Using Satellite Imagery to Automate Building Damage Assessment: A case study of the xBD dataset

Xiao Chen
Department of Civil and Environmental Engineering
Stanford University, Stanford, CA 94305

ABSTRACT

Future climate change is expected to increase the frequency and severity of natural disasters, posing new challenges for emergency management. For some disasters, a crucial component of rapid response efforts is the accurate assessment of building damage. To tackle this challenge, we use satellite imagery to develop a learning process that automates the building localization and damage classification. As the first step of building segmentation, we perform comprehensive experiments by modifying several state-of-the-art deep learning models. We also propose a new procedure to effectively classify building damage. In a number of test scenarios from different disasters, our approach can match or even outperform existing methods in both segmentation and classification tasks. Such promising results suggest that this approach could have broad-ranging applications in disaster response, vulnerability analysis, and insurance markets.

KEYWORDS
Resilience; Buildings; Disasters; Remote Sensing; Satellite Imagery; Deep Neural Networks;

1. Introduction

Natural disasters such as hurricanes, wildfires, flooding or landslides have become more frequent and severe due to the effects of climate change[1]. One consequence of such disasters is that built structures can be dramatically damaged after such events[2–4]. Allocating resources and reconstructing buildings is typical disaster responses that require fast and accurate assessment of affected areas. Traditional response strategies rely on building sensors detecting vibrations [5, 6] and in-field surveys[4], which are either prohibitively expensive to deploy at large scales or overly time-consuming for processing accurate information. However, with the increased availability of satellite imagery, the task of understanding building damage can be performed in a remote and automated fashion using state-of-the-art computer vision approaches[2, 7–9].

In this paper, we automate the process using satellite imagery to identify the locations of damaged buildings as well as to classify the damage severeness. We show that modern deep learning models can perform well, even for disasters that are newly introduced to our model. In the context of building segmentation in disaster relief and humanitarian assistance, prior work commonly focuses on a single type of disaster[10, 11] or on using customized models with a limited transferable ability to scale up[12, 13]. In contrast, our approach makes use of high resolution images at a relatively large scale, incorporating various disasters and standard deep learning architectures (Section 2.1) to streamline the building segmentation process. We carefully design and calibrate the loss function so that our model yields high performance on precision and recall of the building segmentation (Section 3.1). After segmenting out the buildings, we jointly classify the building damage using components of our segmentation model joined with another deep neural network (Section 2.2). We find our classification module can achieve, and in cases exceed, state-of-the-art performance. For example, our method has an F1 score (defined in Section 5) of 0.63 on average cases; our model outperforms models in existing studies by almost three fold. We finally apply our model on a new, out-of-sample disaster (Section 3.2) and show that it can detect building damage accurately with the F1 score of 0.54 that exceeds the performance of the baseline by two times[2]. These results demonstrate that automated building assessment via satellite image is a promising approach that can be adopted in many applications post-disaster, such as assisting emergency managers in allocating resources or improving the insurance claim process.

2. Methodology

We decouple our approach into two main tasks: image segmentation and damage classification. To predict an accurate building mask, we evaluate three deep learning models with four popular encoders. After identifying building locations, we propose a new learning pipeline to accurately classify the damage of buildings.

Email: markcx@stanford.edu; We give sincere thankfulness to Dr. Ram Rajagopal for his insightful feedback and generous support.

