

Noname manuscript No.
(will be inserted by the editor)

Building Damage Annotation on Post-Hurricane Satellite Imagery Based on Convolutional Neural Networks

Quoc Dung Cao · Youngjun Choe

Received: date / Accepted: date

Abstract After a hurricane, damage assessment is critical to emergency managers for efficient response and resource allocation. One way to gauge the damage extent is to quantify the number of flooded/damaged buildings, which is traditionally done by ground survey. This process can be labor-intensive and time-consuming. In this paper, we propose to improve the efficiency of building damage assessment by applying image classification algorithms to post-hurricane satellite imagery. At the known building coordinates (available from public data), we extract square-sized images from the satellite imagery to create training, validation, and test datasets. Each square-sized image contains a building to be classified as either ‘Flooded/Damaged’ (labeled by volunteers in a crowd-sourcing project) or ‘Undamaged’. We design and train a convolutional neural network from scratch and compare it with an existing neural network used widely for common object classification. We demonstrate the promise of our damage annotation model (over 97% accuracy) in the case study of building damage assessment in the Greater Houston area affected by 2017 Hurricane Harvey.

Keywords image classification · neural network · damage assessment · building · remote sensing

Quoc Dung Cao
Department of Industrial and Systems Engineering, University of Washington, Seattle, WA, USA.

Tel.: +1 (217) 979-0117
E-mail: qcao10@uw.edu

Youngjun Choe
Department of Industrial and Systems Engineering, University of Washington, Seattle, WA, USA.
3900 E Stevens Way NE, Seattle, WA 98195, USA
Tel.: +1 (206) 543-1427
E-mail: ychoe@u.washington.edu

1 Introduction

When a hurricane makes landfall, situational awareness is one of the most critical needs that emergency managers face before they can respond to the event. To assess the situation and damage, the current practice largely relies on emergency response crews and volunteers to drive around the affected area, which is also known as windshield survey. Another way to assess hurricane damage level is flood detection through synthetic aperture radar (SAR) images (e.g., see the work at the Dartmouth Flood Observatory [14]), or the damage proxy map to identify regional-level damages on the built environment (e.g., the Advanced Rapid Imaging and Analysis (ARIA) Project by Caltech and NASA [7]). SAR imagery is useful in terms of mapping different surface features, texture, or roughness pattern but is harder for laymen to interpret than optical sensor imagery. The resolutions of virtually all SAR images of today are too coarse to permit the building-level (as opposed to regional-level) damage assessment. Also, satellites equipped with SAR sensors are far fewer than those with optical sensors, making timely and frequent data collection challenging. In this paper, we focus on using optical sensor imagery as a more intuitive and accessible way to analyze hurricane damage by distinguishing damaged buildings from the ones still intact. From here onwards, we will refer to optical sensor imagery as ‘imagery’.

Recently, imagery taken from drones and satellites started to help improve situational awareness from a bird’s eye view, but the process still relies on human visual inspection of captured imagery, which is generally time-consuming and unreliable during an evolving disaster. Computer vision techniques, therefore, can be particularly useful. Given the available imagery, our proposed method can automatically annotate ‘*Flooded/Damaged Building*’ vs. ‘*Undamaged Building*’ on satellite imagery of an area affected by a hurricane. The annotation results can enable stakeholders (e.g., emergency managers) to better plan for and allocate necessary resources. With decent accuracy and quick runtime, this automated annotation process has potential to significantly reduce the time for building situational awareness and responding to hurricane-induced emergencies.

The satellite imagery data used in this paper covers the Greater Houston area before and after Hurricane Harvey in 2017 (Figure 1). The flooded/damaged buildings were labeled by volunteers through the crowd-sourcing project, Tomnod [3]. We then process, filter, and clean the dataset to ensure that it has correct labels and can be learned appropriately by a learning algorithm.

By sharing the dataset and code used in this paper (see the appendix), we hope that other researchers can build upon this study and help further improve computer vision-based damage assessment process. The shared code includes a pre-trained deep-learning architecture that achieves the best classification accuracy (detailed in Section 4). It can facilitate transfer learning either in feature extraction, fine-tuning, or as a baseline model to speed up the learning process for future hurricane events.

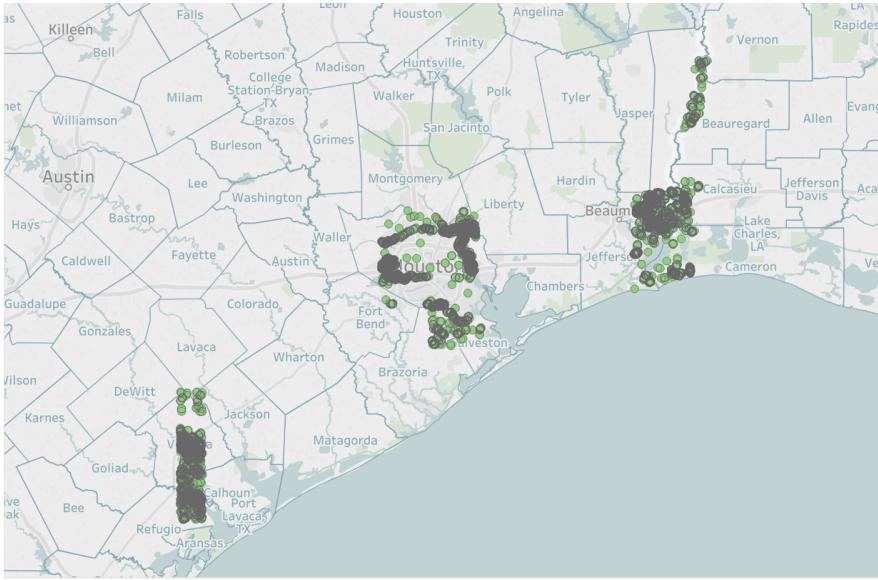


Fig. 1 The Greater Houston area was affected by Hurricane Harvey in 2017. The green circles represent the coordinates of flooded/damaged structures tagged by Tomnod volunteers.

