


Article

Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets

Chuyi Wu ¹, Feng Zhang ^{1,2,*} , Junshi Xia ³, Yichen Xu ¹, Guoqing Li ⁴, Jibo Xie ⁴, Zhenhong Du ² and Renyi Liu ²

¹ School of Earth Sciences, Zhejiang University, 38 Zheda Road, Hangzhou 310027, China; wuchuyi@zju.edu.cn (C.W.); 3170105526@zju.edu.cn (Y.X.)

² Zhejiang Provincial Key Laboratory of Geographic Information Science, Hangzhou 310028, China; duzhenhong@zju.edu.cn (Z.D.); liurenyi@zju.edu.cn (R.L.)

³ Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan; junshi.xia@riken.jp

⁴ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; ligq@aircas.ac.cn (G.L.); xiejb@aircas.ac.cn (J.X.)

* Correspondence: zfcarnation@zju.edu.cn; Tel.: +86-571-8827-3287

Abstract: The building damage status is vital to plan rescue and reconstruction after a disaster and is also hard to detect and judge its level. Most existing studies focus on binary classification, and the attention of the model is distracted. In this study, we proposed a Siamese neural network that can localize and classify damaged buildings at one time. The main parts of this network are a variety of attention U-Nets using different backbones. The attention mechanism enables the network to pay more attention to the effective features and channels, so as to reduce the impact of useless features. We train them using the xBD dataset, which is a large-scale dataset for the advancement of building damage assessment, and compare their result balanced F (F1) scores. The score demonstrates that the performance of SEresNeXt with an attention mechanism gives the best performance among single models, with the F1 score reaching 0.787. To improve the accuracy, we fused the results and got the best overall F1 score of 0.792. To verify the transferability and robustness of the model, we selected the dataset on the Maxar Open Data Program of two recent disasters to investigate the performance. By visual comparison, the results show that our model is robust and transferable.

Keywords: building damage; disaster; remote sensing image; Siamese neural network; U-Net; attention mechanism; change detection



Citation: Wu, C.; Zhang, F.; Xia, J.; Xu, Y.; Li, G.; Xie, J.; Du, Z.; Liu, R. Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets. *Remote Sens.* **2021**, *13*, 905. <https://doi.org/10.3390/rs13050905>

Academic Editor: Chiman Kwan

Received: 21 January 2021

Accepted: 22 February 2021

Published: 28 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the past 20 years, natural disasters have claimed one million lives and caused more trauma, displacement and loss of families and livelihoods [1]. Building damage is the main type of disaster damage, which is used to estimate the location distribution of the affected population [2] and essential for emergency management professionals, helping them direct the rescue teams in a short time to the right locations [2]. It has been proven that remote sensing data are able to derive accurate building damage in a short time [2–4], with low cost and a wide field of view.

Remote sensing (RS) is utilized widely for disaster assessment and the detection of damaged buildings [2,3,5–12]. Frequently used remote sensing images are mainly optical and synthetic aperture radar (SAR) data. SAR data is less affected by weather conditions and has been gradually used for emergency and disaster assessment. Both of the backscatter products and phase data of SAR can be used to detect damaged buildings [8,13]. Compared with optical images, the processing of SAR data is more complicated. Although it is not susceptible to interference from shadows and cloud information, it has many noise, blurry boundaries, no color and less information than multispectral images [14].

Optical remote sensing images can directly reflect the real surface information and are the primary data source in the early stage of remote sensing seismic damage assessments.

Most of the research data that uses optical images for disaster exploration are bitemporal (before and after the disaster) optical images [11]. As known, disasters will cause specific changes to the building, and its characteristic form in the optical remote sensing image often changes significantly. Collapsed buildings become ruins, and the regular geometric shape disappears. The area of the ruins formed is usually larger than the outline area of the original buildings. The disappearance of the geometric shapes causes the shadow of the building to no longer exist, and the texture feature distribution is irregular, or the obvious law is broken. It was found that the spatial features of texture and structure are more critical than spectral information in the algorithm classification [4]. Since a very high resolution (VHR) remote sensing image can provide more detail of the structure and spatial characteristics of the damaged target [15], with the development of RS technology, the spatial resolution of optical images can reach the centimeter level, and the efficiency of image acquisition has also been greatly improved. The detection algorithms of damaged buildings using VHR images include image enhancement, post-classification comparison and machine learning. The image enhancement method of the optical image combined with bitemporal data is to mathematically calculate and compare the pixel value [16]. The post-classification comparison method compares two independent classification results before and after the disaster. Most of them are object-based detection methods. The main advantage of this type of method is minimizing the influence of the radiation difference between two datasets; however, the accuracy depends on the initial classification results [17].

Using VHR images to generate building damage maps, automation and visualization methods are commonly adopted [18]. With technological innovation, numerous scholars have used machine learning to detect damaged buildings and carried out much fruitful work. A. Cooner et al. used a multilayer feedforward network to detect damaged buildings and achieved an error rate of less than 40% [4]. M. Ji et al. optimized the data balance strategy, using the Convolutional Neural Network (CNN) to improve the overall accuracy of detecting collapsed buildings from satellite images [15]. R. Liu et al. proposed an end-to-end network framework combining the CNN and Recurrent Neural Network (RNN) [19]. The CNN is used to extract the spectral spatial features of bitemporal images, while the RNN can effectively analyze the time dependence of bitemporal images and transfer the image features. D. Duarte et al. used residual connections with dilated convolutions that expanded the receptive field; the overall satellite image classification accuracy of the damaged building improved nearly 4% [12]. However, most of these studies only classify buildings into two categories. In order to more accurately assess the damage to buildings after a disaster, some studies have also classified the degree of damage to buildings. E. Weber et al. used Fast-RCNN on the xBD dataset to set the localization and classify at one time to grade the damage of buildings after a disaster, and their overall F1 score reached 0.741 [9].

The U-Net based on a fully convolutional network is considered as one of the standard CNN structures for change detection tasks [20]. The U-Net combines high-level information that can be used to distinguish categories and high-resolution information that provides an accurate positioning and segmentation basis that is suitable for detecting disaster-damaged buildings. Y. Zhan et al. proposed a U-Net-based network that used two-phase SAR images to detect new building structures [21]. S. Ghaffarian et al. used a U-Net with residual connections to detect the statuses of buildings, such as the extinction, addition and expansion, so as to update the building database automatically [22]. Y. Sun et al. proposed a multitask change detection model based on a deep full convolutional network (FCN) to extract the building changes [23]. These studies used many different variants of the U-Net, and all achieved good results; however, they may have a loss of the detailed spatial characteristics. X. Yang et al. proposed the RCNN-U-Net, which can exploit the spatial context and the rich low-level visual features [24]. RCNN was the first to segment the region of interest (ROI) and then perform feature extraction on the region of interest (ROI), but U-Net's feature extraction is for the entire image. In the scene of detecting

damaged buildings, it is undeniable that the global feature of the image contributes to the classification task, because the scale of disaster is larger than that of building. At the same time, according to the first law of geography [25], the part closer to the target building should get more attention when judging the target type. The embodiment of this idea is the attention mechanism. The attention mechanism was proposed in 2014 [26]. It pointed out that human beings selectively focus their attention on each part of the visual space to obtain information at the required time and place and combine the information from different gazes over time to establish an internal representation of the scene. The attention mechanism utilizes this human trait in neural networks. Informally, the attention mechanism provides a neural network with the ability to focus on a subset of its inputs (or features): it selects specific inputs. This helps the model reduce the impact of useless information and increases the contribution of useful information to the results. Attention can be categorized into hard and soft attention, and it is a general idea that does not depend on a specific framework but is currently mainly used in combination with the Encoder-Decoder framework. R. Liu et al. introduced the attention mechanism in change detection and gave weight to features of different times to enhance the changed information of the images, which significantly improved the results [19]. H. Hao et al. proposed the Siam-U-Net-Attn model, which achieved a 0.70 damage F1 score and 0.73 localization F1 score on the xBD dataset [27]. In the task of classification of a hyperspectral remote sensing image, by adding the spectral attention module to the CNN, L. Mou et al. made the model selectively emphasize the useful band and suppress the less useful band [28].

In this paper, we added a soft attention mechanism to Siamese U-Net for exploring the model's performance of different backbones on the degree of building damage. There were three objectives of this study:

1. Explore the performance of multi-artificial neural networks on detecting different damage levels of buildings using both present and post satellite data.
2. Compare the fusion results of different networks and the result of a single network.
3. Evaluate the transferability and robustness of the total model.

For objectives 1 and 2, the xBD dataset [29] is used. As the distribution of four damage level samples are imbalanced, two balancing strategies are adopted, including random under-sampling and a cost-sensitive strategy, which are introduced in Section 2.2.3. For evaluation, the F1 score was used in this study. For objective 3, datasets of Beirut explosion and hurricane Laura are used. The rest of this paper is organized as follows. Section 2 gives the descriptions of all datasets and proposed method. Section 3 presents the experimental results, and the discussion is in Section 4. Finally, the conclusions are drawn in Section 5.

2. Materials and Methods

2.1. Data

2.1.1. xBD Dataset

xBD is a large-scale dataset of building damage assessment used to advance research on humanitarian assistance and disaster recovery. It contains 850,736 building annotations and covers 45,362 km² of images [29]. xBD provides building polygons, labels of damage levels (Figure 1) and satellite images before and after various disaster events.

However, the number of class samples in the xBD dataset is seriously unbalanced. After counting the post-disaster data of the training set (including train and tier3), the proportion of each class in relation to the category of Destroyed Buildings is shown in Table 1. For this, we preprocessed the data before training; the specific operation is in Section 2.2.3.

Damage Level	Structure Description
0 (No Damage)	Undisturbed. No sign of water, structural or shingle damage, or burn marks.
1 (Minor Damage)	Building partially burnt, water surrounding structure, volcanic flow nearby, roof elements missing, or visible cracks
2 (Major Damage)	Partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud.
3 (Destroyed)	Scorched, completely collapsed, partially/completely covered with water/mud, or otherwise no longer present.