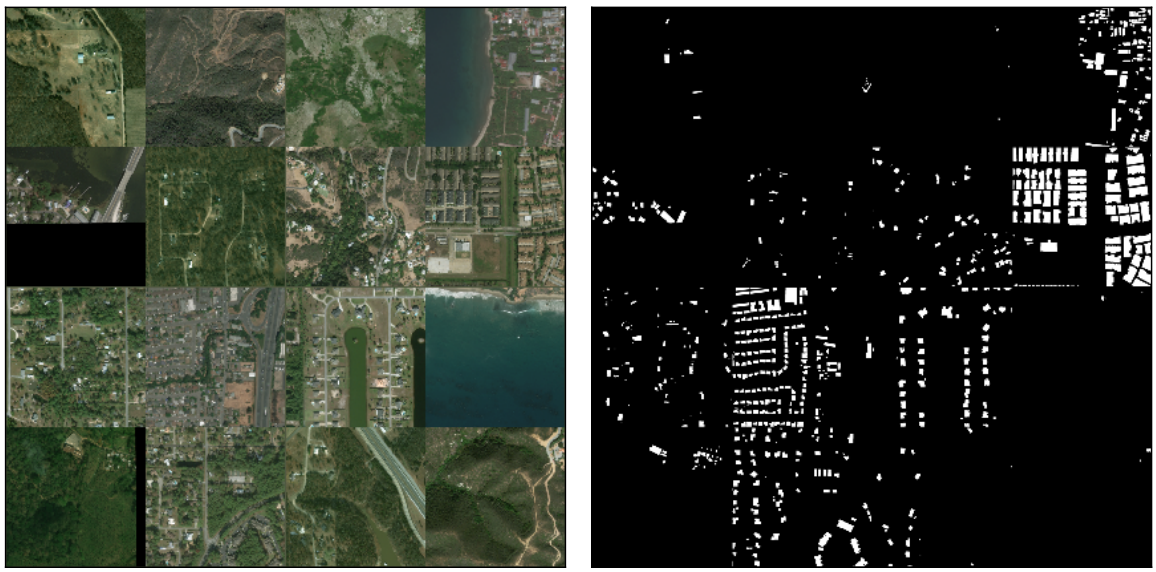


Figure 1. Building localization tasks

Dataset and Setup. We use xBD[2], a newly released open-source dataset that was collected for the building segmentation and damage classification tasks. The dataset consists of 5598 images with $1024 \times 1024 \times 3$ (red, green, and blue channels) resolution spanning across 11 natural disasters, including hurricanes, wildfires, floods, earthquakes, and tsunamis (Tier1 items in Table 1 in [2])[1]. To obtain the ground truth segmentation masks, we use a script[14] to extract the labeled polygons from files and convert them into gray-scale images. We split the dataset of images and labels into training, validation, and testing (holdout) sets at a 80/10/10% ratio. For the classification tasks, we carefully extract buildings according to the labeled polygons in post-disaster images and align them with categorical labels scaled from 0 to 3 as suggested in [2], which yielded 162787 images. Given the post-disaster images, we follow the same 80/10/10% split ratio of training, validation, and testing sets. In our training, the Adam optimizer[15, 16] with weight decay of 0.0005, learning rate of 0.0002, β_1 of 0.9, and β_2 of 0.999 is used in both segmentation and classification models.

2.1. Building Segmentation

To successfully segment buildings, our first challenge is to identify an appropriate model to use in the context of building localization in disasters. Previous work proposed specialized neural networks[10, 12] to detect infrastructure in specific types of disasters. Intuitively, different disasters may have substantially different landscape attributes because of their unique context (e.g., hurricane Harvey in Huston, Texas versus wildfires in Santa Rosa, California). We therefore seek a unified architecture that can perform well across various disaster types and contexts.

We tackle these questions by exploring a number of deep architectures that perform well in image classification and segmentation. We compare three deep model architectures: Unet[17], FPN[18], and PAN[19], all of which to our best knowledge perform well in image segmentation in many applications[10, 20, 21]. To make a fair comparison, we search across four state-of-the-art encoding models (resnet50[22], vgg16_bn[23], se_resnext50_32x4d[24, 25], and densenet121[26]) because these models are widely adopted in the context of classification problems. As shown in Figure 1, we see the images have varied building densities in different locations. To allow the model to zoom into local areas, we randomly crop the raw images down to $512 \times 512 \times 3$. To alleviate the impact of imbalanced pixel distribution (i.e., much fewer positive pixels indicating buildings are located in the images compared with those pixels indicating non-buildings), we evaluate several loss functions (in section 2.1.1) and pick the best one in terms of F1 score. We found that the combo loss performs best among all the options.

2.1.1. Segmentation loss functions

Given that the ground truth building mask Y is Bernoulli distributed such that $P(Y=1)=p$ and $P(Y=0)=1-p$, we have the predicted mask $P(Y=1)=\hat{p}$ and $P(Y=0)=1-\hat{p}$, where p and $\hat{p}$ are the true and predicted probability that a pixel indicates a part of a building. We align $\hat{p}$ with p by minimizing the following loss functions.

Weighted CrossEntropy Loss (WCE). The CrossEntropy loss is defined as $\ell_{CE} = -p\log\hat{p} - (1-p)\log(1-\hat{p})$. Because the imbalance pixel distributions of building vs. non-building. We use the weighted CrossEntropy loss,

[1]The xBD has 22,068 images in total. We use a subset which are well annotated by domain experts.

that is $\ell_{WCE} = -(\gamma_1 p \log \hat{p} + \gamma_0 (1-p) \log(1-\hat{p}))$, where γ_0 and γ_1 are positive numbers. In our training, we set $\gamma_1 = 0.8$ and $\gamma_0 = 0.2$ because positive probability indicating buildings is more important in segmentation.