The remaining of this paper is organized as follows. In Section 2, we present a brief review of convolutional neural networks, machine learning-based damage annotation work on post-hurricane satellite imagery, and challenges in the damage annotation on satellite imagery. Section 3 describes our proposed methodological framework for the damage annotation. Details of the implementation and discussion of the results are presented in Section 4. Finally, Section 5 concludes this paper and draws some future research directions.

2 Background

2.1 Convolutional neural network

The convolutional neural network (CNN) [24] often yields outstanding results over other algorithms for computer vision tasks such as object categorization [20], image classification [11, 23], and object recognition [12]. Variations of CNN have been successfully applied to remote sensing image processing tasks [35] such as aerial scene classification [33, 28, 26], SAR imagery classification [36], or object detection in unmanned aerial vehicle imagery [9].

Structurally, CNN is a feed-forward network that is particularly powerful in extracting hierarchical features from images. The common structure of CNN has three components: the convolutional layer, the sub-sampling layer, and the fully connected layer as illustrated in Figure 2.

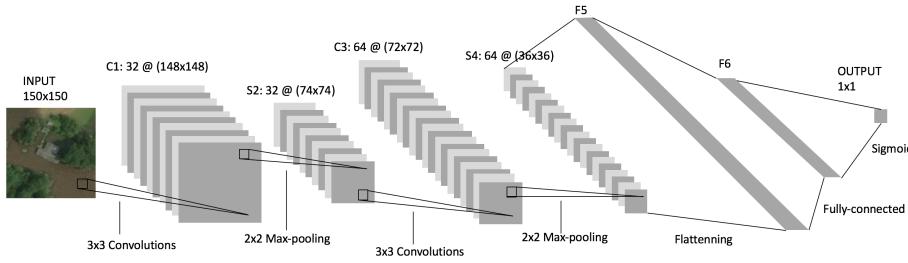


Fig. 2 A convolutional neural network inspired by LeNet-5 architecture in [24]; C: Convolutional layer, S: Sub-sampling layer, F: Fully connected layer; 32@(148x148) means there are 32 filters to extract features from the input image, and the original input size of 150x150 is reduced to 148x148 since no padding is added around the edges during 3×3 convolution operations so 2 edge rows and 2 edge columns are lost; 2x2 Max-pooling means the data will be reduced by a factor of 4 after each operation; Output layer has 1 neuron since the network outputs the probability of one class ('Flooded/Damaged Building') for binary classification.

In the convolutional layer (C in Figure 2), each element (or neuron) of the network in a layer receives information from a small region of the previous layer. A 3×3 convolutional filter will take a dot product of 9 weight parameters with 9 pixels (3×3 patch) of the input, and the resulting value is transformed by an activation function to become a neuron value in the next layer. The same region can yield many information maps to the next layer through many convolutional filters. In Figure 2, at convolutional layer C1, we have 32 filters that represent 32 ways to extract features from the previous layers and form a stack of 32 feature matrices. Another advantage of CNN is its robustness to shift of features in the input images [16]. This is crucial since in many datasets, objects of interest are not necessarily positioned right at the center of the images and we want to learn the features, not their positions.

In the sub-sampling layer (S in Figure 2), the network performs either local averaging or max pooling over a patch of the input. If the sub-sampling layer size is 2×2 such as S2, local averaging will yield the mean of the 4 nearby convoluted pixel values, whereas max pooling will yield the maximum value among them. Essentially, this sub-sampling operation reduces the input feature matrix to half its number of columns and rows, which helps to reduce the resolution by a factor of 4 and the network's sensitivity to distortion.

After the features are extracted and the resolution reduced, the network will flatten the final stack of feature matrices into a feature vector and pass it through a sequence of fully connected layers (F in Figure 2). Each subsequent layer's output neuron is a dot product between the feature vector and a weight vector, transformed by a non-linear activation function. In this paper, the last layer has only 1 neuron, which is the probability of a reference class ('Flooded/Damaged building').

As mentioned, the dot products are transformed by an activation function. This gives a neural network, with adequate size, the ability to model any func-

tion. Some common activation functions include sigmoid $f(x) = \frac{1}{1+e^{-x}}$, rectified linear unit (ReLU) $f(x) = \max(0, x)$, and leaky ReLU $f(x) = \max(\alpha x, x)$, with $0 < \alpha \ll 1$. There is no clear reason to choose any specific function over the others to improve performance of a network. However, using ReLU may speed up the training of the network without affecting the performance [18].

2.2 Machine learning-based damage annotation on post-hurricane satellite imagery

Machine learning on remote sensing imagery is actively researched to assess damage from or susceptibility to various hazards such as earthquake [30], landslide [6, 19], tsunami [29], and wildfire [27]. Such methods showed remarkable promise. However, often leveraging unique characteristics of each hazard type, they are not directly applicable to damage annotation on post-hurricane imagery and sometimes require extensive pre-processing. For example, the work in [30] for post-earthquake damage assessment closely resembles our work. Both pre-event and post-event are used to extract building's roof and texture features. This process requires input from expert operators and rely on texture feature of the buildings in the imagery, which sometimes may not be available depending on means of collection. In another work, [29] classifies regions of pixels into water, vegetation, urban, and bare land, which does not provide the building level granularity as we pursue. Furthermore, since we rely purely on the widely available optical sensor imagery, convolutional neural network is one of the most suitable model classes due to its flexible feature extraction capability and architecture.

Some recent studies used machine learning to assess post-hurricane damages on satellite imagery. A small project studied detecting *flooded roads* by comparing pre-event and post-event satellite imagery [21] but the method is not applicable to other types of damages. Two commercial vendors of satellite imagery also separately developed unsupervised algorithms to detect flooded area using spectral signature of impure water (which is not available from the pansharpened satellite images in our data) [4, 5]. Before deep learning era, a method using a pattern recognition template set was applied to detect hurricane damages in *multispectral* images [8] but the method is not applicable to our pansharpened images.