Figure 1. Damage levels and their descriptions [29].

Table 1. Proportion of each category.

Class	Proportion
Not Building	539.721
No Damage Building	12.963
Minor Damage Building	1.433
Major Damage Building	1.493
Destroyed Building	1

In the process of adjusting the model, the xBD train and tier3 datasets were used for training. Since we wanted to ensure more training samples and more validation data, this will take up more training time. Taking these factors into account, the dataset was randomly divided into 90% training data and 10% validation data, and the test dataset in xBD was used for verification.

2.1.2. Instance Data

For verifying the transferability and robustness of the model, we selected two disasters out of xBD for applying our method. The reason we chose these two disasters is that the date of occurrence were relatively new, the data was available and they are two different types of disasters. In particular, the explosion in Beirut was further evaluated by The Copernicus Emergency Management Service (CEMS). The buildings affected were divided into four categories, and the classification mapping to xBD was as shown in Table 2. The details of these two disasters and the image data involved are shown in Table 3. The remote sensing image data is provided by the Maxar/DigitalGlobe Open Data Program (<https://www.maxar.com/open-data> (accessed on 24 November 2020)) [30].

Table 2. Level of correspondence between xBD and Copernicus Emergency Management Service (CEMS).

xBD Level	CEMS Level
No Damage	Possible Damage
Minor Damage	Moderate Damage, Possible Damage
Major Damage	Severe Damage, Moderate Damage
Destroyed	Destroyed, Severe Damage

Table 3. The two disasters' detail information.

Disaster	Location	Date	Pre-Image Date	Post-Image Date
Beirut Explosion	Beirut, Lebanon	04/08/2020	09/06/2020	05/08/2020
Hurricane Laura	Parts of Louisiana and far-eastern Texas	27/08/2020	03/06/2019	27/08/2020

The preprocessing of the data is as follows. First, the two images before and after the disaster are geo-referenced. Second, they are cut according to the area of interest. In the selection of the area of interest, we try to avoid the clouds. Third, because the resolutions of the two images before and after the disaster are different, they are resampled to the resolution of the images before the disaster ($0.3\text{ m} \times 0.3\text{ m}$) to ensure the correspondence of the pixel positions in space. Fourth, crop the two images to a size of 1024×1024 .

2.2. Methods

Convolutional neural networks can process data in the form of multiple arrays [31]. For classifying a pixel, the CNN-based segmentation method uses pixel blocks in a fixed size window centered on the pixel as the input of the CNN. This method has several disadvantages. Firstly, the storage space required is large. Secondly, the calculation efficiency is low. Thirdly, the window's size limits the extent of the perceptual field. Usually, the window's size is much smaller than the whole image's. Only some local features can be extracted, which leads to the limited performance of classification. In order to overcome the above shortcomings, a full convolutional network (FCN) come into being. The FCN is a special type of CNN and can recover the category of each pixel from the abstract features. That is, it extends from image-level to pixel-level classification. In this paper, U-Net [32] is used as a kind of FCN. Essentially, convolution is feature fusion of a local area that fuses features from spatial dimensions and channel dimensions. For a convolutional neural network, its core calculation is a convolution operator that learns a new feature map from the input feature map through the convolution kernel. Different backbones are used to try to strengthen our model. All backbones used are the residual network and its variants.

2.2.1. Proposed Framework

This section introduces the details of the overall framework. As shown in Figure 2, the architecture of the proposed model is a Siamese neural network that is divided into two parts, and both share the same weight. One part with the pre-disaster images is used to localize buildings, and the other part is used to classify the buildings' damage levels. While training, the localization part is trained first and as the pre-training weight of two parts. This step is marked as ① in Figure 2. Then, the image pairs contained pre- and post-disaster are augmented and input into the network for training simultaneously. This step is marked as ② in Figure 2. Finally, we get an end-to-end damaged building detection network. This is an example of transfer learning. The training model of one problem can be reused as the initialization of another model of a similar problem [33]. For further improving the accuracy, we process the results of the end-to-end network data, use the building mask generated by the building localization network and remove some nonbuilding pixels in the result. This step is marked as ③ in Figure 2. The whole process is shown in Figure 2.

2.2.2. Attention U-Net

The attention mechanism means that, when selecting information, it calculates the weighted average of the N input information and then passes it on to the next block. The decoder part of our U-Net adds the attention mechanism. The specific architecture is shown in Figure 3.

In the decoder part of U-Net in Figure 3, the previous decoder layer's output is originally directly spliced with the output of the corresponding encoder layer as the input of the next decoder layer. After adding the attention block, the input will be processed by the attention gate, which is shown in Figure 4, and then enter into the next decoder layer to express the spatial attention.

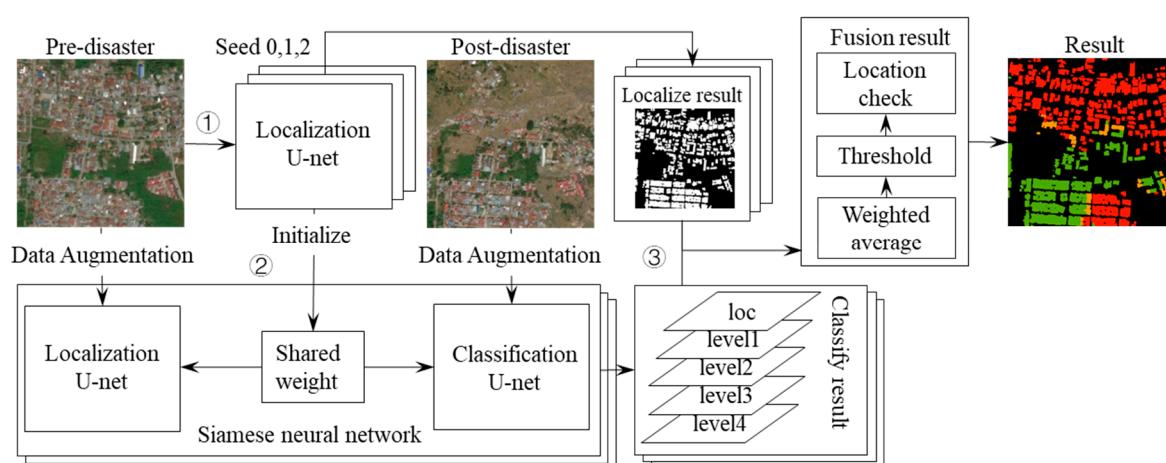


Figure 2. The overall framework. Localization U-Net is used to locate buildings. After pre-training with pre-disaster image, it shares weight with classification U-Net, which is used to classify damage level. The combination of the two is Siamese neural network. We use three random seeds to train it. After inputting pre- and post- disaster images into Siamese neural network, we get three five channels classification results. The first channel is the result of localization, and the last four channels are the probability of each damage level at each location. In the fusion step, the results corresponding to the three seeds are first weighted and averaged, and the weight is determined by the validation accuracy during the training process. Then, the threshold is used to determine the value of each pixel. Finally, for improving the accuracy, the pre-trained classification U-Net localizes the buildings again and get the localize result which be used for the double check of buildings' localization.

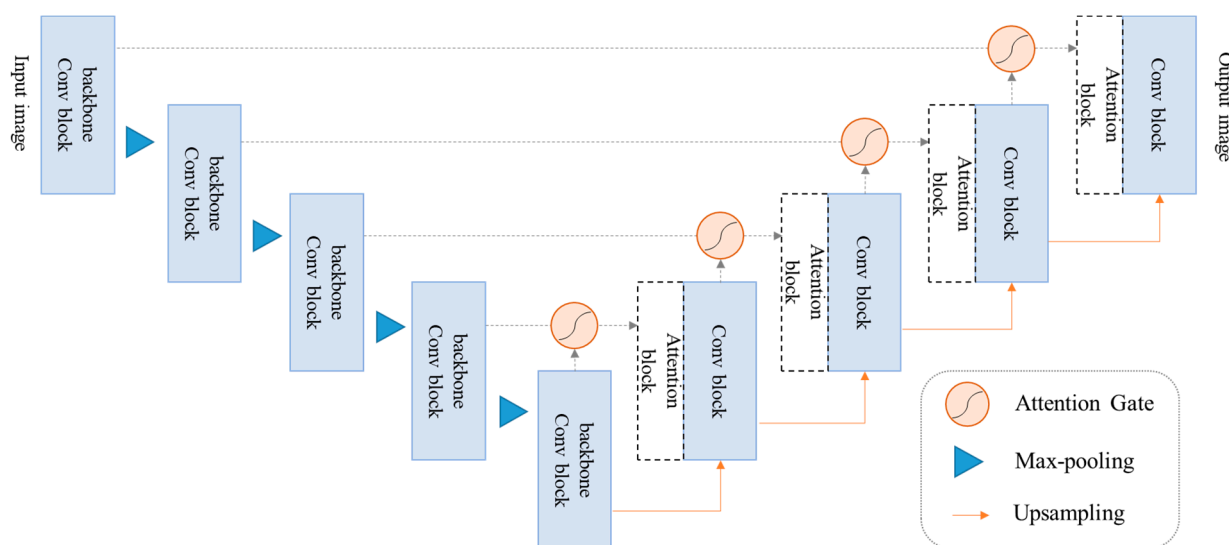


Figure 3. U-Net with the attention mechanism's structure [34]. The encoder performs 4 downsampling. Symmetrically, its decoder upsamples 4 times to restore the features to the original image resolution. The attention gate is placed at the end of the skip connection.

