Expert Systems With Applications 202 (2022) 117268

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A novel attention-based deep learning method for post-disaster building damage classification

Chang Liu [a], Samad M.E. Sepasgozar [b], Qi Zhang [a], Linlin Ge [a,*]

[a] School of Civil and Environmental Engineering, Faculty of Engineering, University of New South Wales, Sydney 2052, Australia
[b] School of Built Environment, Faculty of Arts, Design & Architecture, University of New South Wales, Sydney 2052, Australia

ARTICLE INFO

Keywords:
Building damage classification
Deep learning
Channel attention
Remote sensing
Natural disaster
High resolution imagery

ABSTRACT

Although several past studies proposed deep learning methods to extract pre-disaster building footprints for post-disaster management using remote sensing techniques, building damage classification after a disaster is rarely discussed. Building damage classification using deep learning is a challenging task because imagery data of damaged buildings are limited publicly and damage levels are hard to judge from images. Most building damage classification papers with deep learning methods only judge buildings as collapse or not. Moreover, most deep learning models are designed for non-damaged building segmentation, which might not be suitable for damaged building segmentation. In order to solve the problems, including lack of imagery data, detailed damage levels, and suitable deep learning model for damaged building segmentation, this research aims to propose and evaluate a novel deep learning model to classify building damage into four levels quickly with in-house labelled damaged building information. This research proposes the Squeeze-and-Excitation dual High-Resolution Network (HRNet) model to be trained with open access xBD dataset and the 2010 Haiti Earthquake images labelled at the University of New South Wales. HRNet is adopted twice in the model because two steps need to be completed, building localization and then damage classification. Four comparative experiments are implemented with seven metrics, including combination F1, localization F1, localization precision, localization recall, damage F1, damage precision, and damage recall. The first experiment is to choose where the best place is to add SE channel attention block in the model. The Squeeze-and-Excitation (SE) block is added at four different places in the dual HRNet backbone to have a comparison with the model without SE. According to analyses of the results, the SE-PRE model has the best performance. The second experiment shows that a larger input size of images increases the processing time but performs better. The third and fourth experiments reveal that the Sigmoid function is better than Hard-Sigmoid in SE, and the transfer learning with pre-trained ImageNet weights may not be suitable for the proposed model. The theoretical contribution is proposing a novel deep learning model for building damage classification. Another contribution of this research is creating a Haiti Earthquake image dataset with four building damage levels. The practical contribution is providing a safe and speedy post-disaster building damage classification method with minimal fieldwork to support rescue teams.

1. Introduction

Devastating natural disasters such as earthquakes can cause severe building damage and casualties (Ji et al., 2018). For instance, in the 2010 Haiti Earthquake, nearly half of all structures collapsed or were severely damaged in the epicentral area, including more than 300,000 homes (DesRoches et al., 2011). This means there is a need to develop some tools and techniques that are capable of detecting, analyzing, and classifying building damages for emergency responses. Due to the field restriction, the tools should be able to work remotely and automatically in terms of detection and classification.

The immediate damage estimation and classification after the occurrence of the earthquake will help emergency response plans to save human life. Besides saving lives, classifying building damage provides quick and on time information for post-disaster reconstruction plans, which can help earthquake areas to recover from catastrophes.

As one of the most common approaches after disasters, on-site damage investigation can provide detailed information, but it is time consuming and laborious, with a high risk of working in the field (Tanjung et al., 2020). If building damage levels can be obtained using

* Corresponding author.
E-mail address: l.ge@unsw.edu.au (L. Ge).

https://doi.org/10.1016/j.eswa.2022.117268
Received 13 September 2021; Received in revised form 26 March 2022; Accepted 15 April 2022
Available online 20 April 2022

remote sensing techniques with minimal delay, rescue teams and governments can make post-event decisions with the least on-site observation. Therefore, a quick post-event building damage classification method is critical to post disaster management. Remote sensing with deep learning can help to resolve this issue by obtaining building data remotely (Ji et al., 2018). The fast development of deep learning in computer vision provides a pathway to offer quick classification (Su et al., 2020; Wheeler & Karimi, 2020; Yang et al., 2021).

With the development of deep learning in computer vision, some papers apply deep learning models in building damage segmentation (Ji et al., 2018; Su et al., 2020; Wheeler & Karimi, 2020; Yang et al., 2021). However, applying deep learning in remote sensing image semantic segmentation for disaster management is still a challenging task (Gupta et al., 2019). There are two main problems that need to be solved for building damage classification with deep learning.

The first problem is the insufficiency of labelled image datasets of collapsed buildings. Very limited damaged building satellite image datasets are accessible publicly for applying deep learning algorithms. Popular 2D semantic segmentation datasets in computer vision always do not contain damaged buildings such as ImageNet (Deng et al., 2009). Some available datasets contain information about collapsed or not, which is not enough for post-disaster rescue and management. Therefore, labelled images with more than two levels (damaged or not) are needed for deep learning study.

In order to resolve the above problem, this research adopts two optical satellite image datasets, the online free xBD dataset and own labelled 2010 Haiti Earthquake dataset. Both are labelled at four levels (no, minor, major, and total damage). The xBD dataset was published for 2019 Defense Innovation Unit Experimental (DIUx) xView2 Challenge of building damage classification (Gupta et al., 2019). However, images after earthquakes are insufficient in it, though it contains several post-event images with different natural disasters such as tsunamis, bushfires, and tornados. Hence, the second dataset, the 2010 Haiti Earthquake dataset, is added to the research. Building footprints were drawn, and damage levels were labelled manually. Building damage levels in it are categorized based on the analysis in 2010 (UNITAR/UNOSAT/EC/JRC/WB, 2010). The number of its damage levels is the same as that in xBD dataset. This study drew outlines of buildings and labelled damage levels manually. Hence, this paper uses both manually labelled images and xBD dataset.

The second problem is the lack of adaptation of deep learning models for remote sensing applications in post-disaster building damage assessments. Performances of most models are not good to be applied in building damage classification. This is because these deep learning models are designed without considering damaged buildings though non-damaged building segmentation is widely discussed in the computer vision field (Krupiński et al., 2019; Majd et al., 2019). Moreover, most models are not designed for satellite images initially because these models are proposed for classifying or segmenting indoor or small objects from 2D photos which are not visible in satellite images. In order to solve this issue of the lack of suitable deep learning models, a novel deep learning model that is suitable for building damage classification with satellite images is necessary.

In order to resolve both of these two problems, this research aims to propose and evaluate a novel deep learning model to classify post-event building damage quickly using both public and own labelled damaged building datasets. The model contains two main steps, including building localization and damage classification. The first step is localizing building footprints with pre-disaster images, and then the second step is categorizing damage levels with post-event images according to the footprints from the first step.

2. Evaluation strategies for deep learning models

This section states some well-known deep learning models and their advantages reported in different contexts. This will be followed by discussing a set of appropriate strategies to compare a model in the earthquake contexts. Deep learning has been increasingly applied for image classification purposes in recent years, particularly when AlexNet is introduced in the literature (Krizhevsky et al., 2012). AlexNet improved image classification accuracy from 70%+ of conventional computer vision methods to 80%+, which is a breakthrough in terms of accuracy. The dominance of AlexNet in the classification contest was acknowledged by the ImageNet Large Scale Visual Recognition Challenge 2012 (LSVRC 2012), as a well-known competition in computer sciences, and thus the application of deep learning for image classification in various contexts has been further increased.

One of its ground-breaking contributions is that it first uses the graphics processing unit (GPU) to accelerate training speed. Second, it applies Rectified Linear Unit (ReLU) activation function instead of conventional activation functions to increase accuracy. Third, local response normalization (LRN) was proposed to improve generalization ability. One of the most important tasks of deep learning is to improve the generalization ability. Generalization means the ability of deep learning models to react to new data. If a model has a higher generalization ability, it means the accuracy of this model will be higher for new data. Fourth, its first two fully connected layers use the “dropout” method to decrease the possibility of overfitting.

After that, another famous net, VGGNet, was proposed by the Visual Geometry Group of the University of Oxford (Simonyan & Zisserman, 2014). It was the winner of LSVRC 2014 localization task and the second place of the LSVRC 2014 classification task. Its highlight is that it applies two 3×3 kernels replacing one 5×5 kernel and three 3×3 kernels replacing one 7×7 kernel. They have the same receptive field. This can reduce required parameters during computing. The receptive field is the size of the region on the input layer corresponding to one feature (cell) on the output feature map.

The residual block was proposed in ResNet (He et al., 2016). ResNet achieved the winner of image classification, localization, and detection in ILSVR Challenge 2015. It was also the first place of object detection task and image segmentation task of MS COCO Challenge 2015. ResNet-34 reduced top-1 error by 3.5% on ImageNet validation compared to its plain counterpart (He et al., 2016). One of its significant contributions is the “residual block”. Another one is that it applies batch normalization (BN) for accelerating the training instead of dropout.

The residual block was proposed for addressing two problems: 1) gradient vanishing problem and gradient exploding problem; and 2) degradation problem. Residual blocks are skip-connection blocks to improve the accuracy of deep learning models and show good results based on the ResNet test. The output is the addition of an identity function and residual blocks.

