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 segmentation of flooded buildings can be performed well using only freely available medium-resolution data and can be further 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 Multi3Net, a novel approach for rapid and accurate flood damage segmentation by fusing multiresolution and multisensor satellite imagery in a convolutional neural network. The network consists of multiple deep encoder-decoder streams, in which each individual stream is able to produce 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 once at least a single satellite acquisition has been successful and subsequently be improved upon once 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) and PSP-net (Zhao et al., 2017) augment these using an improved encoder architecture with a higher receptive field and additional modules such as ASPP or PSP.

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.

Multi³Net

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 $3 \times 3$ convolution and a PReLU activation function to learn a mapping from latent space to label space. This way, Multi³Net 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 can be fused using a variety of approaches. We examine early as well as late fusion. In an early fusion approach, we upsample the image tensors of all satellites, concatenate them into one large input tensor, and then process the information within a single network. In a late fusion approach, each image type is fed into a dedicated information processing stream 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 [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
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 the CNN outputs 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.

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 $3 \times 3$ convolutions with PReLU activations and a $1 \times 1$ 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 $96 \times 96$ px 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.

**Multitemporal Fusion** 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. More complex approaches, such as a two stream approach with shared encoder weights similar to Siamese networks [Melekhov, Kannala, and Rahtu, 2016] or subtracting the activations of feature maps, did not improve our early-fusion-with-concatenation approach.

**Data**

We use medium-resolution satellite imagery with a pixel size of 5m–10m acquired before and after Hurricane Harvey 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 creation of these images, example images, and Earth observation terminology can be found in the supplementary material.

Figure 2: The context aggregation module extracts and combines image features at different image resolutions, similar to [Zhao et al., 2017].
122

Table 1 shows that fusing images from all resolutions and sensors across time yielded the best results (75.3% mIoU), but that fusing only globally available medium-resolution Sentinel-1 and Sentinel-2 images performed well as well (59.7% mIoU). 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 the majority of 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 and found that our model performed best (73.4% bIoU) at this task (see Table 2).

129

Table 1: Mean IoU (mIoU), building IoU (bIoU), and pixel accuracy for flooded building segmentation using Multi $^3$ Net.

<table>
<thead>
<tr>
<th>Data</th>
<th>mIoU</th>
<th>bIoU</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-1 + S-2</td>
<td>59.7%</td>
<td>34.1%</td>
<td>86.4%</td>
</tr>
<tr>
<td>VHR</td>
<td>74.2%</td>
<td>56.0%</td>
<td>93.1%</td>
</tr>
<tr>
<td>S-1 + S-2 + VHR</td>
<td><strong>75.3%</strong></td>
<td><strong>57.5%</strong></td>
<td><strong>93.7%</strong></td>
</tr>
</tbody>
</table>

Table 2: Building IoU (bIoU) and pixel accuracy for building footprint segmentation using VHR imagery of Austin in the INRIA aerial labels dataset Maggiolo et al. (2017a).

<table>
<thead>
<tr>
<th>Model</th>
<th>bIoU</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maggiolo et al.</td>
<td>61.2%</td>
<td>94.2%</td>
</tr>
<tr>
<td>Ohleyer (2018)</td>
<td>65.6%</td>
<td>94.1%</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>73.4%</strong></td>
<td><strong>95.7%</strong></td>
</tr>
</tbody>
</table>

132

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.

133

<table>
<thead>
<tr>
<th>Data</th>
<th>mIoU</th>
<th>bIoU</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-1 + S-2</td>
<td>59.7%</td>
<td>34.1%</td>
<td>86.4%</td>
</tr>
<tr>
<td>VHR</td>
<td>74.2%</td>
<td>56.0%</td>
<td>93.1%</td>
</tr>
<tr>
<td>S-1 + S-2 + VHR</td>
<td><strong>75.3%</strong></td>
<td><strong>57.5%</strong></td>
<td><strong>93.7%</strong></td>
</tr>
</tbody>
</table>

Table 1: Mean IoU (mIoU), building IoU (bIoU), and pixel accuracy for flooded building segmentation using Multi $^3$ Net.

134

Conclusions

An increasing number of satellites monitor the Earth’s surface. In disaster response, where fast information extraction is key for local responders to coordinate relief efforts, Earth observation data can be an extremely valuable asset. 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.

134

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 broadly applicable to different types of flood events and could be used to predict damage caused by other disasters as well. It is easy to deploy and 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 further research into image fusion for disaster relief. 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