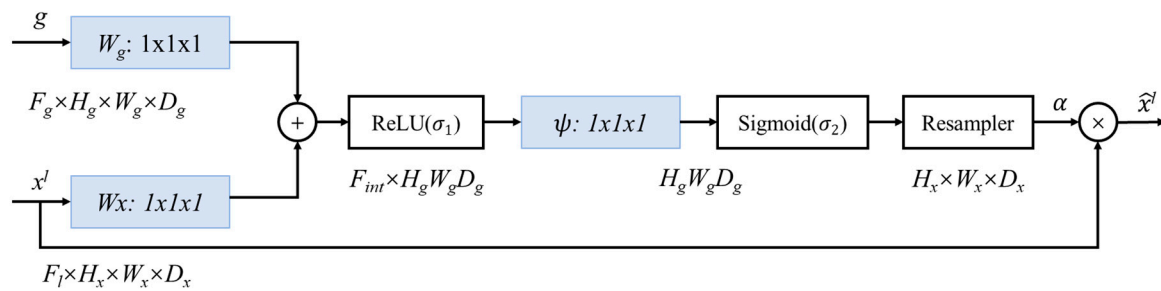


Figure 4. Attention gate structure [34]. F , H and W stand for channel, height and width respectively, and D is the depth of the 3D data block. x^l , the feature map from the encoder layer, is scaled with the attention coefficients (α), which are computed by x^l and g . The previous decoder features in g are added to x^l to determine the focus regions; then, the value of the attention coefficients is between 0 and 1 throughout training.

2.2.3. Data Augmentation

In machine learning algorithms, the ideal situation is that the number of samples in each class is roughly the same. However, in most real scenes, the category distributions are uneven [35]. This research uses a variety of data balancing strategies. The first is the under-sampling strategy. The 1024×1024 image is randomly cropped multiple times, and the crop size is 512×512 . The cropping scheme with the largest sum of pixel values is selected to reduce the sampling frequency of non-buildings. Secondly, a cost-sensitive strategy is adopted. When constructing the loss function, the loss of each category and the total loss are combined, and at the same time, they are given different weights by referring to the proportion of the categories.

In addition, for enhancing the robustness of the model, we also randomly flip, rotate, translate, side view and zoom on the input images; adjust their saturation, contrast and brightness; convert the color space and band order and add Gaussian noise and filtering operations randomly.

2.2.4. Backbones

- ResNet-34 backbone

The first backbone used is ResNet-34 [36], belonging to the residual network (ResNet) pre-trained on the ImageNet [37] dataset. The ability of CNN to retrieve relevant information from images is enhanced with the increase of the network depth [38]. However, if the network is too deep, it will lead to gradient explosion and network degradation. Residual connections [36] solved this problem by feeding a given layer into the previous one. Figure 5 is the structure of a residual connection.

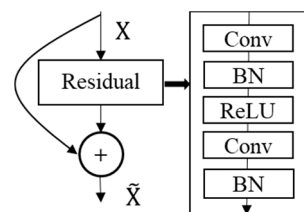


Figure 5. Residual connection. Each residual block is composed by convolutions (Conv), batch normalizations (BN) and rectified linear units (ReLU).

- Squeeze-and-Excitation Networks (SENet) backbone

For convolution operations, a large part of the work is to improve the receptive field—that is, to fuse more features spatially or to extract multi-scale spatial information, such as the multi-branch structure of the Inception network [39]. For the feature fusion of the channel dimensions, the convolution operation basically defaults to fusing all channels of the input feature map. The Group Convolution and Depth-wise Separable Convolution

in the MobileNet network group channels mainly make the model more lightweight and reduce the amount of calculation. The innovation of the SENet network is to focus on the relationship between channels, hoping that the model can automatically learn the importance of different channel features; the SENet can be regarded as the channel-wise attention mechanism. To this end, SENet proposes the Squeeze-and-Excitation (SE) module, as shown in Figure 6.

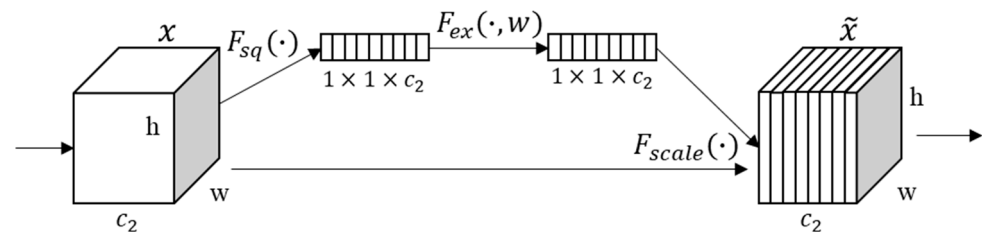


Figure 6. Squeeze-and-Excitation (SE) module [39]. The module is mainly composed of three parts: squeeze, excitation and scale. $F_{sq}(\cdot)$ represents the squeeze transformation, $F_{ex}(\cdot, w)$ represents the excitation transformation and $F_{scale}(\cdot)$ represents the scale transformation.

The core module of SENet is divided into three parts: squeeze, excitation and scale. The squeeze part is used to compress features to $1 \times 1 \times \text{channels}$ in the spatial dimension, which represents the channels' global distribution. The excitation part is reassembled at the gating mechanism, which produces channel-wise weights $W \in \mathbb{R}^{c_2 \times c_2}$. The scale part uses the learned weights to reweigh the importance of each channel to build attention on the channel.

- SEResNeXt backbone

SEResNeXt is a model obtained by applying the SE module to the residual block in ResNeXt. In fact, in the block of ResNet, one residual path becomes multiple residual paths. The success of the Visual Geometry Group Net (VGGNet) [40] and ResNet shows that the method of stacking blocks of the same shape can reduce the number of hyperparameters and achieve state-of-the-art (SOTA) results. The practice represented by GoogleNet and Inception also shows that a fine network design through the split–transform–merge strategy can also achieve very good results. ResNeXt's idea is to combine these two good ideas. ResNeXt does not perform split–transform–merge like the GoogleNet series but simply repeats the same substructure, as shown in Figure 7, so that the split–transform–merge is done; at the same time, there is not much increase in the hyperparameters.

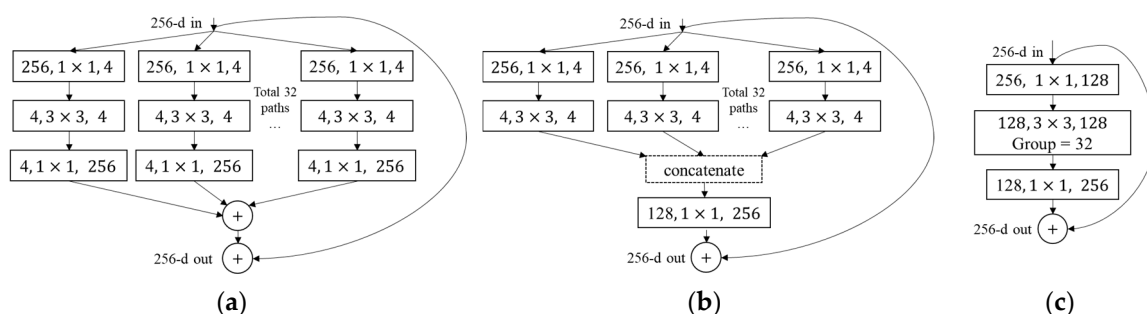


Figure 7. ResNeXt structure [41]. In the figure above, the structure of (a) is the original structure of ResNeXt, and (b,c) are equivalent representations of the structure of (a) in an actual implementation, the structure of (c) which is relatively simple to implement, and the basic block of ResNeXt is realized through the form of grouped convolution.

SEResNeXt is obtained by adding the SE module to the residual block in ResNeXt. The structure of a single residual block combined with the SE module is shown in Figure 8.

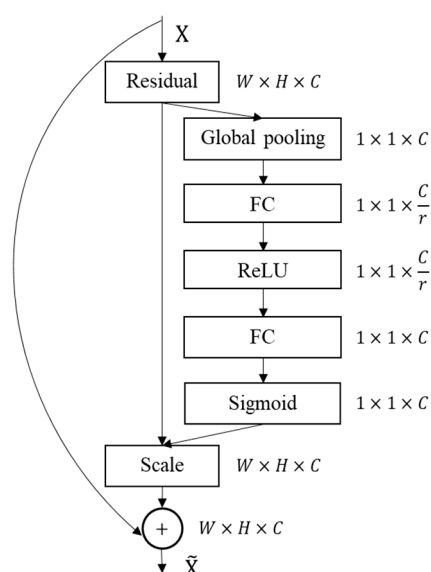


Figure 8. Residual block combined with the SE module. The SE module is composed of the residual block, which was introduced in Figure 5, global pooling block, full connection (FC) block, rectified linear units (ReLU) and activation block (Sigmoid).

- Dual Path Net (DPN) backbone

DPN is a new convolutional network structure that combines the advantages of the ResNet and Dense Convolutional Network (DenseNet). By revealing the equivalence of the ResNet and the Dense Convolutional Network (DenseNet), the author found that ResNet supports element reuse, while DenseNet supports new element exploration. In order to integrate the benefits of these two path topologies, DPN aggregates the functions of the two.

In Figure 9, the phase results of the Dense Net on the left and ResNeXt on the right are added together. The added result is then processed by 3×3 convolution and 1×1 dimension transformation operations; finally, its channels are divided into two parts. The left part is merged with the original input on the left, and the right part is added with the original input on the right. The operation, in this way, is a block formed in which the original input can be the input that entered the network at the beginning or the input of the previous stage.

2.2.5. Fusion

Each model uses three different random seeds for training, so each network has three training weights and its optimal overall F1 score while training. After using the three training weights to predict the verification set, three prediction results can be obtained. Then, the optimal overall F1 score is used as the weight of each result for the fusion based on the weighted average. Finally, the preliminary results of localization and classification are obtained. When fusing the results of different models, the same method is used, except that the weighted results are changed from 3 to 12. Since the localization task network is a single target network, it is more targeted than the classification network, which has two targets in the localization task. We used the localization result to mask the classification result and removed the non-building pixels to get the results. The weighted average method is also used for the fusion of different networks with or without attention.