2.3 Challenges in damage annotation on satellite imagery

There are multiple challenges in damage annotation on satellite imagery. First, satellite imagery resolution is not as high as various benchmark datasets commonly used to train neural networks (NNs) (e.g., ImageNet [23] and traffic signs [12]) with respect to the objects of interest. Dodge & Karam [15] studied the performance of NNs under quality distortions and highlighted that NNs could be prone to errors in blurry and noisy images. Although our dataset is

of relatively high resolution (e.g., one of the satellites capturing the imagery is GeoEye-1, which has 46cm panchromatic resolution [1]), it is still far from the resolution of common-object detection datasets (e.g., animals, vehicles). In fact, the labeling task on satellite imagery is hard even with human visual inspection, which leads to another challenge. The volunteers' annotation could be erroneous. To limit this, the crowd-sourcing platform has a proprietary system that computes the agreement score of each label. In this paper, we ignore this information to gather as many labels as possible and take the given labels as ground truth since limited size of training data could be a critical bottleneck for models with many parameters to learn such as NNs. Third, there are some inconsistencies in image quality. Since the same region can be captured multiple times on different days, the same coordinate may have multiple images of different qualities (e.g., due to pre-processing), as shown in Figure 3. In summary, effective learning algorithms should overcome the challenges from low-resolution images, noisy labels, and inconsistent image qualities.

3 Methodology

In this section, we describe our end-to-end methodological framework from collecting, processing, featurizing data to building the convolutional neural network to classify whether a building in a satellite image is flooded/damaged or not.

3.1 Data description

The satellite imagery of the Greater Houston area was captured by optical sensors with sub-meter resolution, preprocessed (e.g., orthorectification and atmospheric compensation), and pansharpened by the image provider. The raw imagery consists of around four thousand image strips taken on multiple days (each strip is roughly 1GB and has around 400 million pixels with RGB bands). Some strips overlap and have black pixels in the overlapped region. Some images are also covered fully or partially by clouds. Figure 4 shows a typical strip in the dataset and Figure 5 shows some examples of *low quality* images (from the perspective of model training) that we chose to discard.

3.2 Damage annotation

We present here our methodological framework (Figure 6) that starts from raw data input to create damage annotation output. The first step is to process the raw data to create training-ready data by using a cropping window approach. Essentially, the building coordinates, which can be easily obtained from public data (e.g., OpenStreetMap [2]), can be used as the centers of cropping. We use the building coordinates already associated with the damage labels from Tomnod. A window is then cropped from the raw satellite imagery to create a



Fig. 3 Different orthorectification and pre-processing quality of the same location on different days.

data sample. Tomnod volunteers' annotation of flooded/damaged buildings is taken as the ground truth for the positive label, 'Flooded/Damaged building'. At the same coordinates, we crop windows from the imagery captured before the hurricane to create negative data samples, labeled 'Undamaged building'.

The optimal window size depends on various factors including the image resolution and building footprint sizes. Too small windows may limit the background information contained in each sample, whereas too large ones may introduce unnecessary noise. We keep the window size as a tuning hyperparameter in the model. A few sizes are considered such as 400x400, 128x128, 64x64, and 32x32.

The cropped images are then manually filtered to ensure the high quality of the dataset. To let the model generalize well, we only discard the images that can obviously hamper the algorithm's learning process, such as the example images in Figure 5. The cleaned images are then split into training,

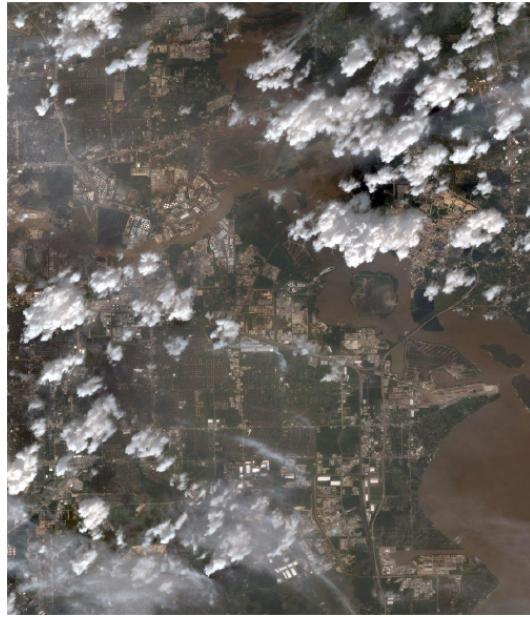


Fig. 4 A typical strip of image in the dataset.

validation, and test sets and fed to a convolutional neural network for damage annotation as illustrated in Figure 6. Validation accuracy is monitored to tune the necessary hyper-parameters (including the window size).

3.3 Data processing

As described above, the data generation starts from a building coordinate. Since there are multiple raw images containing the same coordinates, there are duplicate images with different quality. This can potentially inflate the prediction accuracy as the same coordinate may appear in both the training and test sets. We maintain a set of the available coordinates and make sure each coordinate is associated with a unique, “good-quality” image in the final dataset through a semi-automated process. We first automatically discard the totally blacked out images for each coordinate, and keep the first image we encounter that is not totally black. The resulting set of images are manually filtered to eliminate the images that are partially black or covered by clouds.

3.4 Data featurization

Since we control the window size based on physical distance, there could be round-off errors when converting the distance to the number of pixels. Therefore, we project them into the same feature dimension. For instance, both a



Fig. 5 Examples of discarded images during the data cleaning process due to their potential to hinder model training.

128x128 image and a 127x129 image are projected into 150x150 dimension. The images are then fed through a CNN to further extract useful features, such as edges, as illustrated in Figure 7.

How to construct the most suitable CNN architecture is an ongoing research problem. The common practice, known as *transfer learning*, is starting with a known architecture and fine-tuning it. We experiment with a well-known architecture, VGG-16 [31], and modify the first layer to suit our input dimension. VGG-16 is known to perform very well on the ImageNet dataset for common object classification.