In recent years, the use of attention mechanisms for image classification and semantic segmentation has developed quickly. Examples of the attention mechanism in the literature are SENet (Hu et al., 2018) and Vision Transformer (Dosovitskiy et al., 2020). Attention mechanism was applied in the natural language processing field initially. It has been increasingly used for other applications such as image processing since several computer vision researchers realized its advantages (Dosovitskiy et al., 2020).

High-Resolution Network (HRNet) was proposed for human pose estimation initially (Sun, Xiao et al., 2019). Later, its authors applied HRNet for image classification, semantic segmentation, object detection, and facial landmark detection. Since it has been tested for both classification and segmentation, so this research chooses this model. One great advantage of HRNet is that it maintains high-resolution representations in the network. Several conventional deep learning models decrease input sizes with losing information to decrease the amount of calculation such as UNet (Ronneberger et al., 2015). Compared with those methods, HRNet can obtain features from the original high-resolution input images with keeping information. A higher resolution image contains more features and information than a lower resolution image.

Considering the advantages of abovementioned deep learning models, this paper adopts three different strategies, including residual blocks, Squeeze-and-Excitation (SE) attention mechanism, and GPUs training strategy (He et al., 2016; Hu et al., 2018; Krizhevsky et al., 2012). The HRNet is adopted due to its capability of maintaining the high-resolution of input images,which is important for sensitive earthquake analyses. Since there are two steps required for the analysis, namely, building localization and damage level classification, a dual HRNet is utilized in the present experimentation to cover both steps.

A set of comparative evaluations are required to address the need to identify efficient models' performances. The literature on applying deep learning models in different contexts is rich. However, there are less resources and evidence evaluating different versions of the models using different functions or blocks applicable to earthquake data sources. The following four items are critical to examine the accuracy of a model: the place of inserting a block into the body of a model (Hu et al., 2018), the capability of running with or without pre-trained weights (Liu et al., 2021), the possibility of using various input image sizes (Wang et al., 2020), and efficiency of activation functions in the SE block. The comparative evaluation of different places of a block inserted into a model is recommended to be considered for the evaluation purpose. For instance, Hu et al. (2018) compared the accuracy of different deep learning structures for the image segmentation task by adding a block in different parts of the same backbone. Therefore, in this research, the SE block is inserted in different places in the basic block of HRNet as the first comparative experiment, in order to analyze where the best place is to add it.

Another item that should be examined for each model is the capability of running with or without pre-trained weights to assess the accuracy of the outcome for the context of earthquake building damages. For example, Koo et al. (2020) used the ImageNet dataset, but they have neither applied any comparisons nor reported the reason for using the pre-trained weights in their experimentation. The next item concerns the accuracy of the output, where the sizes of the input images are different in real-life events. There are limited published experimentations reporting this type of comparison because most computers have limitations to process blended and large images without a computer with high specification and expensive GPUs. This comparison is valuable since it can address the question whether smaller-sized images can offer an output with similar accuracy as the larger size images. The last item refers to the evaluation of different activation functions that can affect the optimization of a model. Due to the availability of different activation functions and the gap of sources reporting the effect of each function in different contexts, it is required to examine any selected functions for earthquake building detection purposes. The current paper will fill the gap by conducting a set of four comparisons, which will be discussed in the following sections.

3. Method

3.1. Workflow

This paper proposes a SE-based dual-HRNet model for building damage classification. There are three main stages in this paper, including data pre-processing, model training, and model testing. Four comparative experiments are implemented in the test stage. The workflow is shown in Fig. 1.

3.1.1. Data pre-processing stage

The first stage is data pre-processing. The input data are pre- and post-event images from disasters. These images are xBD dataset and the 2010 Haiti Earthquake dataset. Each image is labelled with all building footprint coordinates and the building damage level of each building. All images are cropped, flipped, rotated, and scaled randomly for data augmentation. The detailed augmentation information of these two datasets is stated in Section 3.2.

3.1.2. Training stage

At the second stage, five structure options of the model are trained. They are SE-Standard, SE-PRE, SE-POST, SE-Identity, and No-SE. The training stage contains two steps, including localizing buildings with pre-event images and classifying building damage into four levels with post-event images (representing no, minor, major, and total damage, respectively). The number of image pairs at this stage is 3249 for training and validating the model. Each pair contains a pre-event image and a post-event image. The structure of the model applies the backbone model twice to connect these two steps, which is the dual-HRNet model. Cross-validation method is applied to find the optimal parameter configuration. The information of the model is stated in Section 3.3. To be specific, key blocks in the proposed model are shown in Section 3.3.1,

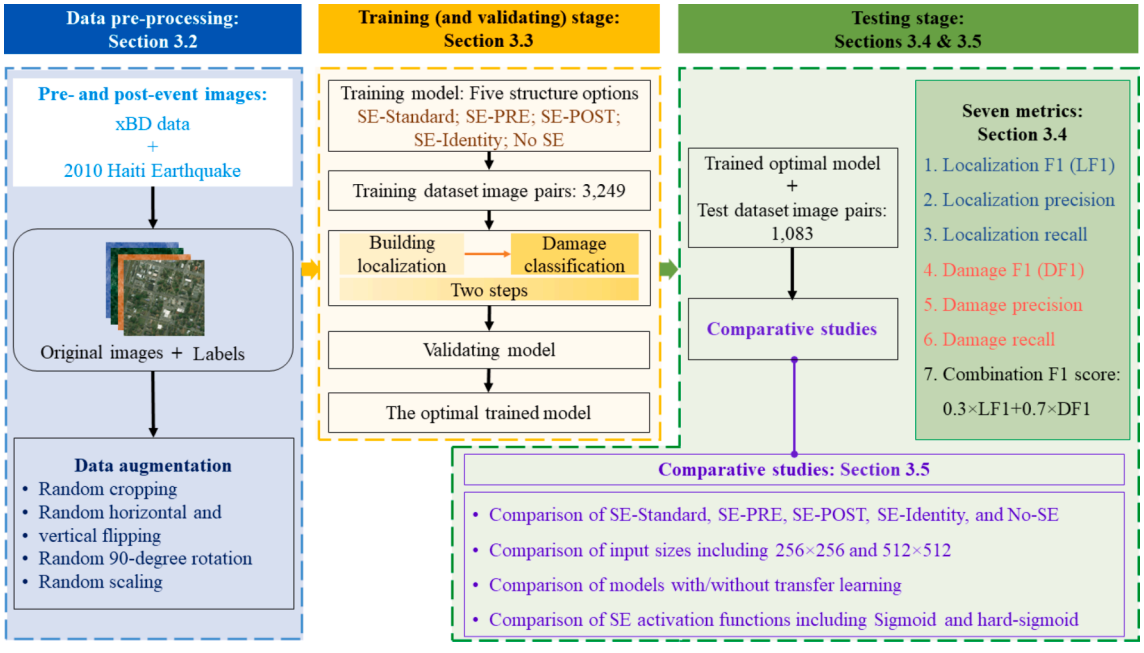


Fig. 1. Workflow of this paper.

and the detailed structure of the model is stated in Section 3.3.2.

3.1.3. Testing stage

At the third stage, the optimal model of each option from the training stage is tested with the test dataset based on seven metrics, including a combination F1 score, three metrics for building localization (localization F1, localization precision, localization recall), and three metrics for damage classification (damage F1, damage precision, and damage recall). The optimal model is chosen from the model with the highest combination F1 score. The number of test dataset image pairs is 1083. The results show the damage level of each building. The results of these seven metrics are recorded for the following comparative studies. The details of metrics are stated in Section 3.4.

As shown in Fig. 1, four types of comparison experiments are implemented in this study during the test stage. The first comparison is to find where the best is to place the SE channel attention block. Four models with different SE added places and one model without SE block are compared. Second, to judge the influence of cropped input size, a comparison of 256 × 256 and 512 × 512 is made. Third, models with and without transfer learning are trained to judge whether the pre-trained ImageNet image classification weights can improve the model performance or not. The last one is the comparison of two activation functions in SE block, including Sigmoid and Hard-Sigmoid. Details of each experiment are stated in Section 3.5.

3.2. Dataset

The dataset of this study contains pre-event non-damaged and post-event damaged building images. Building damage is classified into four levels, including no, minor, major, and total damage. All images in this dataset are high-resolution optical satellite images with red, green, and blue (RGB) bands. They are collected from multiple types of natural disasters, including earthquakes, volcanic eruptions, hurricanes, floods, tsunamis, and wildfire. This dataset contains 8664 images in total, which are 4332 image pairs. Each image pair has two images, namely, a pre-event image and a post-event image. This dataset is a mix of two datasets, xBD dataset, and 2010 Haiti earthquake dataset. 1200 images (600 pairs) come from the Haiti earthquake, and 7464 images (3,732 pairs) are chosen from xBD dataset.

The image number for training is 6,498, including 3,249 image pairs, which are 75% of the whole dataset. Among these training images, 900 images (450 pairs) are collected from the Haiti earthquake, and 5,598 images (2799 pairs) are chosen from xBD dataset. Among them, 10% of training images are chosen as validation dataset randomly. Test images are 25% of the whole dataset, including 300 Haiti earthquake images (150 pairs) and 1,866 xBD images (933 pairs). The details of xBD dataset and the Haiti earthquake dataset are stated in the following sections.