Focal Loss (Focal)[27]. Focal loss is expressed as $\ell_{Focal} = -\big(\alpha(1-\hat{p})^{\gamma} p \log \hat{p} + (1-\alpha)\hat{p}^{\gamma}(1-p)\log(1-\hat{p})\big)$, where α and γ are hyperparameters. Focal loss generalizes the WCE because when $\gamma = 0$, it is equivalent to WCE when a certain prescribed α. We use $\alpha = 0.5$ and $\gamma = 2$ during the training.

Dice Loss[28]. In the case of binary classification of masks, the dice loss can be described as $1 - \frac{2p\hat{p}+\varepsilon}{p+\hat{p}+\varepsilon}$, where $\varepsilon = 1$ is the hyperparameter.

Tversky Loss[29]. Tversky loss, a variant of Dice loss[28], is $\ell_{Tversky} = 1 - \frac{p\hat{p}+\varepsilon}{p\hat{p}+\beta(1-p)\hat{p}+(1-\beta)p(1-\hat{p})+\varepsilon}$, where $\beta = 0.8$ is the weighting hyperparameter and $\varepsilon = 0.05$ ensures the loss stability in our experiments.

Combinations Loss (Combo). We also evaluate the loss combination of Dice and Focal, i.e., $\ell_{Combo} = \mu_1 \ell_{Dice} + \mu_2 \ell_{Focal}$, where μ_1 and μ_2 are positive hyper-parameters. We test multiple combinations and find $\mu_1 = 1, \mu_2 = 10$ yields the best performance.

2.1.2. Data augmentation and prepossessing

We apply random rotation up to 45 degrees and horizontal flips on both input images and label masks. We also add small Gaussian noise with 0 mean and 0.01 standard deviation on each pixel. Random brightness contrast and blur are incorporated with 0.5 probability. We also use the encoders pre-trained on ImageNet[30] as the initial point. We present the detail results in Table 1 in section 3.

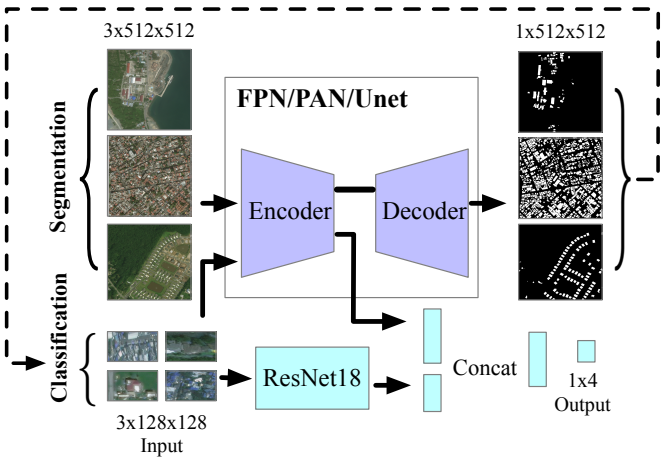
2.2. Damage Classification

To classify the extent of damage of the buildings, we design a novel training pipeline. Specifically, we crop out the buildings in post-disaster images and resize them to $128 \times 128 \times 3$ since many standard neural networks accept these input dimensions. The cropping region depends on whether the predicted mask contains a building or not. (We crop the image even through only few pixels represent buildings). We also use the cropped images according to the ground-truth mask to feed the model. With these down-sized input images, we leverage both the Unet's encoder that learned the weights from the segmentation task and the ResNet18 that is pre-trained on ImageNet[30] to jointly extracting image features. After concatenating the features yielded from two networks, we append a fully-connected layer with the softmax operation to output a one-hot encoded vector where each element represents the probability of an ordinal class. Figure 2a shows our approach of classifying the building damage.

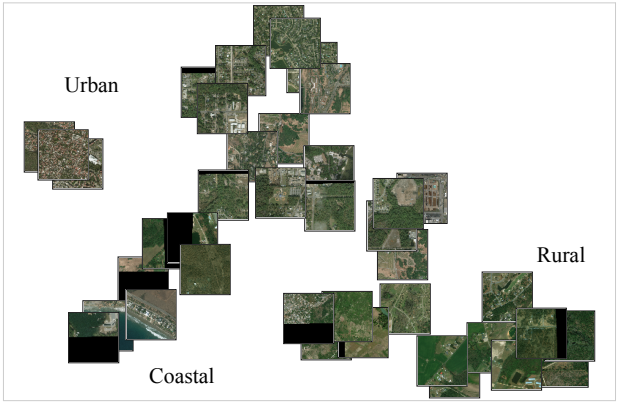
3. Experimental Results

3.1. Localization and Classification

We evaluate segmentation models through several metrics such as *precision*, *recall*, *F1 score*, and *Intersection over Union (IoU)*. The F1 score is the harmonic mean of precision and recall. The results of building localization are presented in Table 1 using Combo loss because it outperforms other loss functions.