However, because of the substantial differences between the common object classification and our flooded/damaged building classification, we also build our own network from scratch. We carefully consider proper hyper-parameters, as similarly done in [25]. Our basis for determining the size and depth of a customized network is to monitor the information flow through the network

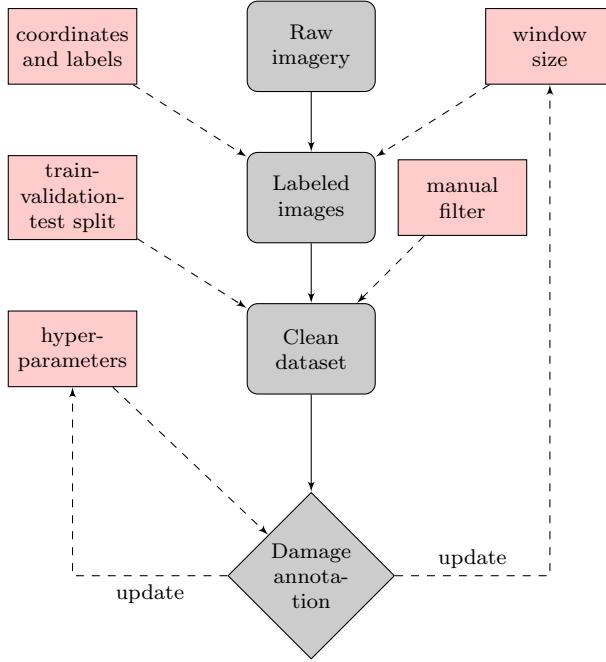


Fig. 6 The damage annotation framework.

and stop enlarging the network when there are too many *dead* filters (*i.e.*, blank filters that do not carry any further information to the subsequent layers in the network). Due to the nature of the rectified linear unit (ReLU), which is defined as $\max(0, x)$, there will be many zero weights in the hidden layers. Although sparsity in the layers can promote the model to generalize better, it may cause the problem on gradient computation at 0, which in turns does not update any parameters, and hurt the overall model performance [18,34]. We see that in Figure 8 after four convolutional layers, about 30% of the filters are dead and will not be activated further. This is a significant stopping criterion since we can avoid a deep network such as VGG-16 to save the computational time and safeguard satisfactory information flow in the network at the same time.

We present our customized network architecture that achieves the best result in Table 1. The network begins with four convolutional and max pooling layers and ends with two fully connected layers.

In our CNN structure, with four convolutional layers and two fully connected layers, there are already about 3.5 million parameters to train, given 67,500 pixels as an input vector for each image. The VGG-16 structure [31], with thirteen convolutional layers, has almost 15 million trainable parameters, which can over-fit, require more resources, and reduce generalization performance on the testing data. In addition, as discussed in [25], the network depth

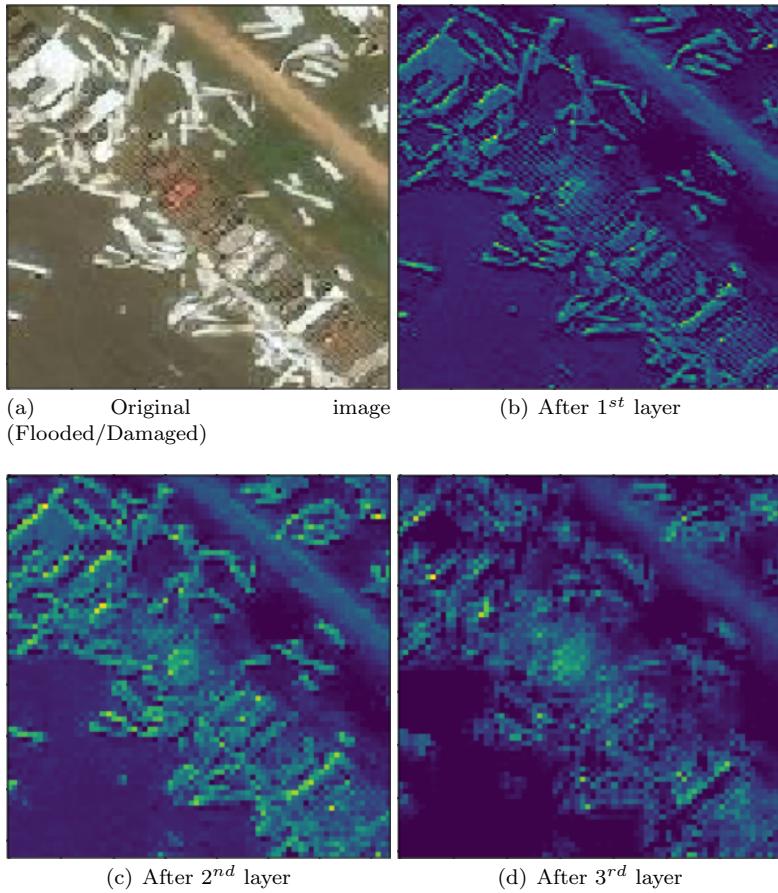


Fig. 7 Information flow within one filter after each convolutional layer. The initial layers act as a collection of edge extraction. At a deeper layer, the information is more abstract and less visually interpretable.

should depend on the complexity of the features to be extracted from the image. Since we have only two classes of interest, a shallower network can be favourable in terms of training time and generalization.

3.5 Image classification

Due to the limited availability of pre-event images and the exclusion of some images (e.g., due to cloud coverage) in the *Flooded/Damaged* and *Undamaged* categories, our dataset is unbalanced with the majority class being *Flooded/Damaged*. Thus, we split the dataset into training, validation, and test datasets as follows. We keep the training and validation sets balanced and leave the remaining data

Table 1 Convolutional neural network architecture that achieves the best result.

| Layer type | Output shape | Number of trainable parameters |
|-----------------------------|---------------|--------------------------------|
| Input | 3@((150x150) | 0 |
| 2-D Convolutional 32@(3x3) | 32@((148x148) | 896 |
| 2-D Max pooling (2x2) | 32@((74x74) | 0 |
| 2-D Convolutional 64@(3x3) | 64@((72x72) | 18,496 |
| 2-D Max pooling (2x2) | 64@((36x36) | 0 |
| 2-D Convolutional 128@(3x3) | 128@((34x34) | 73,856 |
| 2-D Max pooling (2x2) | 128@((17x17) | 0 |
| 2-D Convolutional 128@(3x3) | 128@((15x15) | 147,584 |
| 2-D Max pooling (2x2) | 128@((7x7) | 0 |
| Flattening | 1x6272 | 0 |
| Dropout | 1x6272 | 0 |
| Fully connected layer | 1x512 | 3,211,776 |
| Fully connected layer | 1x1 | 513 |

Note: The total number of trainable parameters is 3,453,121. C@($A \times B$) is interpreted as that there are a total of C matrices of shape ($A \times B$) stacked on top of one another to form a three-dimensional tensor. 2-D Max pooling layer with (2 \times 2) pooling size means that the input tensor's size will be reduced by a factor of 4.