3.2.1. XBD data

Online free xBD dataset was published in 2019 for xView2 Challenge (Gupta et al., 2019). This study chose 5598 images from this dataset for training (2799 pairs of pre- and post-event images) and 1866 images for testing (933 pairs of pre- and post-event images). The chosen images in this study cover several natural disaster events, including volcanic eruption, hurricane, earthquake, flood, tsunami, and wildfire. These images were collected from different satellites, including GeoEye-1, WorldView-2, WorldView-3_VNIR, and QuickBird-2 (Su et al., 2020). The chosen images do not contain any images from the 2010 Haiti earthquake. The size of each image is 1024 × 1024 pixels below a 0.8-meter ground sample distance (GSD) mark. The damage level number is decided by experts invited by the committee that held the xView2 Challenge, which contains four levels mentioned at the beginning of Section 3.2.

3.2.2. 2010 Haiti earthquake data

The 2010 Haiti damage building data were labelled in ArcMap manually at the University of New South Wales (UNSW). An M_w 7.0 earthquake happened in Haiti on 12 January 2010, causing serious building damage (Ji et al., 2018). All images were chosen from Port-au-Prince, which is one of the seriously damaged provinces in Haiti. The size of each image is set as 1024 × 1024 pixels to have the same size as xBD imagery. Images have some overlapping areas with each other. These optical satellite images were downloaded from Maxar/DigitalGlobe Open Data Program (Maxar, 2010). The GSD is 0.8 m. The damage level of each building is according to the damage report of this earthquake provided by UNITAR/UNOSAT/EC/JRC/WB (2010). This report also classified building damage into four levels, which is the same number as the damage level number of xBD data. The building location shapefile is provided by the Operational Satellite Applications Programme (UNOSAT), Joint Research Centre (JRC), and World Bank (WB). The shapefile contains the location point of each building. A building may contain more than one point if it has more than one roof with different heights. These building damage assessment points are projected to the optical images, as shown in Fig. 2. The area in red is the chosen area for labelling. Footprints were drawn and building damage levels were labelled according to these provided damage levels and location points information.

Fig. 3 shows an example of labelled buildings. The colors representing no, minor, major, and total damaged levels are white, yellow, orange, and red, respectively. It should be noticed that some buildings contain more than one location point. If these points of a building were assessed at different levels, the labels would separate a building into parts with different damage levels according to the assessment.

3.2.3. Data pre-processing and augmentation

The size of each image for training is cropped from 1024 × 1024 pixels to 256 × 256 pixels randomly. The choice of 256 × 256 is according to several reasons. First, several widely applied deep learning models are trained with the test of small input images smaller than 300 when these models are proposed initially. For instance, both SENet (Hu et al., 2018) and ResNet (He et al., 2016) adopt 224 × 224 for training cropping. The original size of each image is 1024 × 1024, and 1024 can be evenly divisible by 256, not 224. Hence, this paper selects 256 instead of 224. The 256 × 256 size is enough to cover the identified target range, and the use of smaller input is conducive to reducing the number of parameters, reducing the risk of overfitting, and increasing the operation speed.

Second, based on several attempts, the hardware of this study can afford the training with the input of 256 × 256 pixels for all five structure options shown in Fig. 8 of Section 3.5.1. If the cropped input size is larger, the training time is very long, and the GPU memory is not

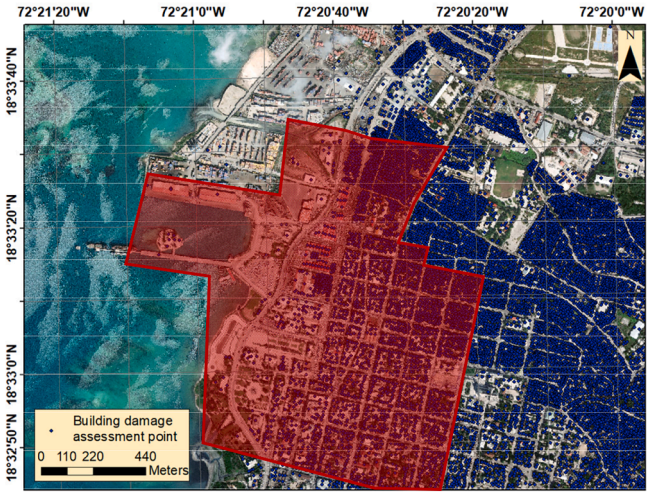


Fig. 2. The selected Haiti Earthquake dataset area (red aera).

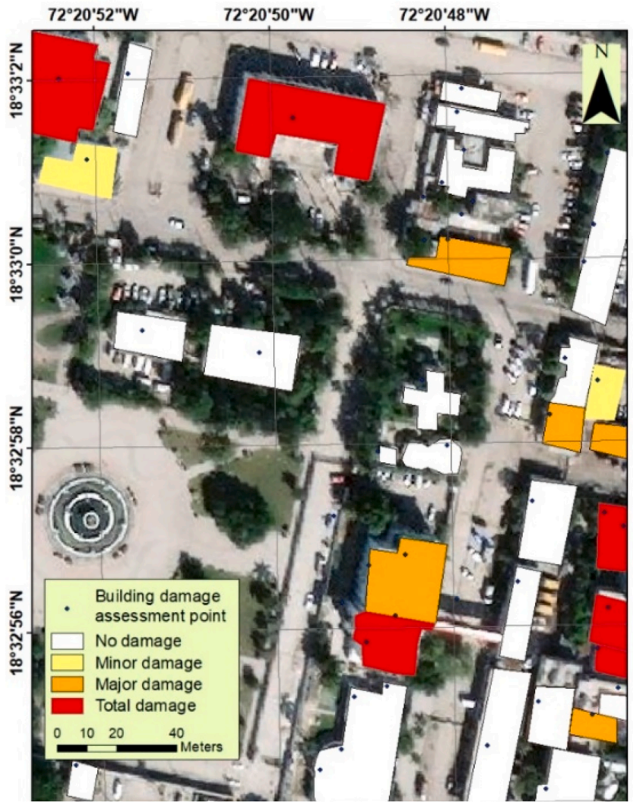


Fig. 3. An example of labelled buildings with location points (UNITAR/UNOSAT/EC/JRC/WB, 2010) and labelled building damage levels.

enough. This is also the reason why this study only uses the standard SE model other than SE-PRE, SE-POST, and SE-identity for the comparative experiment with different pixels as stated in the first paragraph of Section 3.5.2. If 512 × 512 input size is applied in the other structure options, the GPU memory is not enough.

This research applied data augmentation. To reduce the possibility of overfitting, data augmentation, including random horizontal and vertical flipping, random 90-degree rotation, and random scaling (between 0.8 and 1.2), are adopted. Effective data augmentation can not only increase the number of images in the training set but also enrich the diversity of samples. On the one hand, it can avoid the overfitting phenomenon. On the other hand, it can improve the performance of deep learning models based on prior experience.

3.3. Proposed model in this paper

This section introduces the proposed model in this paper. First, the key blocks applied in the model are stated in Section 3.3.1. Second, the structure of this proposed model is explained in Section 3.3.2.

3.3.1. Key blocks in the model

3.3.1.1. Residual block. There are two most applied structures of residual blocks in this research, as shown in Fig. 4, whose input sizes are the same as the output sizes. Structure A is the basic residual block, which was designed for ResNet with 18 or 34 layers initially (He et al., 2016). Structure B is called the bottleneck residual block. It was designed for the network with more layers, including ResNet with 50/101/152 layers (He et al., 2016). Bottleneck residual blocks are variants of basic residual blocks. These bottleneck blocks utilize 1 × 1 convolutions to reduce the number of parameters and calculating times. The design of the bottleneck helps to increase the depth with less parameters of a deep learning model than basic residual blocks.

There are also some other structures of residual blocks applied in this research that require their output sizes to be different from the input sizes. Downsampling is added in the identity part to change the number of channels in these structures. The downsample parts contain convolutional and BN layers.

BN is widely applied before activation functions to speed up the training (Koo et al., 2020), so this research also applied BN. BN was proposed in 2015 (Ioffe & Szegedy, 2015). BN is a technique to normalize inputs to one layer in each batch. If it is added before activation functions, activation functions usually have better performance than those without BN. A model with BN does not need to set the bias in convolutional layers before BN, because bias is useless before BN.