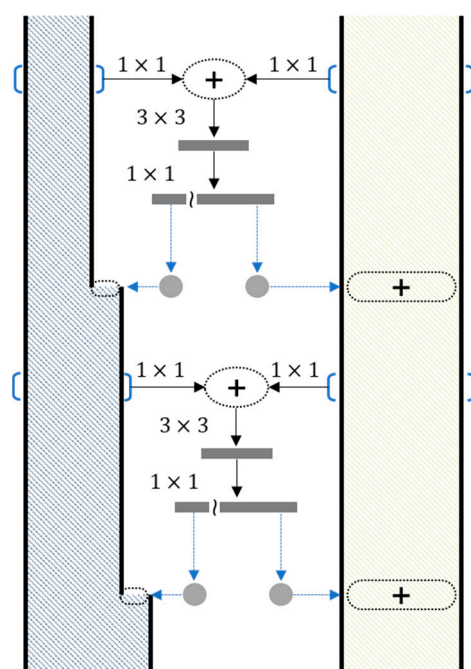


Figure 9. Dual Path Net (DPN) structure diagram [42].

2.3. Metric

In a classification task, a confusion matrix is frequently used to evaluate the accuracy of the information and the performance of a model [43], and accuracy indicators such as precision and recall provide a summary of the information in it. Each row of the confusion matrix represents a prediction category. Each column represents an actual category to which the pixel belongs. TP (True Positive) is the number of pixels that are correctly predicted as this category. FP (False Positive) is when the number of pixels that belong to other categories are wrongly classified as this category. FN (False Negative) is when the number of pixels that belong to this category are mistakenly classified as another one. TN (True Negative) is the number of pixels that are correctly classified as other categories.

The measure of accuracy using the portion of TP and TN does not distinguish between different categories; thus, the overall performance of a multi-class model is not well-described when dealing with an unbalanced dataset. By contrast, the measures of precision and recall reflect the true classification performance, and the F1 score is balanced between the two indicators.

$$Precision = \frac{TP}{TP + FP}, \quad (1)$$

$$Recall = \frac{TP}{TP + FN}, \quad (2)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \quad (3)$$

The evaluation metrics we used is the F1 score calculated by a weighted average of the localization f1 score (lf1) and the damage f1 score (df1), which was provided by xView2 Challenge [29].

$$Overall\ F1 = 0.3 * lf1 + 0.7 * df1, \quad (4)$$

The localization f1 score is the normal F1 score, which is the harmonic mean of precision and recall [44] used to assess the effectiveness of building identification, which is a binary classification task.

$$lf1 = F1. \quad (5)$$

Our model classifies pixels into four labels, so $df1$ is a score of multi-class F measures. The macro-averaged F1 score is a popular performance score that is computed by averaging the per-category scores [45]. It adapts to a large-scale dataset. As the global arithmetic mean of each, it does not adequately represent the performance of the classifier in each category. Our damage F1 score is calculated by taking the harmonic mean of the 4 f1 scores calculated for each damage level [29]. It behaves differently compared to the macro F1 score, as it gives a larger weight to lower numbers.

$$df1 = 4 / \sum_{f1}^{[no_damage_{f1}, minor_damage_{f1}, major_damage_{f1}, destroyed_{f1}]} \frac{1}{(f1 + \epsilon)}. \quad (6)$$

For reference, we also used the Mean Intersection over Union (MIoU) [46] to evaluate the performance of the model.

$$MIoU = \frac{TP}{TP + FP + FN}. \quad (7)$$

$$Localization\ MIoU = MIoU \quad (8)$$

$$\begin{aligned} & Classification\ MIoU \\ &= 4 / \sum_{MIoU}^{[no_damage_{MIoU}, minor_damage_{MIoU}, major_damage_{MIoU}, destroyed\ MIoU]} \frac{1}{(MIoU + \epsilon)} \end{aligned} \quad (9)$$

2.4. Training Implementation

Considering both the resources and efficiency, Adam [47] with a learning rate of 0.0002 is chosen as the optimization algorithm, which has strong robustness in the selection of super parameters. While training the localization task, the training data batch size is set to 16, and 100 epochs are trained on the network. When two tasks are trained at the same time, the training data batch size is set to 10, and the network is trained for 24 epochs. The implementation of the framework network is based on pytorch [48], and two NVIDIA GTX 1080ti GPUs with 8G memory are used for training and verifying. We used the weights of networks trained with ImageNet provided by pytorch to initialize the network.

3. Results

3.1. Compare Models

We trained a total of two groups of U-Net models with eight different backbones—namely, the group with the attention mechanism and the group without. The different backbones used were introduced in Section 2.2.4. Under the premise that the data is divided into a training set and validation set using random seeds, each model used three different random seeds for training. As shown in Tables 4 and 5, the results of each model of the verification set are shown. The classify result index uses the overall F1 score introduced in Section 2.3, the localization result index uses the ordinary F1 score and the overall index is a 0.3 localization F1 score and 0.7 classify F1 score.

Table 4. Indicators of different models without the attention (w/o A) mechanism. MIoU: Mean Intersection over Union.

Backbone Type (BT)	Overall F1 (OF1)	LocalizationF1 (LF1)	ClassificationF1 (CF1)	No Damage	Minor Damage	Major Damage	Destroyed	Localization MIoU (LMIoU)	Classification MIoU (CMIoU)
resNet	0.636	0.856	0.541	0.863	0.351	0.474	0.789	0.748	0.371
SEresNeXt	0.755	0.860	0.710	0.916	0.502	0.741	0.834	0.754	0.551
DPN	0.739	0.735	0.741	0.920	0.553	0.742	0.865	0.581	0.589
SENet	0.772	0.863	0.734	0.912	0.544	0.741	0.857	0.759	0.579

Table 5. Indicators of different models with the attention (w A) mechanism.

BT	OF1	LF1	CF1	No Damage	Minor Damage	Major Damage	Destroyed	LMIoU	CMIoU
resNet	0.744	0.856	0.696	0.880	0.501	0.719	0.818	0.748	0.533
SEresNeXt	0.787	0.868	0.752	0.920	0.583	0.750	0.847	0.767	0.603
DPN	0.781	0.870	0.742	0.919	0.553	0.759	0.852	0.769	0.590
SENet	0.779	0.859	0.745	0.903	0.569	0.751	0.853	0.753	0.594

As can be seen from Tables 4 and 5, SENet and SEresNeXt both show better overall performances without and with the attention. For the classification task, their performances with the attention mechanism are better than those without the attention mechanism. In the task of building localization, DPN with the attention mechanism shows the best performance, reaching an F1 value of 0.870. Observation shows that whether the attention mechanism is added has no uniform impact on the localization accuracy, but it will improve the accuracy of the classification. In order to further observe the results of each model, we selected three sample images in the verification set to compare the results, as shown in Figures 10–12.

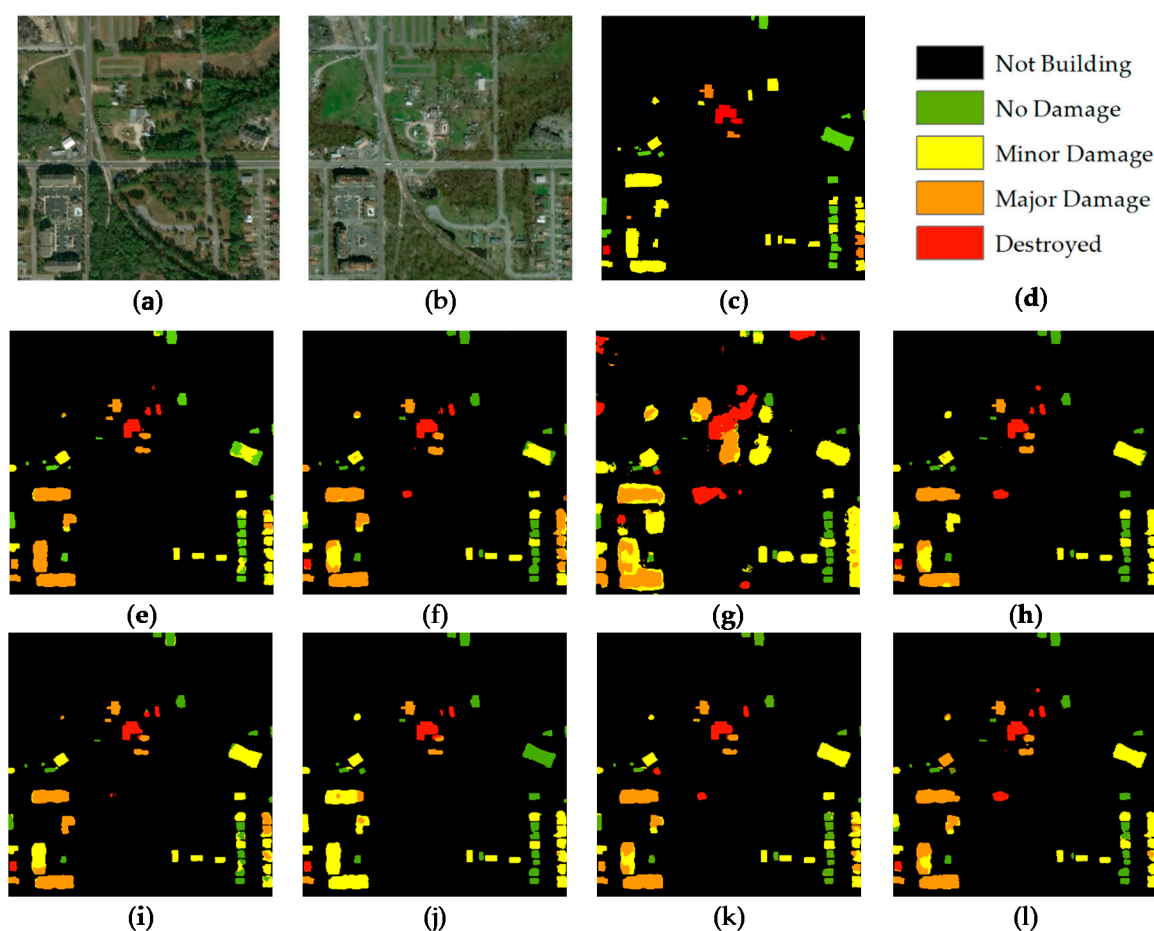


Figure 10. Hurricane-michael_00000256 in the verification set. (a) A pre-disaster image, (b) a post-disaster image, (c) the ground truth (GT) and (d) the legend. (e) The result of ResNet (w/o A or without attention), (f) the result of SEresNeXt (w/o A), (g) the result of DPN (w/o A), (h) the result of the Squeeze-and-Excitation network (SENet) (w/o A), (i) the result of ResNet (w A or with attention), (j) the result of SEresNeXt (w A), (k) the result of DPN (w A) and (l) the result of SENet (w A).