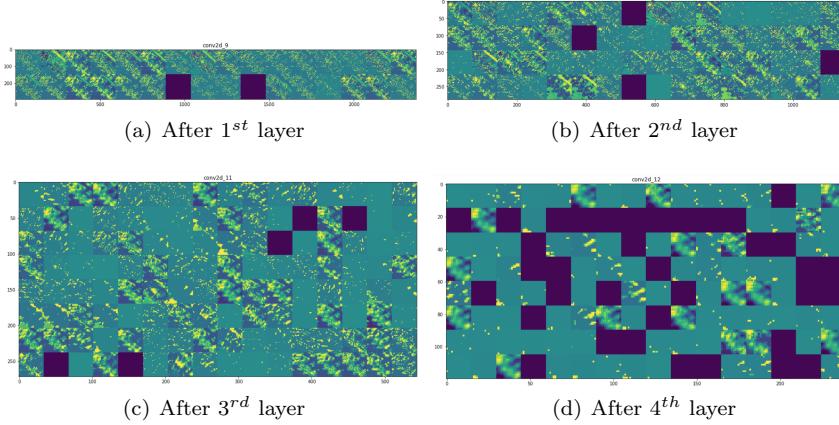


Fig. 8 Information flow in all filters after each convolutional layer. The sparsity increases with the depth of the layer, as indicated by the increasing number of dead filters.

to construct two test sets, a balanced set and an unbalanced (with a ratio of 1:8) set.

The first performance metric is the classification accuracy. In contrast to the balanced test set, we note that the baseline accuracy for the *unbalanced* test set is $8/9 = 88.89\%$ (greater than the random guess accuracy, 50%), which can be achieved by annotating all buildings as the majority class *Flooded/Damaged*. In addition, as the classification accuracy is sometimes not the most pertinent performance measure, we also monitor the area under the receiver operating characteristic curve (AUC), which is a widely-used criterion to measure the

classification ability of a binary classifier under a varying decision threshold [13].

4 Implementation and Result

We train the neural networks using the *Keras* library with TensorFlow backend with a single NVIDIA K80 Tesla GPU. The network weights are initialized using Xavier initializer [17]. The mini batch size for the stochastic gradient descent optimizer is 32.

After the data cleaning process, our dataset contains 14,284 positive samples (*Flooded/Damaged*) and 7,209 negative samples (*Undamaged*) at unique geographical coordinates. 5,000 samples of each class are in the training set. 1,000 samples of each class are in the validation set. The rest of the data are reserved to form the test sets, i.e. in the balanced test set, there will be 1,000 samples of each class, and in the unbalanced test set, there will be 8,000 samples of *Flooded/Damaged* class and 1,000 samples of *Undamaged* class.

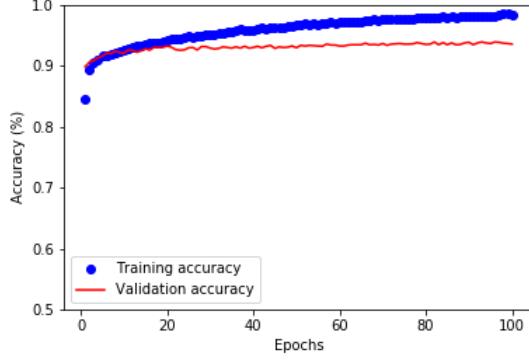
Due the expensive computational cost of training the CNN, we investigate selected combinations of the hyper-parameters in a greedy manner, instead of tuning all the hyper-parameters through a full grid search or full cross-validation. For example, we investigate the performance of a model with multiple window sizes (400x400, 128x128, 64x64, and 32x32) and select the 128x128 window size.

We also implement a logistic regression (LR) on the featurized data to see how it compares to fully connected layers. Although LR under-performs in most cases, it still achieves good accuracy (little over 90% in Table 2). This illustrates that the image featurization through the network works well enough that a simple algorithm like LR can perform well on this data.

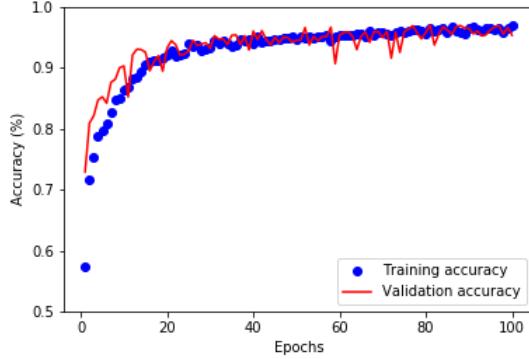
For activation functions in the CNN, a rectified linear unit (ReLU) is a common choice, thanks to its simplicity in gradient computation and prevention of vanishing gradient, which is common with other activation functions such as sigmoid or hyperbolic tangent. But, as seen in Figure 8, clamping the activation at 0 could potentially cause a lot of filters to be dead. Therefore, we also consider using a leaky ReLU activation with $\alpha = 0.1$ based on the survey in [34]. However, leaky ReLU turns out to not significantly improve the accuracy in our implementation (Table 2).

