric coherence for an image pair without flood-related changes acquired on June 6, 2017 and August 23rd, 2017, as well as for an image pair with flood-induced scene changes acquired on August 23rd, 2017 and September 4th, 2017 using Equation (1). Examples of coherence images generated this way are shown in Figures 4a and 4b. As the third radar component, 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 pre- and post-disaster into single, three-band images, respectively. The multiband images are then fed into the respective network streams.

Figures 4c and 4d in *Figures Addendum* show pre- and post-event images obtained from the Sentinel-2 satellite constellation on August 20, 2017 and September 4, 2017. Sentinel-2 measures the surface reflectances in 13 spectral bands of 10m, 20m, and 60m resolutions. We apply bilinear interpolation to the 20m band images to obtain an image representation with 10m ground sampling distance.

Finally, we extract rectangular tiles of size 960m$\times$960m from the set of satellite images to use as input samples for the network. This process is repeated on a 100m$\times$100m grid to produce overlapping tiles for model training and testing. The large tile overlap can be interpreted as an offline data augmentation step.

Method Addendum

Network Training We initialize the encoder with the weights of a ResNet34 (He et al., 2016) model pre-trained on ImageNet (Deng et al., 2009). In case of more than three input channels in the first convolution (due to the 10 spectral bands of the Sentinel-2 satellite), we initialize further channels with the average over the first convolutional filters of the RGB channels. In the following,

[2] `https://www.digitalglobe.com/opendata`

all networks are trained with the Adam optimizer (Kingma and Ba, 2014) using a learning rate of 0.01. The network parameters are optimized using a cross entropy loss

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

between ground truth $\boldsymbol{y}$ and prediction $\hat{\boldsymbol{y}}$. We anneal the learning rate according to the poly policy (power$= 0.9$) introduced in (Chen et al., 2018) and stop training upon loss convergence. We randomly sample 8 tiles of size 960m$\times$960m resolution (96px$\times$96px for optical satellite imagery, 192px$\times$192px for radar) from the dataset and use a batch size of 8 for the network training. We augment our training dataset by randomly rotating and flipping the image vertically and horizontally in order to create additional samples. We first train our network to segment building footprints and then re-use the weights for training on the class of flooded buildings.

Results Addendum

Building Footprint Segmentation—VHR Only 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), containing VHR RGB satellite imagery. Table 2 shows the performance of our approach on the Austin partition of the INRIA aerial labels dataset (Maggiori et al., 2017a). 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 multi-layer-perceptron stream. Ohleyer (2018) employ a Mask-RCNN (He et al., 2017) instance segmentation network for the task. Our model performed better than the current state-of-the-art, obtaining 7.8% higher bIoU than Ohleyer (2018).

Building Footprint Segmentation—Single Sensors In this section we present the results for building footprint segmentation based on imagery from individual sensors. Table 3 shows that the information conveyed in optical bands has the most influence on the performance of out network. Both approaches based on Sentinel-2 and VHR optical imagery performed better than the model trained on Sentinel-1 radar data. This experiment also shows that the usage of higher resolution data improves the quality of predictions.

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%**

Table 3: Building footprint segmentation results for images obtained from individual sensors with different resolutions.

Building Footprint Segmentation—Image Fusion The fusion of multiresolution and multisensor satellite imagery further improves the prediction quality. The results in Table 4 show that the highest accuracy was achieved when all data sources were fused.

Data	mIoU	bIoU	Accuracy
S-1 + S-2	76.1%	70.5%	87.3%
S-1 + S-2 + VHR	**79.9%**	**75.2%**	**89.5%**

Table 4: Building footprint segmentation results for a fusion of images obtained from multiple sensors.

Figure 6 shows qualitative results for building footprint segmentation when fusing images from multiple sensors. The model using Sentinel-1 and Sentinel-2 data produces accurate predictions (76.1% mIoU), but its performance improves by 3.8% when VHR imagery is fused with the other multisensor data.

Figures Addendum

(a) Sentinel-1 (192 px) coherence pre-event

(b) Sentinel-1 (192 px) coherence post-event

(c) Sentinel-2 (96 px) pre-event

(d) Sentinel-2 (96 px) post-event

(e) Very high-res. (1560 px) post-event

Figure 4: One image tile of 960m×960m is used as network input. Figures 4a and 4b illustrate Sentinel-1 coherence images before and after the flooding event, whereas Figures 4c and 4d show a RGB representation of multispectral Sentinel-2 optical data. Figure 4e, shows the high level of spatial details in a very high-resolution image.

(a) VHR imagery with dataset boundaries

(b) OpenStreetMap building footprints

(c) Annotated flooded buildings

Figure 5: Images illustrating the size and extent of the dataset (Figure 5a), available rasterized ground truth annotations as OpenStreetMap building footprints (Figure 5b), and expert-annotated labels of flooded buildings (Figure 5c).

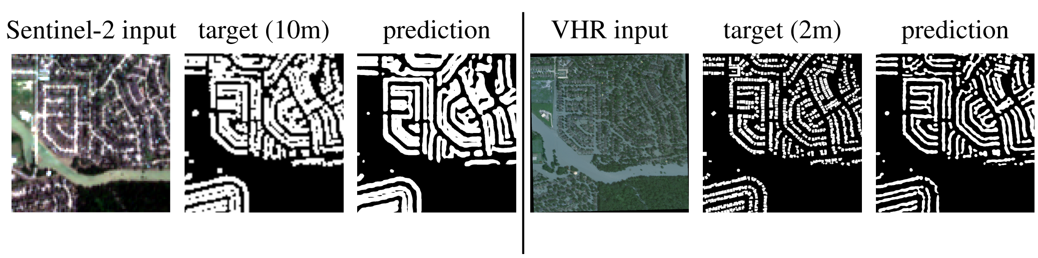

Figure 6: Prediction targets and prediction results for building footprint segmentation using Sentinel-1 and Sentinel-2 inputs fused at a 10m resolution (left panel) and using Sentinel-1, Sentinel-2, and VHR inputs fused at a 2m resolution (right panel).