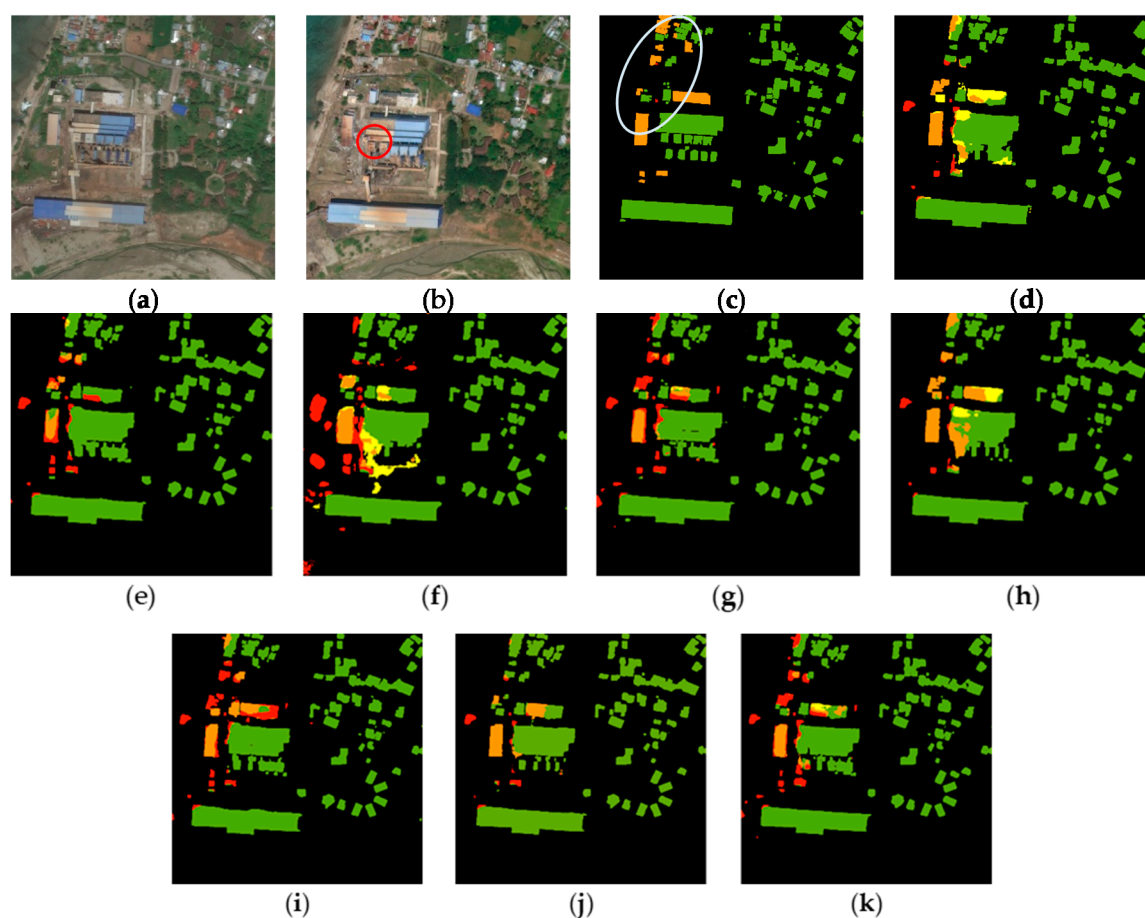


Figure 11. Palu_00000004 in the verification set. (a) A pre-disaster image, (b) a post-disaster image and (c) the ground truth (GT). (d) The result of ResNet (w/o A), (e) the result of SResNeXt (w/o A), (f) the result of DPN (w/o A), (g) the result of SENet (w/o A), (h) the result of ResNet (w A), (i) the result of SResNeXt (w A), (j) the result of DPN (w A) and (k) the result of SENet (w A).

Figure 10 was chosen to observe the discrimination of minor-damaged buildings. It can be seen that the model has a low detection accuracy for minor damage. The minor damage building in the lower left corner of Figure 10 is classified as major damage in most models; only in Figure 10j is it judged as minor damage. In Figure 10a,b, the observation directions are different, so the location of buildings in the two images cannot be completely overlapped, which may cause an error of judgment. The error is also related to the degree of damage that is continuous but is artificially divided into discrete levels. The appearance of minor-damaged buildings is not obvious and diverse. For example, the damaged parts are on the sides of the buildings, which cannot be observed by remote sensing images.

Figure 11 is used to observe the discrimination of nondamaged and major-damaged buildings. It can be seen that the performance of each model in the detection of nondamaged buildings is stable, but when distinguishing major-damaged buildings, it is easy to make misjudgments. For the buildings circled in Figure 11c, there are misjudged pixels in nearby buildings, which can be seen in every model's results. As can be seen from Figure 11b, the part of the building in the red circle has changed in texture and color compared to Figure 11a. Its appearance may have changed due to various factors, but the building itself does not reach the level of minor damage. This is related to the limitation of optical image data, which is easily inferred by the color information, and false changes are detected. Compared with the result without the attention mechanism, the result with the attention mechanism is more accurate in the detection of major-damaged buildings.

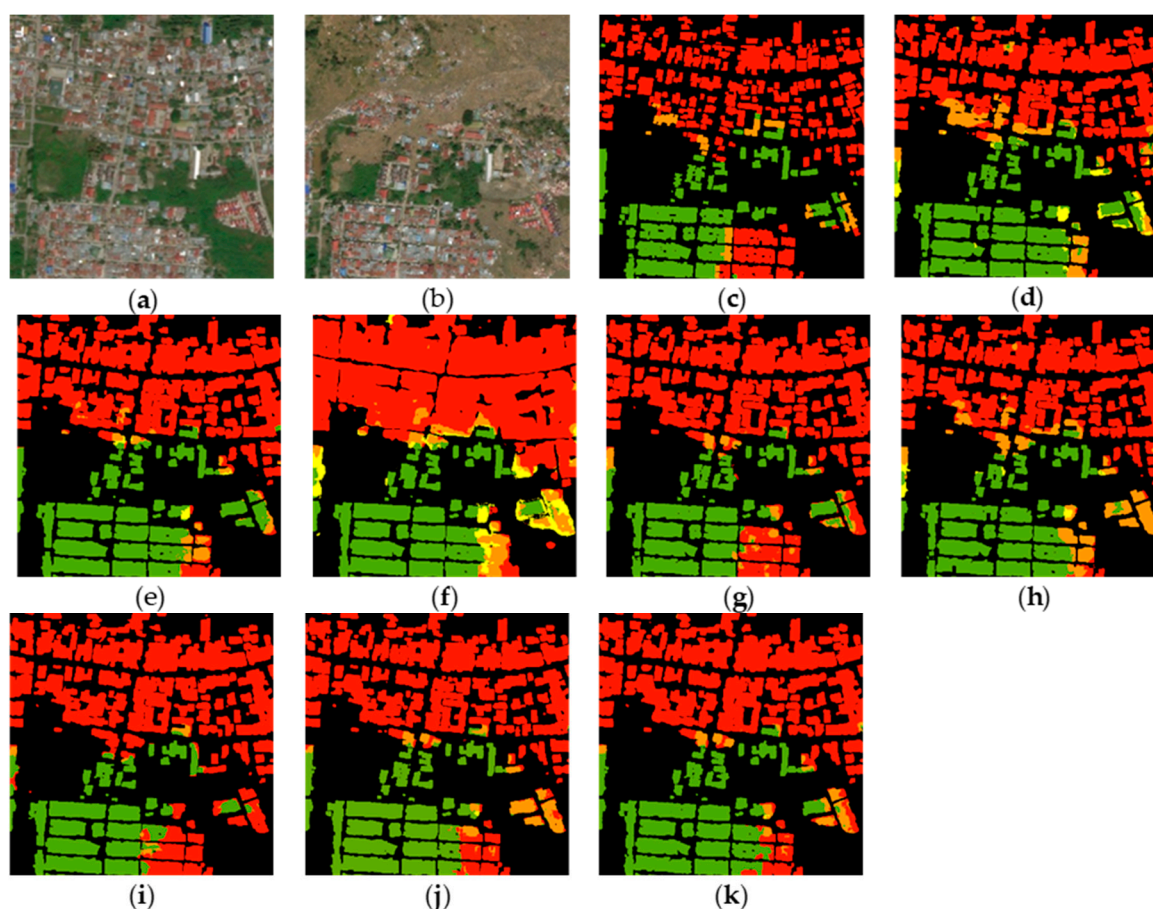


Figure 12. Palu_00000181 in the verification set. (a) A pre-disaster image, (b) a post-disaster image and (c) the GT. (d) The result of ResNet (w/o A), (e) the result of SResNeXt (w/o A), (f) the result of DPN (w/o A), (g) the result of SENet (w/o A), (h) the result of ResNet (w A), (i) the result of SResNeXt (w A), (j) the result of DPN (w A) and (k) the result of SENet (w A).