To counter over-fitting, which is a recurrent problem of deep learning, we also adopt data augmentation in the training set through random rotation, horizontal flip, vertical and horizontal shift, shear, and zoom. This can effectively increase the number of training samples to ensure better generalization and achieve better validation and test accuracy (Note that we do not perform data augmentation in the validation and test sets). Furthermore, we also employ 50% dropout and L2 regularization with $\lambda = 10^{-6}$ in the fully connected layer. Dropout [32] is an effective method to prevent over-fitting, especially in neural networks with many neurons. The method prevents neurons from remembering too much training data by dropping out randomly chosen neurons

and their connections during the training time. L2 regularization is one of the regularization techniques that has been shown to perform better on ill-poised problems or noisy data. Early application of the regularization in computer vision can be traced back to edge detection in images where the changes in intensity in an image are considered noisy [10]. These measures are shown to fight over-fitting effectively and significantly improve the validation accuracy in Figure 9.



(a) Without drop-out and image augmentation, over-fitting seems to happen after about 10 epochs as the validation accuracy separates from the training accuracy.

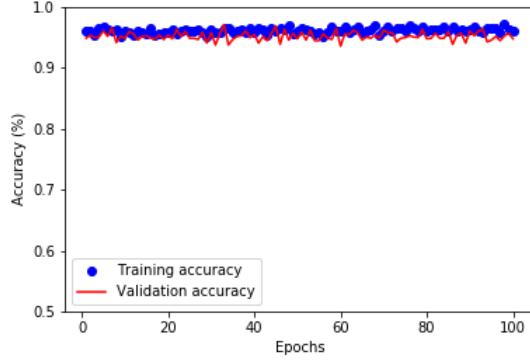


(b) No apparent sign of over-fitting can be seen as the validation accuracy follows the training accuracy.

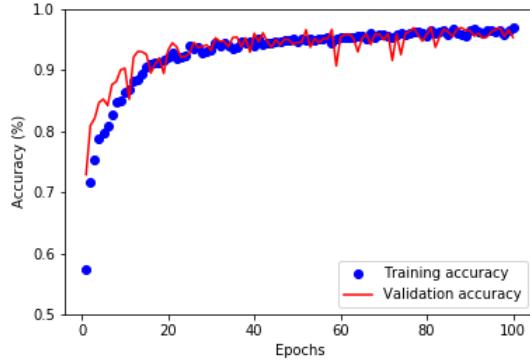
Fig. 9 Over-fitting is prevented using data augmentation, drop-out, and regularization.

As mentioned in Section 3.4, we consider using a pre-built architecture VGG-16 (transfer learning) and building a network from scratch. In Figure 10, we see that the deeper and larger network can achieve a high-level validation accuracy earlier, but the accuracy pretty much plateaus (*i.e.*, over-fitting hap-

pens) after a few epochs. Our simpler network can facilitate learning gradually, where the validation accuracy keeps increasing to achieve a higher value than the deeper network, and takes about 75% less training time.



(a) Transfer learning using pre-built network



(b) Custom network

Fig. 10 Comparison between using a pre-built network and our network. The two networks almost have the same level of performance except our network achieves a slightly better accuracy with a much smaller network size. It is also noticeable that due to large number of pre-trained parameters, the bigger network achieves high accuracy right at the beginning but fails to improve subsequently.

We use two adaptive, momentum-based optimizers, RMSprop and Adam [22], with the initial learning rate of 10^{-4} . Adam generally leads to about 1% higher validation accuracy and less noisy learning in our implementation.

Table 2 summarizes the performances of various models. The best performing model is our customized network with data augmentation and dropout using Adam optimizer, which can achieve 97.08% accuracy on the unbalanced test set. The AUC metric is computed and shows a satisfying result of 99.6%

on the unbalanced test set. The F1 Score metric also shows a consistent choice of the best model.

Table 2 Model performance.

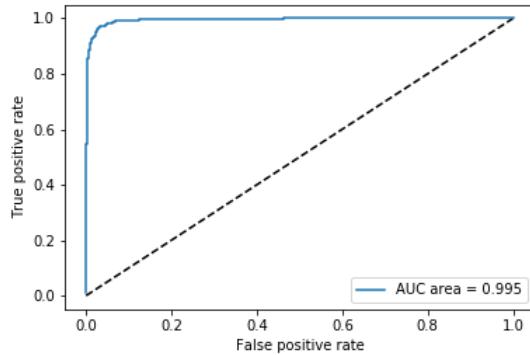
| Model | Validation Accuracy | Test Accuracy (Balanced) | Test Accuracy (Unbalanced) | F1 Score |
|--|---------------------|--------------------------|----------------------------|---------------|
| CNN | 95.8% | 94.69% | 95.47% | 0.9575 |
| Leaky CNN | 96.1% | 94.79% | 95.27% | 0.9558 |
| CNN + DA + DO | 97.44% | 96.44% | 96.56% | 0.9674 |
| CNN + DA + DO (Adam) | 98.06% | 97.29% | 97.08% | 0.9723 |
| Transfer + DO | 93.45% | 92.8% | 92.8% | 0.9304 |
| Transfer + DA + DO | 91.1% | 88.49% | 85.99% | 0.8800 |
| LR + L2 | 93.55% | 92.2% | 91.45% | 0.7713 |
| SVM + L2 | 92.02% | 91.85% | 90.95% | 0.7002 |
| Transfer + DA + FDO | 96.5% | 95.34% | 95.73% | 0.9594 |
| Leaky + Transfer + DA + FDO + L2 | 96.13% | 95.59% | 95.68% | 0.9598 |
| Leaky + Transfer + DA + FDO + L2(Adam) | 97.5% | 96.19% | 96.21% | 0.9643 |

Legend: CNN: Convolutional Neural Network; Leaky: Leaky ReLU activation function, else, the default is ReLU; DA: Data Augmentation; LR: Logistic Regression with features built by convolutional operations; L2: L2 regularization; SVM: Support vector machine classifier; (Adam): Adam optimizer, else, the default is RMSprop optimizer; DO: 50% dropout only in the fully connected layer; FDO: Full dropout, *i.e.*, 25% dropout after every max pooling layer and 50% in the fully connected layer; Transfer: Transfer learning using VGG-16 architecture.