The calculation details of BN are shown in Eq. (1). The output of BN is y. $E[x]$ is the mean, and $Var[x]$ is the variance. ε is the number to avoid

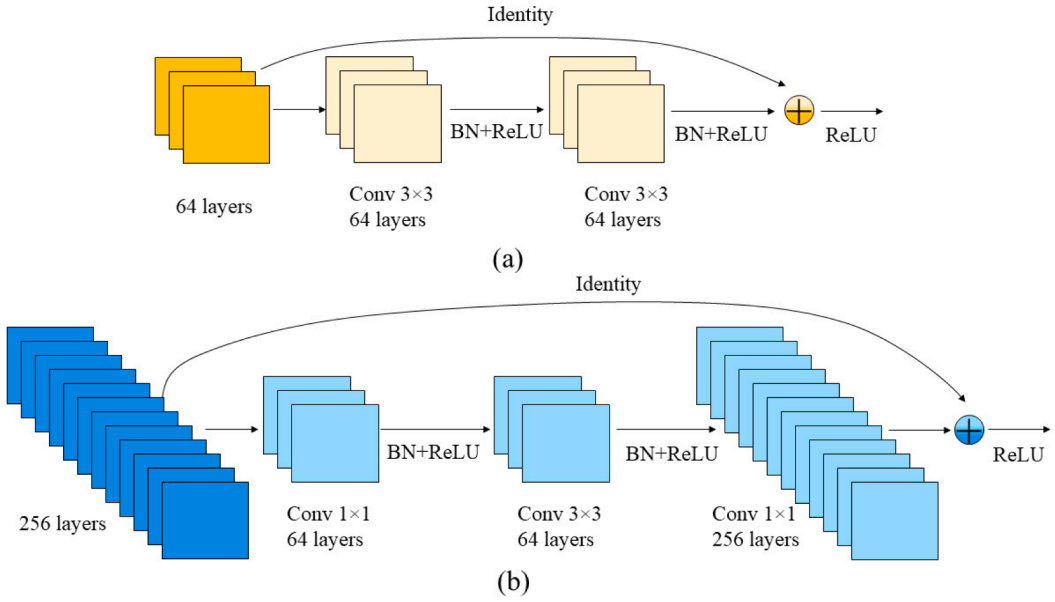


Fig. 4. Residual block structures. (a) Structure A: basic residual block; (b) Structure B: bottleneck residual block.

the devisor being zero. γ and β are learnable parameter vectors of size C. γ is 1 and β is 0 by default. The input shape should be (N, C, H, W), which represents the batch size, channel number, height, and weight, respectively. The performance of BN is better with larger batch sizes. With considering the hardware condition, this research chooses 4 × 4 as the batch size.

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} * \gamma + \beta \tag{1}$$

In this research, a defined BN layer in the Torch library (Paszke et al., 2021) is added between convolutional layers and the ReLU activation function. ReLU is a piecewise linear function. Its output is positive if the input is positive. Otherwise, the output will be zero.

3.3.1.2. Channel attention block. The attention mechanism in computer vision draws lessons from the visual attention mechanism in the human visual system. The attention block used in this research, SE block, was proposed in the SENet deep learning model (Hu, Shen, & Sun, 2018). SENet won first place in ILSVRC 2017 Classification Challenge. It mainly studies the correlation between channels and selects the attention for channels. Although it slightly increases the amount of calculation in computers, the effect is better shown in SENet. Fig. 5 shows the structure of the SE block applied in this research. First, a H × W × C block is converted to 1 × 1 × C block after global pooling. After that, with the fully connecting operation, ReLU activating, and another fully connecting operation, a 1 × 1 × C block with attention channels is achieved. Then, this research chooses both Sigmoid and Hard-Sigmoid activation functions for training to compare the performance of each other. The details of these two functions and the comparative study will be introduced in Section 3.5.2. After that, the block is scaled up to the original size. The hyper parameter r is 16 in this research according to the number in SENet.

3.3.1.3. HRNEt. HRNet structure contains four stages. High-to-low resolution convolutions are connected in parallel in this structure. As shown in Fig. 6, the light-yellow block, the original high-resolution input with all image information, is kept from the beginning to the end of the structure. Gradually scaled-down images are added at the first layer of each stage except the first stage. To be specific, the sizes of the light-orange, light-red, and dark-red blocks decrease gradually with reduced resolution.

This research adopts HRNetV2 (Sun, Zhao, et al., 2019), which is the second version of HRNet. HRNetV2 adds all channels with different resolutions together at the end of the structure. HRNetV1, the first version, was designed for human pose estimation at the beginning, so it is not suitable for image classification and segmentation. In order to be more adaptive for image semantic segmentation, HRNetV2 was designed by concatenating the (upsampled) representations soon after HRNetV1. The structure of HRNetV2 is shown in Fig. 6.

3.3.2. Structure of the proposed model

The structure of the model proposed in this paper is shown in Fig. 7. A dual HRNet with added SE model is designed. The "dual HRNet" in this model means a parallel structure with two HRNet. This research kept using "dual-HRNet" which is from the fifth place of xView 2 Challenge (Koo et al., 2020). This dual HRNet structure is used for the two steps in this model, including building localization and damage classification, respectively, as shown in the blue and the orange rectangles of Fig. 7. The first HRNet (shown in blue) gives the outputs of building locations. Only pre-event images are used in it. The second HRNet (shown in orange) accesses building damage levels. Building footprints in it are according to the locations from the results of the first HRNet model. Post-event images are exploited in the second HRNet with the building location results based on pre-event images. Hence, the results of this structure contain both building locations and damage levels.

These two HRNet are fused by adding the output channels together of one stage at the beginning of the next stage. The main function of the convolution layer is to extract features, which can provide deeper features through multi-layer convolution. Therefore, this study keeps all output layers of the two HRNet for saving information. SE channel attention block is added at each basic residual block in the dual HRNet. The detailed structures of the four adding SE options are given in Section 3.5.1.

The version of HRNet in this paper is HRNetV2W32. The structure of HRNetV2 has been explained in Section 3.3.1.3. "W32" means that numbers of convolutional layers in the four stages are 32, 64, 128, and 256, respectively.

Pretrained ImageNet classification weights are added in this research. The training with pre-trained weights is called transfer learning. Pretrained weights are often be used in deep learning to save training time and have good results (Kolar et al., 2018). These weights are the most suitable ones for a completed task which is similar to the task of this paper. For instance, pretrained ImageNet classification weights suit the image classification task with ImageNet dataset. If one task is similar to that, these weights can be added during the training, and the results may be better than training from scratch.

The output size of a convolutional layer is shown in Eq. (2) (Dumoulin & Visin, 2016).

$$n_{out} = \frac{n_{in} + 2p - k}{s} + 1 \tag{2}$$

n_{in}: number of input features. n_{out}: number of output features. k: kernel size. p: padding size. s: stride size.

Stochastic Gradient Descent (SGD) optimizer is applied with the base learning rate of 0.05. It is a hyper parameter to adjust the weights of the model, and how to use it is shown in Eq. (3). The momentum is 0.9, and the weight decay is 0.0001. The training epochs are 500, and the trained models are recorded every 50 epochs. Training 500 epochs is according to the number of training epochs from other papers which also use xBD dataset. For instance, Koo et al. (2020) trained 250 epochs, and Wheeler and Karimi (2020) trained 100 epochs. Hence, 500 epochs are enough, and this research checks the validation results during the training to avoid overfitting. Recording the model every 50 epochs is for observing the training details. The model was trained on one Nvidia 2080 Ti GPU server with 60G memory in Linux system. Considering the condition of hardware, the batch size per GPU is 4 for both training and testing. The

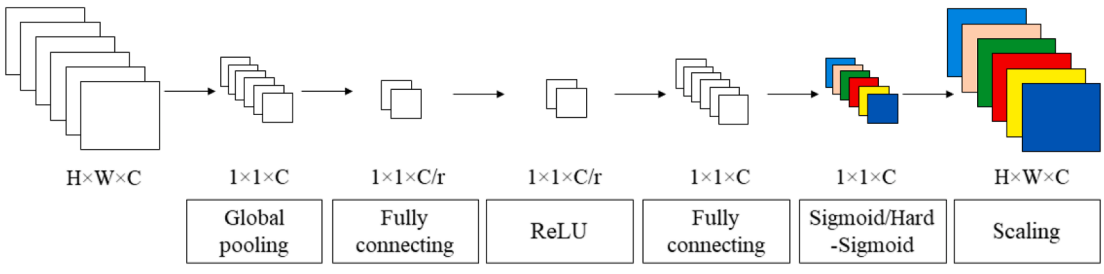


Fig. 5. SE block in this research. H: Height; W: Width; C: Number of channels.

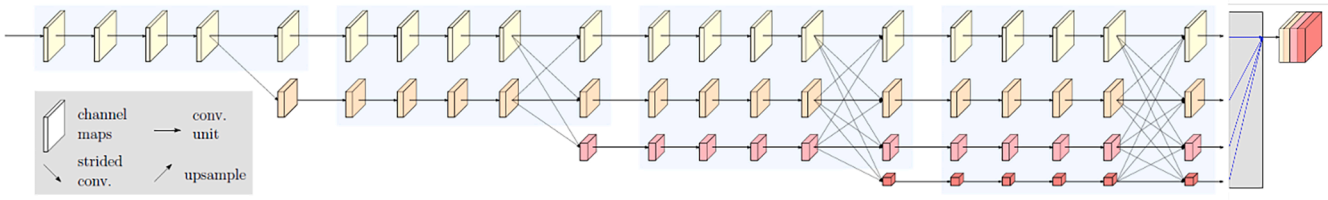


Fig. 6. Structure of HRNetV2 (Sun, Zhao, et al., 2019).