(a) Schematic of localizing and classifying damage of the buildings. We pick Unet to segment buildings after comparison. We use the encoder to yield an image embedding concatenated with the embedding generated by resnet18. After passing through the linear layer, we get the 4-dimensional vector representing 4 damage categories.



(b) Embedded view of the images. Although our goal is predicting building masks, we discover that the outputs from encoder of the Unet implicitly separate the different landscape semantics of images. In the two dimensional projection the upper-left has *urban* images, the lower-left has *coastal* images, and the lower-right has *rural* images.

Figure 2. Panel (a) describes the training pipeline of the building segmentation and damage classification task. Panel (b) display the image embedding generated by the learned encoder from segmentation task.

In general, Unet performs better than FPN and PAN because it can yield a 20% performance gain (e.g., Unet+vgg16_bn and PAN+densnet121 in F1 score). Although different encoder structures yield comparable results under the same network architecture, we find that the encoder of vgg16_bn with batch normalization performs best

in the context of building segmentation in terms of F1 score. In addition, the Unet combined with the vgg16_bn encoder achieves the state-of-the-art performance in terms of IoU scores both in background and building detection. An additional benchmark model, such as DeepLab(v3)[31], is also used in our building segmentation since it is widely adopted in other context. But the DeepLab(v3) performs worse than the Unet plus vgg16_bn encoder.

Using the model that performed best from Table 1, we present visual examples to demonstrate the effectiveness of the model qualitatively in Figure 3. We notice that for some regularly shaped buildings, our approach appears to identify the building with a high degree accuracy. Even when the land cover in the background is soil or forest, the building prediction descriptively aligns with the real observations. However, in cases where there is high building density across the whole image, the prediction has a higher, but still small, false positive rate.

We use the encoder in the learned Unet from the segmentation task to perform the feature extraction because it encapsulates the image information in a compressed vector. We apply average pooling at the output layer of the encoder to convert a tensor into batched vectors. We also use spectral normalization[32] on linear layers because such an operation regularizes the model weights and yields robust results. We consider F1 score as the primary metric because it is a widely-adopted metric for the imbalanced data. The classification results are presented in the Table 2, which decouples the metrics down to damage categories. To further understand the effects of the encoder of the Unet, we plot out the raw images on top of the first two main components of the sampled embeddings in Figure 2b. It reveals that different land covers can be semantically segmented out separately.