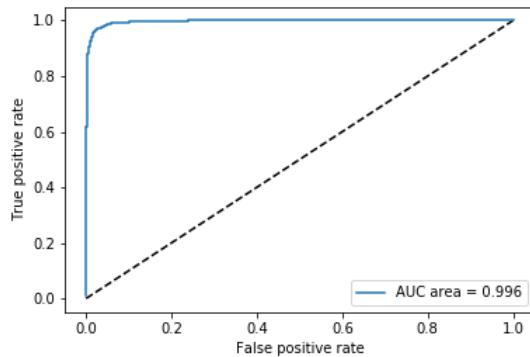
Although the overall result is satisfactory, we also investigate a few typical cases where the algorithm makes wrong classification to see if any intuition can be derived. Figure 12 shows some of the false positive cases. We hypothesize that the algorithm could predict the damage through flood water and/or debris edges. Under such hypothesis, the cars in the center of Figure 12(a), the lake water in Figure 12(b), the cloud covering the house in Figure 12(c), and the trees covering the roof in Figure 12(f) can potentially mislead the model. For the false negative cases in Figure 13, it is harder to make sense out of the prediction. Even through careful visual inspection, we cannot see Figures 13(a)(b) as being flooded/damaged. These could potentially be labeling mistakes by the volunteers. On the other hand, Figures 13(e)(f) are clearly flooded/damaged, but the algorithm misses them.

5 Conclusion and Future Research

We demonstrated that convolutional neural networks can automatically annotate flooded/damaged buildings on post-hurricane satellite imagery with high accuracy. While our data is specific to the geographical condition and building properties in the Greater Houston area during Hurricane Harvey, the model can be further improved and generalized to other future hurricane events in



(a) AUC of balanced test set



(b) AUC of unbalanced test set

Fig. 11 AUC for the balanced and unbalanced test sets using our best performing model—CNN + DA + DO (Adam)—in Table 2.

other regions by collecting more positives samples from other past events and negative samples from other areas. From more data, we can obtain both more robust models as well as sets of hyper-parameters. As mentioned in Section 3, the cropping window size is dependent on the resolution of the available imagery (the mode of collection), as well as the typical building footprint size in the region. From our grid search on the window size, as long as the window captures the building adequately, the performance does not vary significantly so the data collection operator can still have some flexibility about the collection methods (e.g., using drones).

For faster disaster response, a model should be able to process and annotate on low-quality images. For example, images taken right after a hurricane landfall can be covered largely by cloud. Also, image providers might not have enough time to pre-process images well due to the urgency of situation. We will



Fig. 12 False positive examples (the label is *Undamaged*, whereas the prediction is *Flooded/Damaged*).

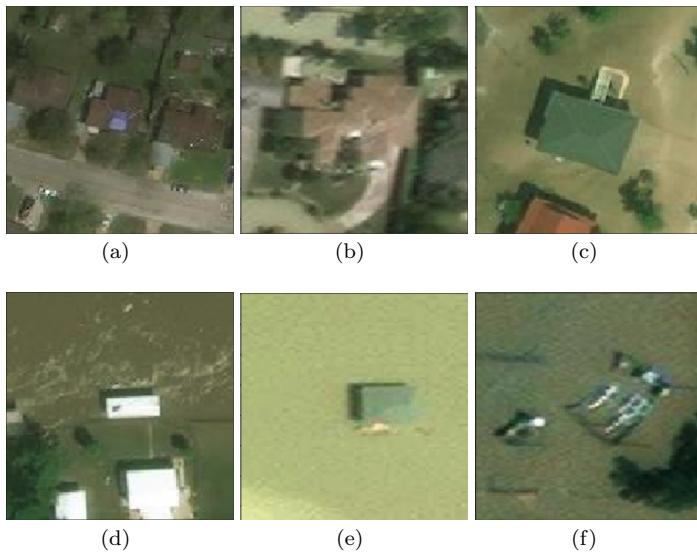


Fig. 13 False negative examples (the label is *Flooded/Damaged*, whereas the prediction is *Undamaged*).

investigate how a model can be made robust against such noise and distortion to reliably annotate damages.

Although the current work extracts the positive and negative samples based on different temporal information of the dataset, it would be more realistic to gather the samples from different spatial information using the same timestamp after the event happens. This would make the data closer to the actual scenario when the method is targeted to be deployed after an event happens. This direction warrants further investigation in future work. We also wish to extend the model to the annotation of road damages and debris, which could help plan effective transportation routes of medical aids, foods, or fuels to hurricane survivors.

Acknowledgements This work was partially supported by the National Science Foundation (NSF grant CMMI-1824681). We would like to thank DigitalGlobe for data sharing through their Open Data Program. We also thank Amy Xu, Aryton Tediarjo, Daniel Colina, Dengxian Yang, Mary Barnes, Nick Monsees, Ty Good, Xiaoyan Peng, Xuejiao Li, Yu-Ting Chen, Zach McCauley, Zechariah Cheung, and Zhanlin Liu in the Disaster Data Science Lab at the University of Washington, Seattle for their help with data collection and processing.

Appendix: dataset and code

The dataset and code used in this paper are available at the first author's Github repository <https://github.com/qcao10/DamageDetection>. The dataset is also available at the IEEE DataPort (DOI: 10.21227/sdad-1e56).