Figure 12 is used to observe the judgment of destroyed buildings. It can be seen that the location and classification of the damaged area for each model are basically the same, but the details are different. For example, all models have detected that the large area above the image is destroyed, but the detection results of the state of the lower right house are inconsistent. Compared with the ground truth, Figure 12g,i shows a better performance, which is the results of SENet (w/o A) and SResNeXt (w A). As shown in Figure 12f,h, DPN (w/o A) and resNet (w A) perform poorly. In the localization task, except for the poor performance in Figure 12f, which is the result of DPN (w/o A), the classification effects of other models are similar. Among them, the boundary is the clearest, and the least sticky is Figure 12h, which is the result of ResNet (w A).

3.2. Fusion Results

In pursuit of high precision, we try to explore whether the fusion results of different networks would be more accurate than the result of a single network. For this reason, we divide the models into two groups according to whether we added attention and integrate the building localization and classification results of four networks. In the fusion process, the four network contribution weight ratios are 1:1:1:1, and the fusion results are shown in Figure 13 and Table 6.

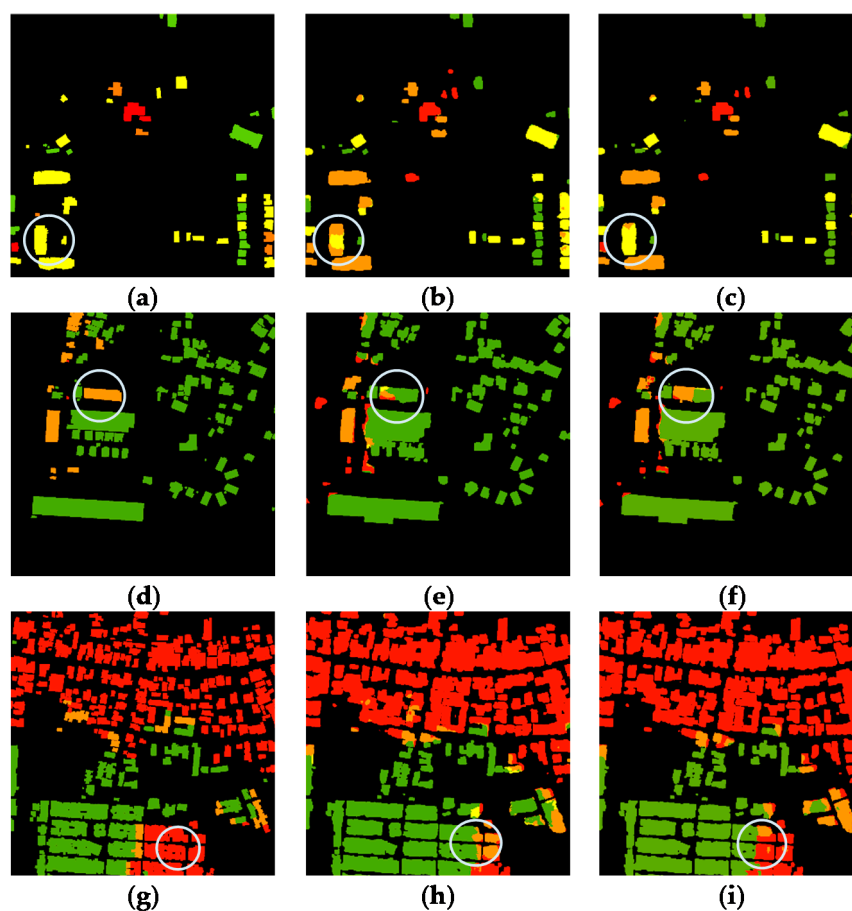


Figure 13. The fusion result of hurricane-michael_00000256, palu_000000004 and palu_00000181. (a) The GT of hurricane-michael_00000256, (b,c) the results of the model w/o A and model w A of hurricane-michael_00000256, (d) the GT of palu_000000004, (e,f) the results of the model w/o A and model w A of palu_000000004, (g) the GT of palu_00000181 and (h,i) the results of the model w/o A and model w A of palu_00000181.

Table 6. Fusion result indicators.

Group	OF1	LF1	CF1	No Damage	Minor Damage	Major Damage	Destroyed	LMIoU	CMIoU
Without Attention	0.769	0.862	0.728	0.914	0.527	0.751	0.855	0.758	0.573
With Attention	0.792	0.871	0.758	0.916	0.582	0.768	0.859	0.771	0.611

Comparing Figure 13 and Table 6, the fusion results of the model with the attention mechanism had an increase in the overall accuracy compared with the results of the single model, which increased by 0.005 compared with the highest overall accuracy of the single model. The result of fusion without the attention mechanism is lower than the highest overall accuracy of single model but only 0.003 lower than SENet. Comparing the accuracy of the two groups, the group with the attention mechanism was higher, indicating that the attention mechanism had a certain effect on improving the overall accuracy of the model. Comparing Figure 13b,c,e,f,h,i, respectively, the fusion results showed a stable classification result. However, they had some differences in the details. As in the circled part in Figure 13, the fusion results of the model with the attention mechanism detected more accurate damage areas than the results of the model without the attention mechanism. In terms of localization, the fusion results of the model the with attention mechanism had clearer boundaries and fewer adhesions between multiple buildings than without

the attention mechanism. In summary, the fusion of multiple models with the attention mechanism is beneficial to the improvement of the accuracy.

3.3. Transferability and Robustness

For verifying the robustness and transferability of model, two disasters not in the training and verification set were selected. One was the explosion accident in Beirut, Lebanon on 4 August 2020, and the other was Hurricane Laura, which occurred on 27 August 2020; see Section 2.1.2 for the specific introduction. We selected one region of interest for the explosion in Beirut and one for Hurricane Laura. The pre- and post-disaster images of each disaster were input into our model, and the results are shown in Figures 14 and 15.

In the case of the Beirut explosion, for the classify task, the results in Figure 14 are more accurate in detecting the destroyed area. However, in the upper left part of all results, a piece was detected as nondamaged or major-damaged. The original satellite image Figure 14a shows that there is a shed here, and the roof still exists after the explosion in Figure 14b, but the wall under it may have collapsed completely, which makes it impossible to detect this situation from a top view. These situations generally occur when the building's upper layer collapses directly to the bottom floor [49]. This shows that the optical satellite image has limitations in building damage detection. Compared with the ground truth in Figure 14m, almost all the results do not accurately classify the disaster damage grade of pixels in the lower right corner, and they tend to overestimate it.

In the localization task, the dense buildings in the lower right corner of the image are not recognized. This part of the building is high, showing the effect of side shooting in the image, and there are large areas of shadows, which limits the model's recognition of the buildings. In the red circle of Figure 14a, there are ships docked at the port, which are identified as buildings in Figure 14k with no attention mechanism. But in Figure 14l with the attention mechanism, it is excluded from the building localization. This proves that the attention mechanism is helpful for localization tasks to distinguish between ships and buildings.

Since there is no public official ground truth of Hurricane Laura, we manually label the ground truth (seen in Figure 14m) following the rules of xBD. Compared to Beirut's explosion, the image of Hurricane Laura is covered by thin clouds, and the texture is more blurred. Unlike the Beirut port, the scale of the buildings is smaller. From the results of Hurricane Laura in Figure 15, the difference between the fusion results with and without the attention mechanism is very small, and both models present similar interpretation levels for each house. By comparing all the results, there is not much difference in the building localization tasks. Some tiny buildings or buildings covered by trees are likely to be missed. As shown in Figures 14 and 15 for determining the level of damage, compared with the existing disasters in the training set, the performances of the two disasters' classification tasks are poor. However, it is basically possible to distinguish between damaged and nondamaged buildings, and the error in determining the level of damage is mostly within one level.