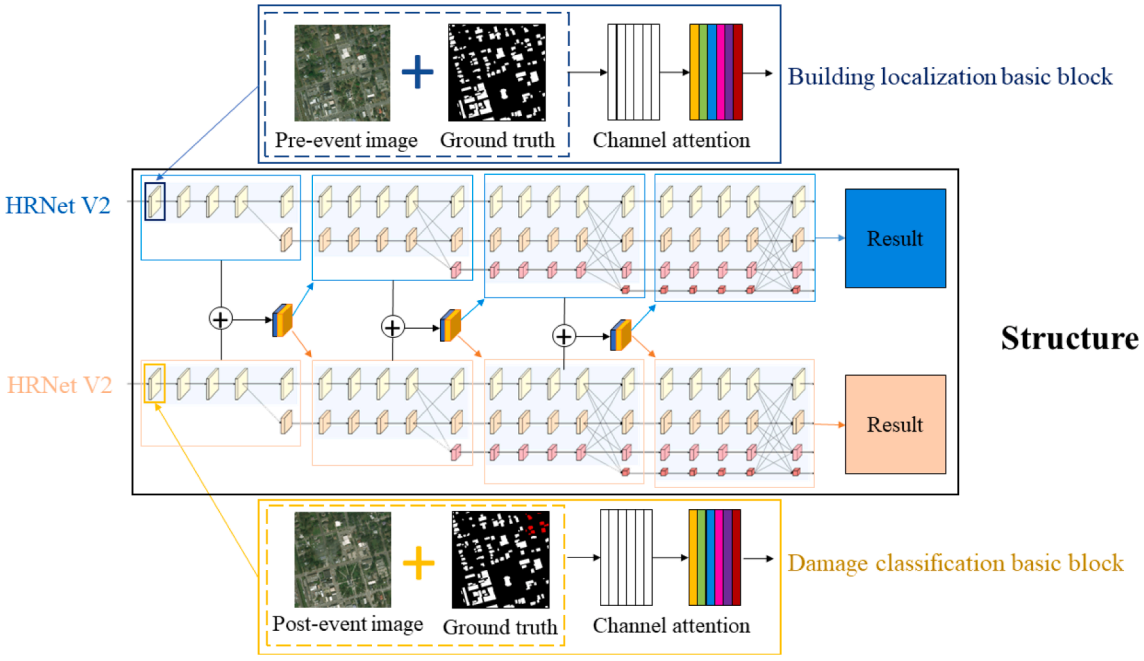


Fig. 7. Structure of the proposed model.

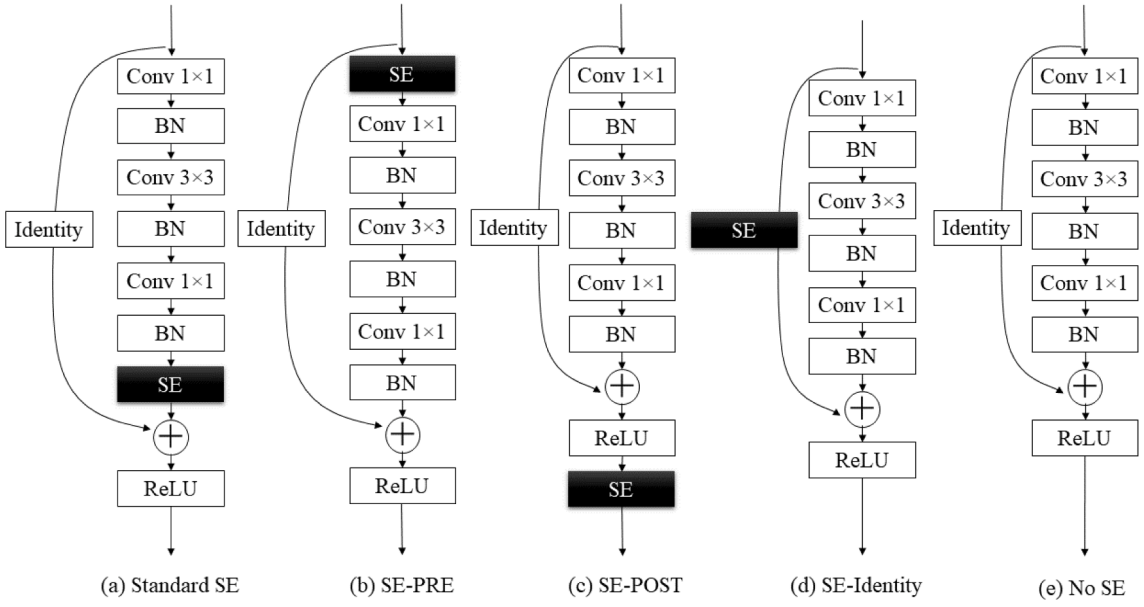


Fig. 8. The blocks with different options of inserting SE attention.

training loss is the sum of localization loss and damage classification loss by calculating Lovasz-softmax loss (Berman et al., 2018).

$$new_{weight} = existing_{weight} - learning_{rate} \times gradient \tag{3}$$

3.4. Performance metrics

This paper attempts to test the model performance for each model using seven metrics, including a combination F1 score, three metrics for localization, and three metrics for damage classification. The combination F1 score is the main metric to judge the performance of the model. This is because it contains both building localization and damage classification results. The three metrics for building localization are localization F1 (LF1) score, localization precision, and localization recall. The three metrics for damage classification are damage F1 (DF1) score, damage precision, and damage recall. Since this study recorded a model every 50 epochs ending at the 500th epoch as mentioned in Section 3.3.2, so each model has ten recorded test results with these seven metrics.

The equations of the first three metrics, including F1 score, precision, and recall for localization, are presented from Eqs. (4) to (6). True positive (TP) represents a building pixel that is segmented correctly. False positive (FP) means a non-building pixel segmented as building. True negative (TN) is a non-building pixel that is correctly segmented as non-building, and false negative (FN) is a building pixel segmented as non-building wrongly.

$$F1 = \frac{2}{precision^{-1} + recall^{-1}} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

The next three metrics, F1 score, precision, and recall for damage classification, are computed as the harmonic mean of those scores of the four damage levels, as shown from Eqs. (7) to (9), respectively. In these three equations, ε is 10^{-6} to avoid the denominator being 0 and i means the building damage level from one to four.

$$DF1 = \frac{4}{\sum_{i=1}^{4} \frac{1}{F1_i + \varepsilon}} \tag{7}$$

$$Damage\ \ precision = \frac{4}{\sum_{i=1}^{4} \frac{1}{precision_i + \varepsilon}} \tag{8}$$

$$Damage\ \ recall = \frac{4}{\sum_{i=1}^{4} \frac{1}{recall_i + \varepsilon}} \tag{9}$$

The F1, precision and recall of each damage level from one to four use the same equations as the equations for building localization step as shown from Eqs. (4) to (6). However, the meanings of TP, TN, FP, and FN for damage classification are different than their meanings for building localization. TP means a pixel contained in this damage level is classified by the model correctly. TN means one pixel that is not included in this damage level is classified correctly. FP means that a building pixel which is not contained in this damage level but wrongly classified as this level, and FN represents one pixel in this level is wrongly classified as another level by the model.

The last metric, the combination F1 score, is applied according to the xView2 Challenge (DIUx, 2019). The combination F1 score calculates a weighted average of LF1 and DF1, as shown in Eq. (10). The numbers are chosen as 0.3 and 0.7 because xView2 Challenge applied these numbers. The percentage of LF1 or DF1 is designed by experts from that xView2 Challenge. Parts of images applied in this research are collected from xBD dataset published for xView2 Challenge as mentioned in Section 3.2.1.

$$Combination\ \ F1\ \ score = 0.3 \times LF1 + 0.7 \times DF1 \tag{10}$$

3.5. Comparative experimentations at the test stage

3.5.1. Experimentation of SE block integrations

Four structure options with different SE added places in the model are compared with the original backbone model in this experiment. Since the final output size of the SE block is the same as its input size, it can be added anywhere in the model backbone without the need to change other layers. In this research, SE is added in the following four options in each residual block of HRNet shown from Fig. 8 (a) to (d) for the comparative study. In other words, SE is added both in basic and bottleneck residual blocks in the codes. Fig. 8 (e) is the initial block without SE block. The bottleneck residual block mentioned in Section 3.3.1 is chosen as the example to show the added places of SE block shown in Fig. 8. The first option shown in Fig. 8 (a) is adding SE after residual blocks, and then the results are added with the identity function. The second option is adding SE before residual blocks shown in Fig. 8 (b). The third one is adding SE after the addition of the identity function and residual blocks shown in Fig. 8 (c). The last one is adding SE and residual blocks together, so the SE block takes the place of the identity part, as shown in Fig. 8 (d). The input size of this comparative study is 256 × 256.

Although two papers (Li, Tian et al., 2020; Li, Wang et al., 2020) have also applied the SE attention mechanism, the targets of their papers are scene classification and human pose estimation, respectively. These applications are different from the applications in this research, and the SE block added places are different. Another difference is that they added the SE block in only one place, while this research applies not only their methods but also others. Moreover, although no attention mechanism related codes can be found in the codes provided by (Li, Wang et al., 2020) in GitHub (which is only HRNet codes), it added SE block parallel with residual block according to the figure in its paper. Its structure is shown in Fig. 8 (d). SE-HRNet added SE block before the summation, as shown in Fig. 8 (b). Hence, their model structures are different from the model of this paper even though some parts are similar. This research attempts more possible combination modes than them.

3.5.2. Experimentation of other hyper parameters

Besides the comparison of SE added places, three types of comparative experiments are implemented in this research. The first one is the input size. Cropped input images with both 512 × 512 pixels and 256 × 256 pixels after data augmentation are trained for comparison of the model performance with different input sizes. As explained in Section 3.2.3, the choice of these input sizes is considering the hardware performance and previous experience. The “Standard SE” structure is chosen as the training model in this test experiment.