References

1. GeoEye-1 satellite sensor. <https://www.satimagingcorp.com/satellite-sensors/geoeye-1/>, visited 2019-04-27
2. OpenStreetMap. <https://www.openstreetmap.org/>, visited 2019-04-27
3. Tomnod. <http://blog.maxar.com/news-events/2019/in-the-blink-of-an-eye-looking-back-on-nine-years-with-tomnod>, visited 2019-04-27
4. Anatomy of a Catastrophe. <https://www.planet.com/insights/anatomy-of-a-catastrophe/> (2017)
5. Unsupervised flood mapping. <http://gbdxstories.digitalglobe.com/flood-water/> (2017)
6. Ada, M., San, B.T.: Comparison of machine-learning techniques for landslide susceptibility mapping using two-level random sampling (2LRS) in Alakir catchment area, Antalya, Turkey. *Natural Hazards* **90**(1), 237–263 (2018). DOI 10.1007/s11069-017-3043-8. URL <https://doi.org/10.1007/s11069-017-3043-8>
7. Advanced Rapid Imaging and Analysis (ARIA). <https://aria.jpl.nasa.gov/about> (2020)
8. Barnes, C.F., Fritz, H., Yoo, J.: Hurricane disaster assessments with image-driven data mining in high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* **45**(6), 1631–1640 (2007). DOI 10.1109/TGRS.2007.890808
9. Bazi, Y., Melgani, F.: Convolutional SVM networks for object detection in UAV imagery. *IEEE Transactions on Geoscience and Remote Sensing* **56**(6), 3107–3118 (2018). DOI 10.1109/TGRS.2018.2790926
10. Bertero, M., Poggio, T.A., Torre, V.: Ill-posed problems in early vision. *Proceedings of the IEEE* **76**(8), 869–889 (1988). DOI 10.1109/5.5962

11. Cireşan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J.: Flexible, high performance convolutional neural networks for image classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11, pp. 1237–1242. AAAI Press (2011). DOI 10.5591/978-1-57735-516-8/IJCAI11-210. URL <http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-210>
12. Cirean, D., Meier, U., Masci, J., Schmidhuber, J.: A committee of neural networks for traffic sign classification. In: The 2011 International Joint Conference on Neural Networks, pp. 1918–1921 (2011). DOI 10.1109/IJCNN.2011.6033458
13. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization. In: Proceedings of the 16th International Conference on Neural Information Processing Systems, pp. 313–320. MIT Press, Cambridge, MA, USA (2003). URL <http://dl.acm.org/citation.cfm?id=2981345.2981385>
14. Dartmouth Flood Observatory (DFO). <http://floodobservatory.colorado.edu/> (2020)
15. Dodge, S., Karam, L.: Understanding how image quality affects deep neural networks. In: 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1–6 (2016). DOI 10.1109/QoMEX.2016.7498955
16. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics **36**(4), 193–202 (1980). DOI 10.1007/BF00344251. URL <https://doi.org/10.1007/BF00344251>
17. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Y.W. Teh, M. Titterington (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (2010). URL <http://proceedings.mlr.press/v9/glorot10a.html>
18. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: G. Gordon, D. Dunson, M. Dudk (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 15, pp. 315–323. PMLR, Fort Lauderdale, FL, USA (2011). URL <http://proceedings.mlr.press/v15/glorot11a.html>
19. Hong, H., Shahabi, H., Shirzadi, A., Chen, W., Chapi, K., Ahmad, B.B., Roodposhti, M.S., Yari Hesar, A., Tian, Y., Tien Bui, D.: Landslide susceptibility assessment at the Wuning area, China: a comparison between multi-criteria decision making, bivariate statistical and machine learning methods. Natural Hazards (2018). DOI 10.1007/s11069-018-3536-0. URL <https://doi.org/10.1007/s11069-018-3536-0>
20. Huang, F., LeCun, Y.: Large-scale learning with SVM and convolutional nets for generic object categorization. In: Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, vol. 1, pp. 284–291 (2006). DOI 10.1109/CVPR.2006.164
21. Jack, K.: Road inspector using neural network. <https://github.com/jackkwok/neural-road-inspector> (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012). URL <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
24. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. In: Shape, Contour and Grouping in Computer Vision, pp. 319–. Springer-Verlag, London, UK, UK (1999). URL <http://dl.acm.org/citation.cfm?id=646469.691875>
25. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., Chen, M.: Medical image classification with convolutional neural network. In: 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), pp. 844–848 (2014). DOI 10.1109/ICARCV.2014.7064414

26. Liu, Y., Zhong, Y., Qin, Q.: Scene classification based on multiscale convolutional neural network. *IEEE Transactions on Geoscience and Remote Sensing* pp. 1–13 (2018). DOI 10.1109/TGRS.2018.2848473
27. Lu, J., Liu, Y., Zhang, G., Li, B., He, L., Luo, J.: Partition dynamic threshold monitoring technology of wildfires near overhead transmission lines by satellite. *Natural Hazards* **94**(3), 1327–1340 (2018). DOI 10.1007/s11069-018-3479-5. URL <https://doi.org/10.1007/s11069-018-3479-5>
28. Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P.: High-resolution aerial image labeling with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing* **55**(12), 7092–7103 (2017). DOI 10.1109/TGRS.2017.2740362
29. Mehrotra, A., Singh, K.K., Nigam, M.J., Pal, K.: Detection of tsunami-induced changes using generalized improved fuzzy radial basis function neural network. *Natural Hazards* **77**(1), 367–381 (2015). DOI 10.1007/s11069-015-1595-z. URL <https://doi.org/10.1007/s11069-015-1595-z>
30. Ranjbar, H.R., Ardalan, A.A., Dehghani, H., Saradjian, M.R.: Using high-resolution satellite imagery to provide a relief priority map after earthquake. *Natural Hazards* **90**(3), 1087–1113 (2018). DOI 10.1007/s11069-017-3085-y. URL <https://doi.org/10.1007/s11069-017-3085-y>
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Computing Research Repository* **abs/1409.1556** (2014)
32. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014). URL <http://jmlr.org/papers/v15/srivastava14a.html>
33. Xia, G., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X.: AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* **55**(7), 3965–3981 (2017). DOI 10.1109/TGRS.2017.2685945
34. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. *Computing Research Repository* **abs/1505.00853** (2015). URL <http://arxiv.org/abs/1505.00853>
35. Zhang, L., Zhang, L., Du, B.: Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine* **4**(2), 22–40 (2016). DOI 10.1109/MGRS.2016.2540798
36. Zhang, Z., Wang, H., Xu, F., Jin, Y.: Complex-valued convolutional neural network and its application in polarimetric SAR image classification. *IEEE Transactions on Geoscience and Remote Sensing* **55**(12), 7177–7188 (2017). DOI 10.1109/TGRS.2017.2743222