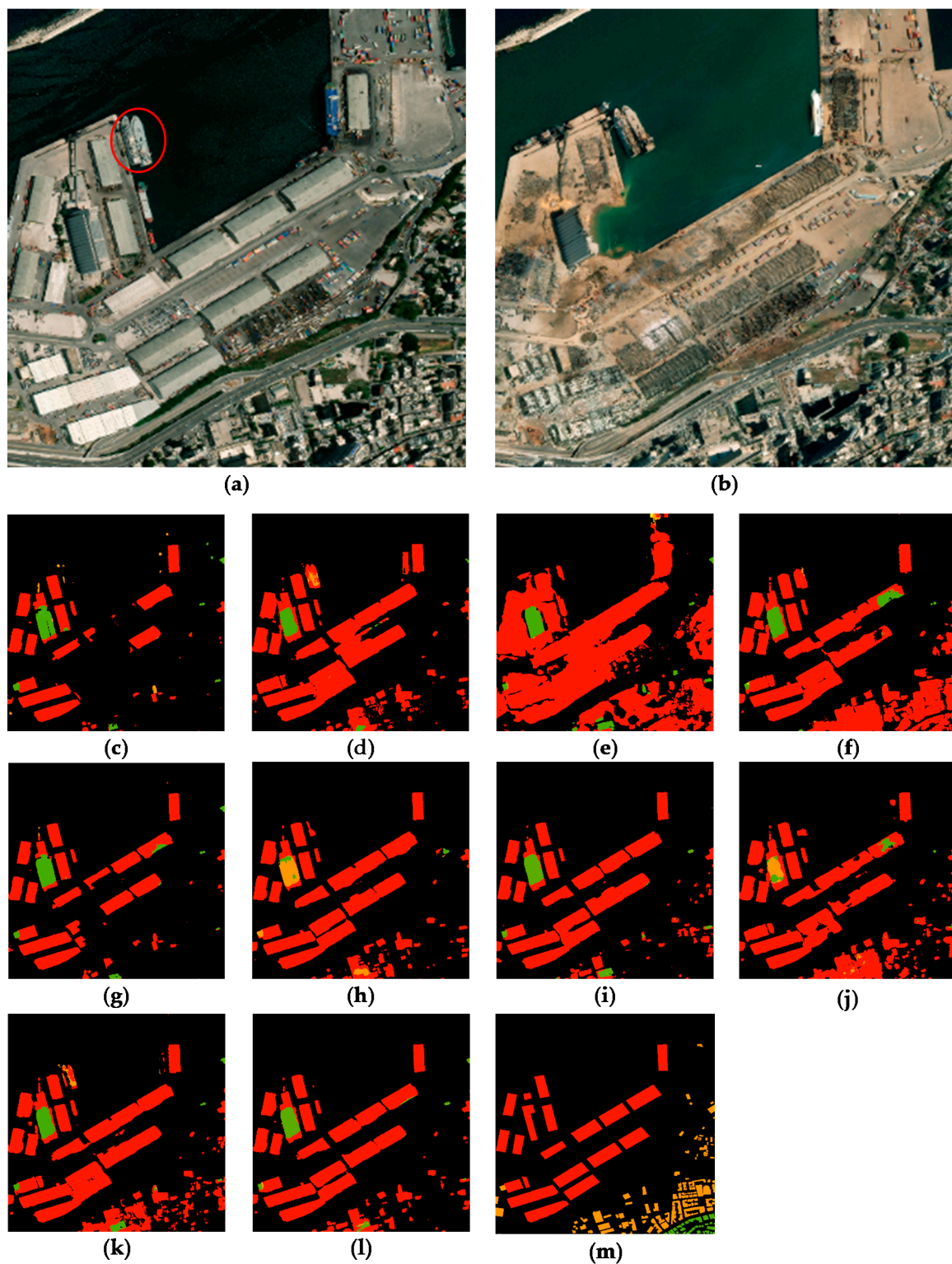


Figure 14. Results of Beirut explosion. (a) A pre-disaster image, (b) a post-disaster image, and (m) the GT. (c) The result of ResNet (w/o A), (d) the result of SResNeXt (w/o A), (e) the result of DPN (w/o A), (f) the result of SENet (w/o A), (g) the result of ResNet (w A), (h) the result of SResNeXt (w A), (i) the result of DPN (w A), (j) the result of SENet (w A), (k) the fusion result without the attention mechanism and (l) the fusion result with the attention mechanism.

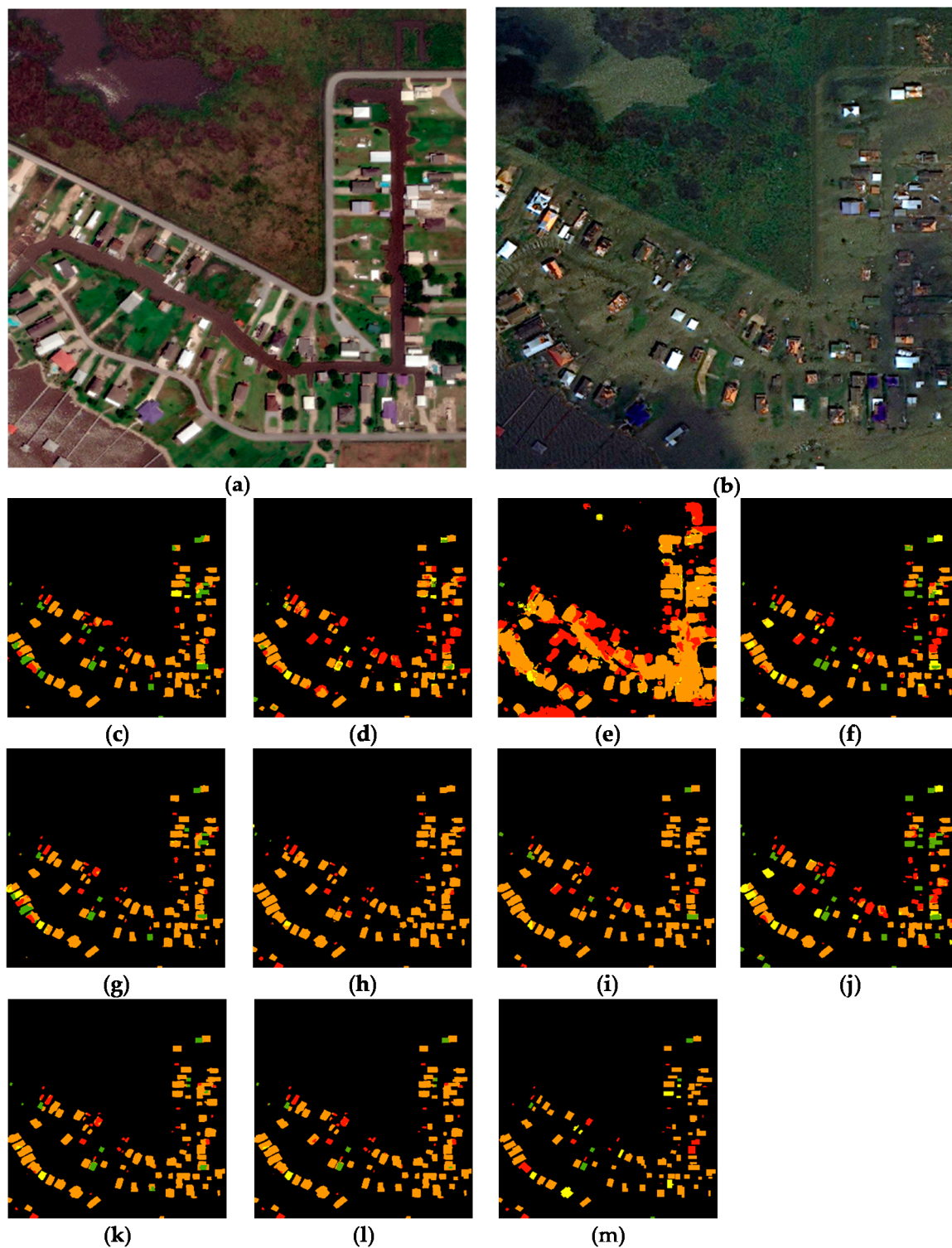


Figure 15. Results of Hurricane Laura. (a) A pre-disaster image, (b) a post-disaster image and (m) is the GT. (c) The result of ResNet (w/o A), (d) the result of SResNeXt (w/o A), (e) the result of DPN (w/o A), (f) the result of SENet (w/o A), (g) the result of ResNet (w A), (h) the result of SResNeXt (w A), (i) the result of DPN (w A), (j) the result of SENet (w A), (k) the fusion result without the attention mechanism and (l) the fusion result with the attention mechanism.

4. Discussion

It can be seen from the indicators in Section 3.1 that the U-Net with the SE module maintains better performance in the group with or without the attention mechanism, which may be due to the characteristics of the SE module; that is, the attention mechanism of

the channel. The SE module can let the network know what channels are more important for the current task. Indeed, compared to images with dozens of channels, our data does not seem to have much need to choose important channels, but it does not mean that this operation is completely meaningless. It turns out that it can improve the accuracy. At the same time, we can expect that, when the number of channels increases, the attention mechanism on the channels will have a greater effect.

In addition to the band attention mechanism, we also added the spatial attention mechanism. The accuracy of the network with the attention mechanism on the localization task did not change much. However, the accuracy on the classification task improved. For the task of building localization, the spatial attention mechanism was not very helpful, and the original convolutional neural network was enough to achieve good results. For classification tasks, the changes in the global characteristics of the image may affect the classification results of its disaster types. According to Tobler's First Law [25], "Everything is related, but nearby things are more related than distant things". The environment around a building can also be used as one of the basis for determining its damage level. As defined in Figure 1, if a house is surrounded by water/mud, no matter how it looks on the outside, it will also be classified as damaged.

The localization and classification of disaster-damaged buildings is a technology that supports post-disaster rescue. For this reason, the performance of the network on untrained disaster images is critical. Therefore, we conducted research on the transferability and robustness of the model. See Section 3.3. It can be concluded that our model performs well on different types of disasters that have not been trained, but there is still room for improvement. At the same time, the pre-trained model on the xBD dataset can be used as the basis for future building disaster detection research without the need to train the model from scratch. When using the trained model to process a pair of 1024×1024 images, it takes no more than one second to get the result and can save valuable time during disaster relief.

Tables 4 and 5 showed that the accuracy of minor damage in the four categories is the lowest, which is related to the fact that the remote sensing image does not contain information on the side of the building, such as the surrounding walls. The misjudgment caused by this lack of information is also reflected in the disaster of the Beirut explosion. If the related street view or oblique photographic image can be added, the accuracy of the model can be further improved. Regarding the reasons for restricting the best accuracy, we think there are roughly three points. One is the limitation of data; that is, the viewing angle of the remote sensing image is limited as mentioned above. The second is the limitation of the method; that is, the method used in this article is not perfect to adapt to this problem. The third is the problem of level classification. Even experts can hardly determine the damage level to some houses, because the damage itself is difficult to divide, leading to errors in data calibration.

In terms of accuracy assessment, evaluating the disaster damage level is not based on each pixel of the building as a unit but mostly based on an entire building as a unit. Therefore, the indicators and loss values used during model training can be improved, such as calculating the loss with a single building as a unit and training model with the loss.

5. Conclusions

Neural networks have been widely used in damaged building detection after disasters [4,13,16,17,32,43,47]. However, most of the studies focus on the binary classification about whether buildings collapsed or not, and most models give the same attention to each feature, which makes it more difficult for some important features to play a full role. In order to make the model focus on more important part for disaster-damaged building classifications, in this study, we described a variety of U-Nets using different backbones with the attention mechanism. These networks can automatically detect damaged buildings in satellite images and assess their level. We trained different networks using xBD and compared their F1 scores on the verification set. Among them, the performance of SEresNeXt

with the attention mechanism on two dimensions is the best, the overall F1 score reaching 0.787. For further improving the accuracy, we fused the results of four models and got better results on the fusion model with the attention mechanism than all the other models, and the overall F1 score reached 0.792. This result proves that the attention mechanism is helpful for the detection of damaged buildings. In order to verify the performance of models on untrained disasters, two disasters not in the training and verification sets were selected to verify the model's portability and robustness. The results showed that our model had good robustness and portability on localization and classification tasks, but there is still space for improvement.