The second comparison is the results between training with and without transfer learning. To be specific, one model is training with HRNetV2W32_ImageNet_pretrained weights, and the other one without transfer learning is training from scratch. Although training with pre-trained weights always has a better result than without a pre-trained model theoretically, it is still necessary to check the results. Therefore, this comparative experiment is implemented in this research.

The third one is comparing Sigmoid and Hard-Sigmoid activation functions in SE block. Sigmoid is shown in Eq. (11). Here “e” is Euler’s number.

$$S(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

Hard-Sigmoid is the segment-wise linear approximation of Sigmoid,

which is far less computationally expensive than Sigmoid both in software and specialized hardware implementations, as reported by Courbariaux et al. (2015). Equation (12) shows its formula.

$$\sigma(x) = clip\left(\frac{x+1}{2}, 0, 1\right) = \max\left(0, \min\left(1, \frac{x+1}{2}\right)\right) \tag{12}$$

4. Results

The results of the four comparative experiments (refer Section 3.5) are presented in this section. As mentioned in Section 3.3.2, all results are obtained based on the experiments with the test dataset. The performance of each model is compared based on its optimal model among 10 recorded models in this research.

The outputs of the test stage are presented as RGB images and evaluation scores. The RGB images contain the footprint of buildings and the damage level of each building. The evaluation scores are based on seven metrics as introduced in Section 3.4.

In each comparative experiment, the combination F1 score is the main metric to judge the performance of each model since it considers both building location and damage classification results. This research not only computes the combination F1 score of each option but also visualizes the scores as line charts of ten intermediate epochs from 50 to 500 that show the score of the training progress for each experiment. Moreover, F1 scores, precisions, and recalls for localization and classification of the optimal model during training are also shown as line charts in this section. The details of all measures for each epoch are all recorded in supplementary documents for reliability checks and reference. Four comparisons are stated one by one as follows.

4.1. Comparison of SE block integrations

This section presents the outcome of the damage level classification using five structure options (refer Section 3.5.1) as shown in the purple part of the workflow of Fig. 1 in Section 3.1. In this section, the RGB image results are analyzed first and shown in Fig. 9. Then, the results of all the seven metrics are analyzed, and the summary is presented in tables or as line charts. A sample of the pre-event image is shown in Fig. 9 (a), which includes buildings and vast green vegetation. Fig. 9 (b) shows what areas have been destroyed due to the disaster by applying

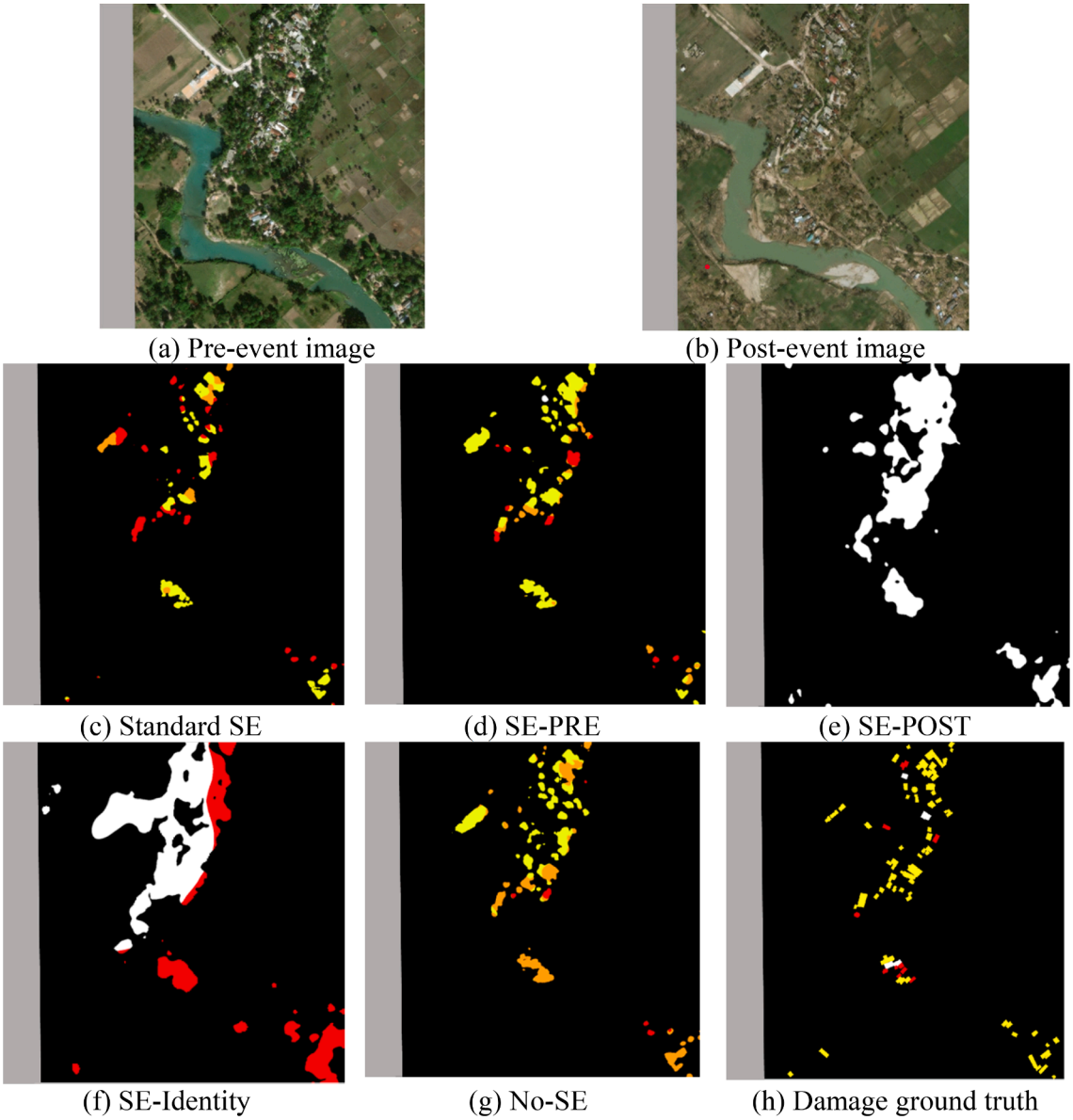


Fig. 9. Sample result: "hurricane-matthew_00000010" in xBD dataset.

the algorithm to the testing sample. Fig. 9 (c) to (g) are visualized results of the five options (refer Fig. 8 in Section 3.5.1). To validate the performance of the results of each option with the ground truth, Fig. 9 (c) to (g) are compared with Fig. 9 (h). As mentioned in Section 3.2, white, yellow, orange, and red colors in these images represent no, minor, major and total damage levels, respectively. Grey means no data.

Fig. 9 (c) to (g) show that all the five options can segment buildings from a small 256 × 256 image, which is hard to be judged by human eyes. Standard SE, SE-PRE, and No-SE perform well for identifying both building locations and damage classifications among the five options. These three options can detect building footprints and classify damage into four levels. However, SE-POST and SE-Identity models are not desirable for both building localization and building damage classification, as shown in Fig. 9 (e) and (f). The models with these two options only detected rough locations of buildings, while the other three show more detailed location information. Besides, SE-POST did not detect any damages in this example. SE-Identity only detected two damage levels, including no damage and total damage. Hence, Standard SE, SE-PRE, and No-SE perform much better than SE-POST and SE-Identity.

In addition to the above qualitative analyses of image results, the results of the chosen metrics give a quantitative analysis. Combination F1 scores are analyzed first. Combination F1 scores of the five structure options are listed in Table 1. Since the model was recorded every 50 epochs, ten models were saved for each option. Only the optimal model among these ten in each structure option was used for comparison. Detailly, the model with the highest combination F1 score of each option during the test was chosen for comparing with other options' models. The epoch number listed in Table 1 is the epoch of the optimal model during the training. Scores of all ten recorded models for each option are listed in Supplementary Table 1. Combination F1 scores of the four models with SE are from 12.84% to 62.06%, as shown in Table 1. Two options, standard SE and SE-PRE, perform better than the No-SE option (the original residual block without SE). SE-PRE has the highest combination F1 score with 62.06%, which is 49.22% higher than the lowest SE-Identity with 12.84%. Hence, it could be stated that SE-PRE performs best of these five options. SE-PRE offers the best result, which is 5.41% higher than the result of the standard SE model. This may be because SE-PRE gives the channel attention before the convolution. The scores also show that the performances of SE-POST and SE-Identity are not as good as expected because their F1 scores are lower than that of No-SE.

Fig. 10 shows the combination F1 scores of the whole 500 epochs. It can be shown that standard SE and SE-PRE models have better results than No-SE results from the beginning to the end. Therefore, the adding of SE block with these two options can help to improve the model performance. However, The scores of SE-POST and SE-Identity models are much lower than that of No-SE.

After the analysis of combination F1 scores, the results of other metrics are analyzed. Fig. 11 presents the results of all the other six metrics of these five options to compare their performances. Compared with the results of all these six metrics at the damage classification step, the results at the building localization step have higher values no matter which model is used. Moreover, F1s, precisions, and recalls of SE-POST and SE-Identity models are nearly zero at the damage classification step. This reflects that these two models cannot be used for damage classification. Performances of these two models are also not good in the building localization step. The performance of the SE-Identity model is the worst among all models according to these results shown from Fig. 11 (a) to (f).