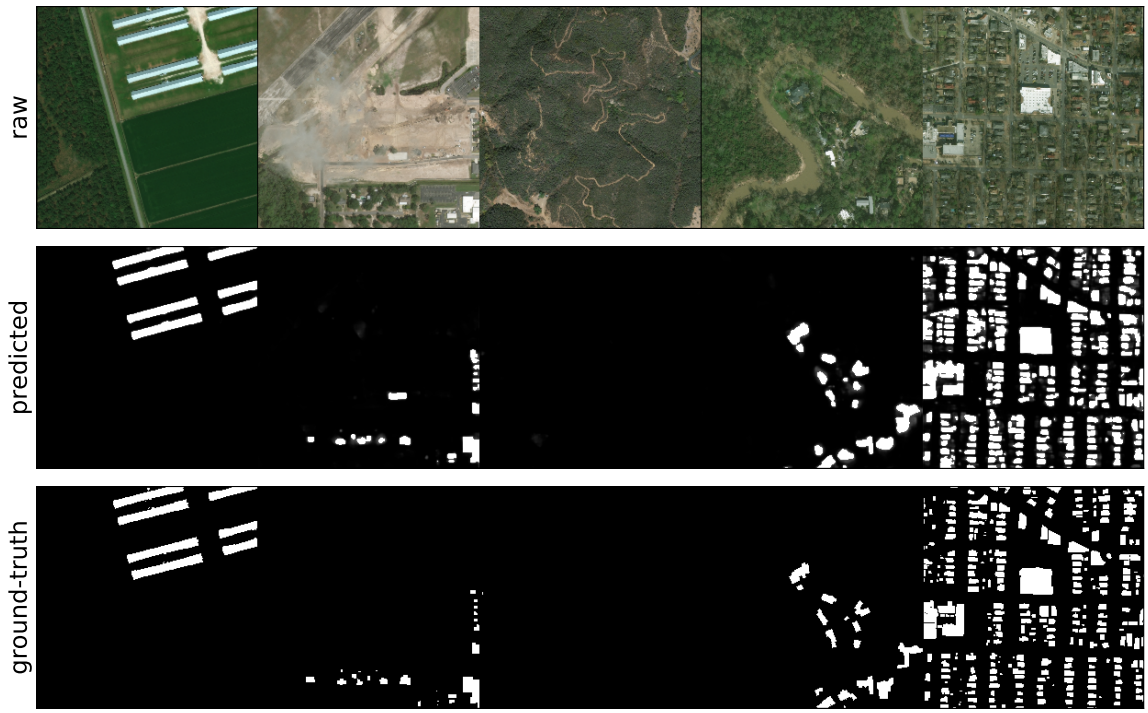


Figure 3. Display the predicted and ground-truth building masks

As shown in Table 2, our approach has averaged 0.63 F1 score which outperforms the baseline F1 score of 0.265 in [2]. The "No Damage" class has high score because there are ample examples of buildings that have been labeled as non-damaged. The minor and major damage have low recall because of the imbalanced data distribution. We also uncovered small discrepancies between those two classes by visually inspecting the images.

3.2. Generalizing new disasters

Ultimately, we are interested in utilizing the deep learning model to provide high-resolution identification of building damage after a future disaster. To this end, we evaluate the generalization capabilities of the model where we attempt to make predictions on data for which the model has not been explicitly trained.

In this experiment, we held out the data of the disaster Hurricane Harvey, and then test building damage in this disaster. The classification results are presented in the following Table 3. We find that the results are lowering-performing compared to the results where we uniformly sample all disasters (Table 2). When Hurricane Harvey is held-out, the overall F1 score, for instance, is 15% lower than that score obtained in the uniform sampling scenario that includes Hurricane Harvey. Yet, such result is expected as the pixel distribution from sample images

Table 1. Building localization performances on test set. Each setting has three runs yielding the mean statistics.

Model	Precision	Recall	F1 score	IoU (background)	IoU (buildings)
Spacenet[2]	✗	✗	✗	**0.97**	0.66
DeepLab(v3)[31]	0.794	0.767	0.780	0.92	0.62
FPN + vgg16_bn	0.818	0.761	0.788	0.91	0.61
FPN + resnet50	0.785	0.724	0.753	0.89	0.55
FPN + se_resnext50_32x4d	0.809	0.742	0.774	0.89	0.57
FPN + densenet121	0.791	0.756	0.773	0.90	0.60
PAN + vgg16_bn	0.754	0.658	0.703	0.81	0.52
PAN + resnet50	0.675	0.690	0.682	0.80	0.50
PAN + se_resnext50_32x4d	0.713	0.672	0.691	0.81	0.49
PAN + densenet121	0.689	0.651	0.669	0.79	0.5
Unet + vgg16_bn†	**0.851**	0.832	**0.841**	**0.97**	**0.68**
Unet + resnet50	0.783	0.795	0.789	0.89	0.65
Unet + se_resnext50_32x4d	0.779	**0.858**	0.816	0.91	0.64
Unet + densenet121	0.785	0.836	0.809	0.92	0.66

†The best performed model in our building segmentation task in terms of F1.

are different between different places. This could make the image embedding features less predictive when we encounter the shifted distribution of pixels. However the F1 score of 0.54 still outperforms the baseline[2].

Table 2. Classifying building damage

Categories	Precision	Recall	F1 Score
No Damage	0.89	0.87	0.88
Minor Damage	0.64	0.51	0.57
Major Damage	0.71	0.43	0.53
Destroyed	0.59	0.52	0.55
Average	0.71	0.58	0.63

Table 3. Classifying damage in hurricane Harvey (Hold-out)

Categories (Harvey)	Precision	Recall	F1 Score
No Damage	0.8	0.72	0.76
Minor Damage	0.57	0.48	0.52
Major Damage	0.62	0.35	0.45
Destroyed	0.48	0.39	0.43
Average	0.62	0.48	0.54

4. Conclusion and Discussion

In the aftermath of a natural disaster, comprehensive data on building damage is difficult to obtain, with such a problem compounded by the severity, extent, and unique characteristics of the event. To address this, we made use of increasingly available images to achieve an accurate assessment of building damage via deep learning. Additionally, our approach can be easily automated and scaled up in a cost-effective way (e.g., without a labor-intensive survey). We find our best experimental result can achieve the state-of-the-art performance[2] shown in Table 1 with a classification outcome having an F1 score of 0.63, almost a three-fold improvement compared to previous studies. Such results are a promising step forward in remote sensing and building detection.