A future research direction should consider specialized network training according to disaster types to improve the accuracy of different types of disasters. The classification of a building object can also be considered, which is more in line with the actual situation. We plan to consider more types of disasters, especially large-scale and high-frequency disasters. We also plan to study some technologies to make the model adapt to different data sources, such as lower resolution remote sensing data, street view data from different perspectives, etc.

Author Contributions: Conceptualization, C.W. and F.Z.; methodology, C.W. and J.X. (Junshi Xia); software, C.W. and Y.X.; writing—original draft preparation, C.W.; writing—review and editing, F.Z., J.X. (Junshi Xia), G.L. and J.X. (Jibo Xie); visualization, Y.X.; supervision, Z.D.; project administration, F.Z.; funding acquisition, F.Z., R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2019YFE0127400); KAKENHI (19K20309) and the National Natural Science Foundation of China (41671391, 41922043 and 41871287).

Acknowledgments: The authors would like to thank the editors and anonymous reviewers for their valuable comments, which helped us improve this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Frankenberg, E.; Sumantri, C.; Thomas, D. Effects of a natural disaster on mortality risks over the longer term. *Nat. Sustain.* **2020**, *3*, 614–619. [\[CrossRef\]](#)
2. Menderes, A.; Eren, A.; Sarp, G. Automatic Detection of Damaged Buildings after Earthquake Hazard by Using Remote Sensing and Information Technologies. *Procedia Earth Planet. Sci.* **2015**, *15*, 257–262. [\[CrossRef\]](#)
3. Xu, J.Z.; Lu, W.; Li, Z.; Khaitan, P.; Zaytseva, V. Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks. *arXiv* **2019**, arXiv:1910.06444.
4. Cooner, A.J.; Shao, Y.; Campbell, J.B. Detection of urban damage using remote sensing and machine learning algorithms: Revisiting the 2010 Haiti earthquake. *Remote Sens.* **2016**, *8*, 868. [\[CrossRef\]](#)
5. Dell'Acqua, F.; Gamba, P. Remote sensing and earthquake damage assessment: Experiences, limits, and perspectives. *Proc. IEEE* **2012**, *100*, 2876–2890. [\[CrossRef\]](#)
6. Wang, T.L.; Jin, Y.Q. Postearthquake building damage assessment using multi-mutual information from pre-event optical image and postevent SAR image. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 452–456. [\[CrossRef\]](#)
7. Vu, T.T.; Ban, Y. Context-based mapping of damaged buildings from high-resolution optical satellite images. *Int. J. Remote Sens.* **2010**, *31*, 3411–3425. [\[CrossRef\]](#)
8. Adriano, B.; Xia, J.; Baier, G.; Yokoya, N.; Koshimura, S. Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 Sulawesi earthquake and Tsunami in Palu, Indonesia. *Remote Sens.* **2019**, *11*, 886. [\[CrossRef\]](#)
9. Weber, E.; Kané, H. Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion. *arXiv* **2020**, arXiv:2004.05525.
10. Chen, S.A.; Escay, A.; Haberland, C.; Schneider, T.; Staneva, V.; Choe, Y. Benchmark Dataset for Automatic Damaged Building Detection from Post-Hurricane Remotely Sensed Imagery. *arXiv* **2018**, arXiv:1812.05581.
11. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99. [\[CrossRef\]](#)
12. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. In Proceedings of the ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Copernicus GmbH, Riva del Garda, Italy, 4–7 June 2018; Volume 4, pp. 89–96.

13. Guo, H.D.; Wang, X.Y.; Li, X.W.; Liu, G.; Zhang, L.; Yan, S.Y. Yushu earthquake synergic analysis using multimodal SAR datasets. *Chin. Sci. Bull.* **2010**, *55*, 3499–3503. [\[CrossRef\]](#)
14. Mao, Y.; Zhu, B.; Zhang, Q.; Wang, J. Urban Change Detection Based on High Resolution SAR and Optical Remote Sensing Data. *Urban Geotech. Investig. Surv.* **2019**, *5*, 17–20.
15. Ji, M.; Liu, L.; Buchroithner, M. Identifying collapsed buildings using post-earthquake satellite imagery and convolutional neural networks: A case study of the 2010 Haiti Earthquake. *Remote Sens.* **2018**, *10*, 1689. [\[CrossRef\]](#)
16. Pesaresi, M.; Gerhardinger, A.; Haag, F. Rapid damage assessment of built-up structures using VHR satellite data in tsunami-affected areas. *Int. J. Remote Sens.* **2007**, *28*, 3013–3036. [\[CrossRef\]](#)
17. Gamba, P.; Dell’Acqua, F.; Odasso, L. Object-oriented building damage analysis in VHR optical satellite images of the 2004 Tsunami over Kalutara, Sri Lanka. In Proceedings of the 2007 Urban Remote Sensing Joint Event, Paris, France, 11–13 April 2007; pp. 1–5.
18. Tong, X.; Hong, Z.; Liu, S.; Zhang, X.; Xie, H.; Li, Z.; Yang, S.; Wang, W.; Bao, F. Building-damage detection using pre- and post-seismic high-resolution satellite stereo imagery: A case study of the May 2008 Wenchuan earthquake. *ISPRS J. Photogramm. Remote Sens.* **2012**, *68*, 13–27. [\[CrossRef\]](#)
19. Liu, R.; Cheng, Z.; Zhang, L.; Li, J. Remote sensing image change detection based on information transmission and attention mechanism. *IEEE Access* **2019**, *7*, 156349–156359. [\[CrossRef\]](#)
20. Khelifi, L.; Mignotte, M. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. *IEEE Access* **2020**, *8*, 126385–126400. [\[CrossRef\]](#)
21. Zhan, Y.; Fu, K.; Yan, M.; Sun, X.; Wang, H.; Qiu, X. Change Detection Based on Deep Siamese Convolutional Network for Optical Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1845–1849. [\[CrossRef\]](#)
22. Ghaffarian, S.; Kerle, N.; Pasolli, E.; Arsanjani, J.J. Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data. *Remote Sens.* **2019**, *11*, 2427. [\[CrossRef\]](#)
23. Sun, Y.; Zhang, X.; Huang, J.; Wang, H.; Xin, Q. Fine-Grained Building Change Detection From Very High-Spatial-Resolution Remote Sensing Images Based on Deep Multitask Learning. *IEEE Geosci. Remote Sens. Lett.* **2020**, 1–5. [\[CrossRef\]](#)
24. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.K.; Zhang, X.; Huang, X. Road detection and centerline extraction via deep recurrent convolutional neural network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [\[CrossRef\]](#)
25. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234. [\[CrossRef\]](#)
26. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. *Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2204–2212.
27. Hao, H.; Baireddy, S.; Bartusiak, E.R.; Konz, L.; LaTourette, K.; Gibbons, M.; Chan, M.; Comer, M.L.; Delp, E.J. An attention-based system for damage assessment using satellite imagery. *arXiv* **2020**, arXiv:2004.06643.
28. Mou, L.; Zhu, X.X. Learning to Pay Attention on Spectral Domain: A Spectral Attention Module-Based Convolutional Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 110–122. [\[CrossRef\]](#)
29. Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Doshi, J.; Heim, E.; Choset, H.; Gaston, M. xBD: A Dataset for Assessing Building Damage from Satellite Imagery. *arXiv* **2019**, arXiv:1911.09296.
30. Open Data Program. Available online: <https://www.maxar.com/open-data> (accessed on 24 November 2020).
31. Castello, R.; Roquette, S.; Esguerra, M.; Guerra, A.; Scartezzini, J.L. Deep learning in the built environment: Automatic detection of rooftop solar panels using Convolutional Neural Networks. *J. Phys. Conf. Ser.* **2019**. [\[CrossRef\]](#)
32. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Munich, Germany, 5–9 October 2015; Volume 9351, pp. 234–241.
33. Ebrahim, M.; Al-Ayyoub, M.; Alsmirat, M.A. Will Transfer Learning Enhance ImageNet Classification Accuracy Using ImageNet-Pretrained Models? In Proceedings of the 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 11–19 June 2019; pp. 211–216. [\[CrossRef\]](#)
34. Oktay, O.; Schlemper, J.; Folgoc, L.L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention U-Net: Learning Where to Look for the Pancreas. *arXiv* **2018**, arXiv:1804.03999.
35. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; Volume 2016, pp. 770–778.
37. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Kai, L.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; 2009; pp. 248–255.
38. Telgarsky, M. Benefits of depth in neural networks. In Proceedings of the Journal of Machine Learning Research: Workshop and Conference Proceedings, New York, NY, USA, 23–26 June 2016; Volume 49, pp. 1517–1539.
39. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [\[CrossRef\]](#)
40. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.

41. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 5987–5995.
42. Chen, Y.; Li, J.; Xiao, H.; Jin, X.; Yan, S.; Feng, J. Dual path networks. In Proceedings of the NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing, Long beach, CA, USA, 3–9 December 2017; pp. 4468–4476.
43. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **1997**, *62*, 77–89. [[CrossRef](#)]
44. Van Rijsbergen, C.J. Information retrieval; 2nd ed.; Butterworth, 1978. *J. librariansh.* **1979**, *11*, 237.
45. Yang, Y. An evaluation of statistical approaches to text categorization. *Inf. Retr. Boston* **1999**, *1*, 69–90. [[CrossRef](#)]
46. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.O.; Villena-Martinez, V.; Garcia-Rodriguez, J. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
47. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings; International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
48. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703.
49. Corbane, C.; Saito, K.; Dell'Oro, L.; Bjorgo, E.; Gill, S.P.D.; Piard, B.E.; Huyck, C.K.; Kemper, T.; Lemoine, G.; Spence, R.J.S.; et al. A comprehensive analysis of building damage in the 12 January 2010 MW7 Haiti earthquake using high-resolution satellite and aerial imagery. *Photogramm. Eng. Remote Sens.* **2011**, *77*, 997–1009. [[CrossRef](#)]