Table 1
Combination F1 scores of all five options.

Structure option	Standard SE	SE-PRE	SE-Post	SE-Identity	Original No-SE
Combination F1 Score	56.65%	62.06%	16.49%	12.84%	53.93%
Epoch	500	500	350	400	500

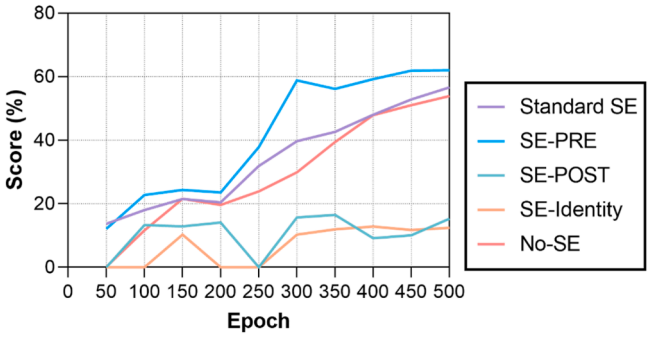


Fig. 10. Combination F1 scores of all five options at every 50 epochs.

Fig. 11 (a) shows that both standard SE and SE-PRE can achieve stable F1 scores from the 300th epoch. Their performances are better than No-SE for building localization. The results of SE-POST and SE-Identity are much lower than No-SE. Hence, only standard SE and SE-PRE models can positively influence the model performance of building localization. Localization precision results are shown in Fig. 11 (b). Similar to LF1, the values of localization precision of SE-Identity are nearly zero from the first to around the 250th epoch. Localization recalls of standard SE and SE-PRE do not increase much during the training, as shown in Fig. 11 (c). Fig. 11 (d) to (f) also show that standard SE and SE-PRE perform much better than SE-POST and SE-Identity.

Possible reasons for undesirable results of SE-POST and SE-Identity are discussed in this paragraph. The reason that SE-POST has bad results probably is that the SE block does not have the benefit or even has a bad influence on the model if it is added after ReLU. The reason for the SE-Identity model might be that the Identity block does not need channel attention. Therefore, the results show that the added attention block can help improve the accuracy of the model, but it depends on the place of the attention block.

4.2. Comparison of input size after data augmentation

In this section, the results of the standard SE model applied on two input samples of 512 × 512, and 256 × 256 sizes are presented. The method of this comparative experiment is stated in the first paragraph of Section 3.5.2. First, this section analyzes the output visualization image results. Second, the detailed results of seven metrics are analyzed to have a comparison of model performances with different input sizes.

Fig. 12 shows an example of visualization results compared with the ground truth. The ground truth in Fig. 12 (c) is the same as that in Fig. 9 (h). The results show that the standard SE model can detect building footprints and building damage levels using both input samples with different sizes. While the RGB images are useful to show the level of damages in an efficient way, quantitative performance analyses are also carried out to compare the size effect on the performance of the model, and the quantitative analyses are discussed as follows.

Table 2 shows the combination F1 score results. Similar to Table 1, the epoch listed in Table 2 is the epoch of the optimal model during the training. The highest F1 score with 512 × 512 pixels is 70.17% at the 400th epoch. The highest F1 score with 256 × 256 pixels is 56.65% at the 500th epoch. F1 with 512 × 512 pixels is higher 13% than that with 256 × 256 pixels. Hereafter, this research will use "512" and "256" to represent the two models with input sizes of 512 × 512 pixels and 256 × 256 pixels for brevity, respectively.

Fig. 13 shows the trend of combination F1s, DF1s, and LF1s. It is obvious that the "512" results are higher than the "256" results in all epochs with all three types of F1s. All LF1 lines are placed higher than DF lines, so it can be stated that this model shows better performance for building localization than building damage classification disregard of the size of images.

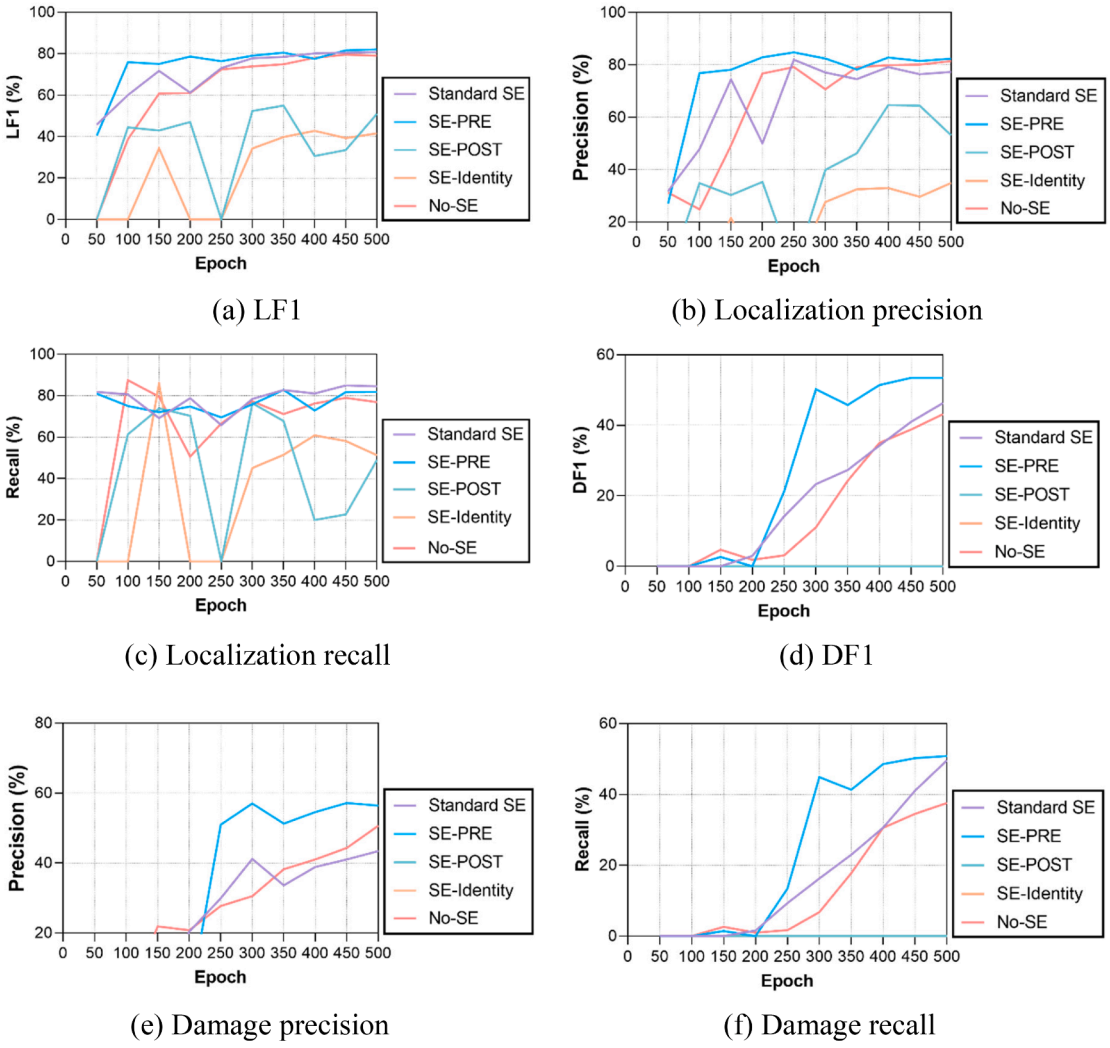


Fig. 11. F1s, precisions and recalls of all five options.

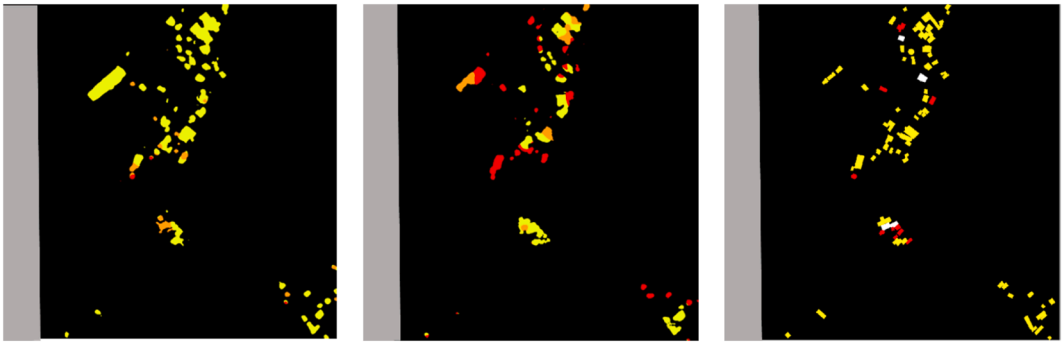


Fig. 12. Image example with different input sizes.

Table 2
Combination F1 score of each input size after data augmentation.