Within the general task of infrastructure mapping, we believe several potential directions could be valuable for the future work. First, with more data and a few additional training techniques such as adversarial learning, classification results can be improved. Second, transfer learning from other datasets, such as spacenet[33], to the building assessment task is potentially a more effective way to create models when learning to associate ground-level features and other observations with satellite imagery. Lastly, being able to generalize the damage classification in new disasters remains a topic to be explored more thoroughly in future research.

5. Appendix

We use the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) to characterize the following metrics specifically.

$$precision = \frac{TP}{TP+FP} \quad (1) \qquad recall = \frac{TP}{TP+FN} \quad (2)$$

$$F_\beta \text{ score} = (1+\beta^2)\frac{precision * recall}{\beta^2 precision + recall} \quad (3) \qquad IoU = \frac{TP}{TP+FP+FN} \quad (4)$$

when $\beta = 1$ it is $F1$ score.

References

[1] Rolnick D, Donti PL, Kaack LH, et al. Tackling climate change with machine learning. arXiv preprint arXiv:190605433. 2019;.

[2] Gupta R, Hosfelt R, Sajeev S, et al. xBD: A dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:191109296. 2019;.

[3] Duarte D, Nex F, Kerle N, et al. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences. 2018;4(2).

[4] Xu JZ, Lu W, Li Z, et al. Building damage detection in satellite imagery using convolutional neural networks. arXiv preprint arXiv:191006444. 2019;.
[5] Rudner TG, Rußwurm M, Fil J, et al. Multi3net: segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. In: Proceedings of the AAAI Conference on Artificial Intelligence; Vol. 33; 2019. p. 702–709.
[6] Balafas K, Kiremidjian AS, Rajagopal R. The wavelet transform as a gaussian process for damage detection. Structural Control and Health Monitoring. 2018;25(2):e2087.
[7] Chen SA, Escay A, Haberland C, et al. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. arXiv preprint arXiv:181205581. 2018;.
[8] Oshri B, Hu A, Adelson P, et al. Infrastructure quality assessment in africa using satellite imagery and deep learning. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 616–625.
[9] Gomes C, Dietterich T, Barrett C, et al. Computational sustainability: Computing for a better world and a sustainable future. Communications of the ACM. 2019;62(9):56–65.
[10] Cao QD, Choe Y. Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. arXiv preprint arXiv:180701688. 2018;.
[11] Cooner AJ, Shao Y, Campbell JB. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 haiti earthquake. Remote Sensing. 2016;8(10):868.
[12] Foulser-Piggott R, Spence R, Eguchi R, et al. Using remote sensing for building damage assessment: Geocan study and validation for 2011 christchurch earthquake. Earthquake Spectra. 2016;32(1):611–631.
[13] Li Y, Hu W, Dong H, et al. Building damage detection from post-event aerial imagery using single shot multibox detector. Applied Sciences. 2019;9(6):1128.
[14] Carnegie Mellon Univ. xview2-baseline [https://github.com/DIUx-xView/xview2-baseline]; 2019.
[15] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.
[16] Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017;.
[17] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention; Springer; 2015. p. 234–241.
[18] Lin TY, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117–2125.
[19] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation. arXiv preprint arXiv:180510180. 2018;.
[20] Roy AG, Navab N, Wachinger C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. IEEE transactions on medical imaging. 2018;38(2):540–549.
[21] Lin TY, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In: European conference on computer vision; Springer; 2014. p. 740–755.
[22] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.
[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132–7141.
[25] Xie S, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1492–1500.
[26] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.
[27] Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2980–2988.
[28] Sudre CH, Li W, Vercauteren T, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2017. p. 240–248.
[29] Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging; Springer; 2017. p. 379–387.
[30] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition; IEEE; 2009. p. 248–255.
[31] Chen LC, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:170605587. 2017;.
[32] Miyato T, Kataoka T, Koyama M, et al. Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations; 2018.
[33] Christie G, Fendley N, Wilson J, et al. Functional map of the world. In: CVPR; 2018.