Input size	512 × 512	256 × 256
Combination F1 Score	70.17%	56.65%
Epoch	400	500

Fig. 14 shows the results of all the other metrics, including F1s, precisions, and recalls in both localization and damage steps. Fig. 14 (a) to (c) are localization results, and (d) to (f) are results of building damage. All these six results show the "512" model has better performance than the "256" model, because all "512" results are higher than "256" results at every recorded epoch. It should be noted that all six "512" lines turn flat, or the fluctuation has decreased since the 350th epoch. The "256" lines do not have this phenomenon. The following

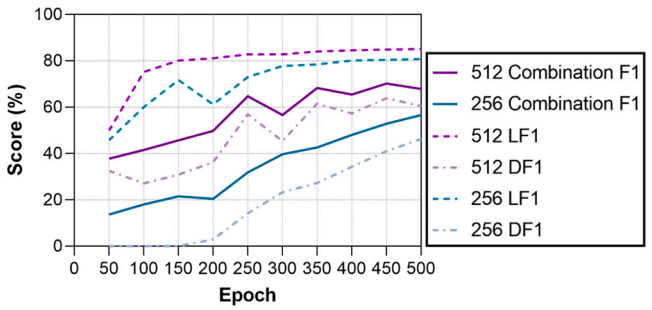


Fig. 13. Combination F1 scores, LF1s, and DF1s with two different input sizes.

paragraphs will analyze localization results first and then results on the damage.

All "256" lines have more fluctuated than "512" lines in the localization step as shown in Fig. 14 (a). LF1s are analyzed first. The highest "512" LF1 is 85.10% at the 500th epoch. LF1s from 350th to 500th are similar, which are 84.08%, 84.57%, 84.90% and 85.10%. The highest "256" LF1 is 80.79% at the 500th epoch. Similar to "512" LF1, the difference of the results from 350th to 500th is less than previous epochs (78.44%, 80.09%, 80.44%, 80.79%). There is no decrease of LF1 for both "512" and "256" results.

The second metric is localization precision. As shown in Fig. 14 (b), the line charts of both "512" and "256" localization precision results have fluctuated. The highest "512" result is 83.96% at the 500th epoch, and the highest "256" result is 81.99% at the 250th epoch.

The result of the last metric for localization is shown in Fig. 14 (c). The two localization recall lines fluctuated more than LF1 and localization precision lines. The highest "512" recall is 87.02% at the 400th epoch, and the highest "256" recall is 84.94% at the 450th epoch.

In the second step, namely, the damage classification step, all the "256" lines of the three metrics are less fluctuated than "512" lines. The trend of the DF1 line chart is similar to that of the combination F1 line chart in both "512" and "256" results. This is because the DF1 has a higher proportion than LF1 in Equation (10), which are 70% and 30%, respectively. The highest "512" DF1 is 63.86% at the 450th epoch, and the highest "256" DF1 is 46.31% at the 500th epoch. There is no decrease of "256" results, while "512" results show the model performance is increasing in a fluctuation.

As for the damage precision results, "256" line increases more slowly than "512" line, which is 0 from 0 to 150 epochs. The trend of "512" damage precision is similar to the trend of "512" DF1, especially from the 200th epoch. The highest "512" damage precision is 65.27% at the 450th epoch, and the highest "256" damage precision is 43.41% at the 500th epoch.

The trends of two damage recall lines in Fig. 14 (f) are very similar to the trends of two DF1 lines in Fig. 14 (d). The highest "512" recall is 62.5% at the 450th epoch, and the highest "256" recall is 49.63% at the 500th epoch.

The model of 512 × 512 pixels at the 450th epoch could be said to be the best model among all 512 × 512 pixels, because its value is the highest for five out of the seven results. The best localization model of 512 × 512 pixels could be said between 450 and 500 epochs since their values are quite close to each other. The optimal damage classification model of 512 × 512 pixels is the model training with 450 epochs among all recorded models because its performances are the best with all damage classification metrics, including DF1, damage precision, and damage recall. As for the models of 256 × 256 pixels, the model at 500th epoch is the best both in the localization and damage classification since it performs best with all the metrics. Moreover, the results show that localization results are higher than damage classification results with all metrics.

4.3. Comparison of transfer learning and non-transfer learning

This section compares the results of transfer learning and non-transfer learning models with 512 × 512 pixels. The method of this experiment is introduced in the second paragraph of Section 3.5.2. Combination F1 score results are discussed first, as shown in Table 3. The highest F1 score with transfer learning is 69.85% at the 400th epoch, and that with non-transfer learning is 70.17% at the 450th epoch. The detailed information is listed in Supplementary Table 3. Fig. 15 (a) shows the combination F1 scores of these models. The results reflect that the transfer learning model does not have obvious advantages over the

Table 3
Combination F1 scores with and without transfer learning.

Model	Transfer learning	Non-transfer learning
Combination F1 Score	69.85%	70.17%
Epoch	400	450

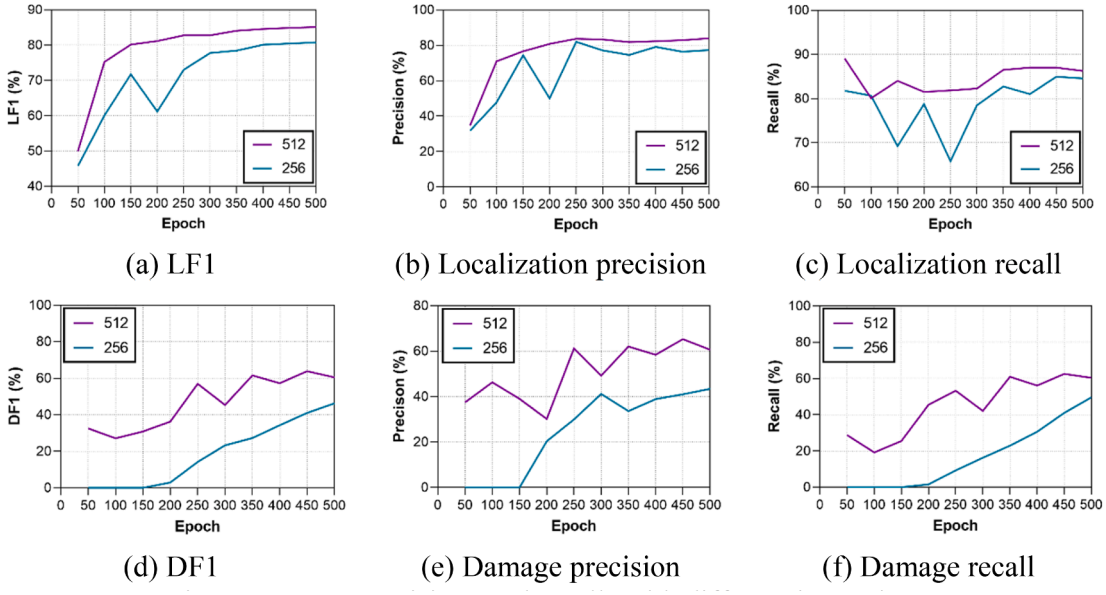


Fig. 14. F1s, precisions and recalls with different input sizes.

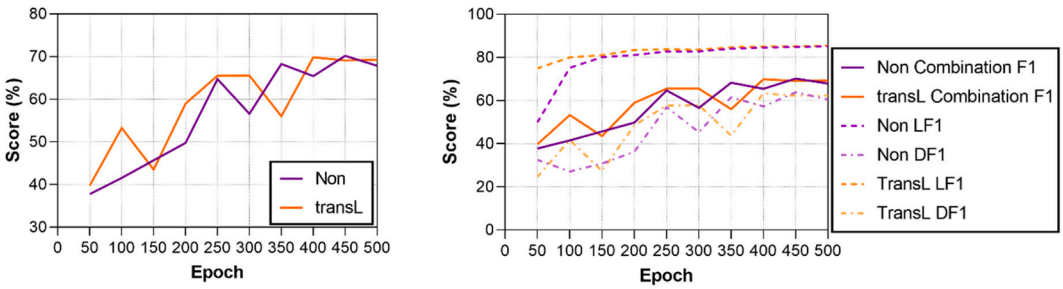


Fig. 15. F1 scores of transfer and non-transfer learning models.

non-transfer learning model. In Fig. 15 (b), "Non" represent the trained model without transfer learning. "TransL" means that the model is trained with transfer learning. Fig. 15 (b) shows that these two models have similar performance for building localization. The model with transfer learning has better performance at the initial stage, but their performances turn similar gradually. That is, the highest score without transfer learning is even higher than that with transfer learning.

Fig. 16 shows the detailed results at both localization and classification steps. Fig. 16 (a) to (c) display building localization results. Fig. 16 (a) shows LF1s of these two methods. If the number of training epochs is less than 300, the benefit with the pre-trained model is obvious, but after 300 epochs, the LF1s difference is less than 1% in each recorded epoch. The highest LF1 with transfer learning is 85.47%, which is 0.37% higher than the highest LF1 without transfer learning (85.10%) as shown in Supplementary Table 3. Fig. 16 (b) and (c) show the localization precisions and recalls, respectively. The trends of these results are similar to the trend of Fig. 16 (a), whose difference is very large at the beginning, but the difference decreases quickly. All the results of these three metrics show that 350 training epochs may be enough no matter the model is with transfer learning or not. This is because the results improve slowly after 350 epochs.

Fig. 16 (d) to (f) show damage classification results. The trends of these results are different from the trends in the building localization step. The initial values of the three metrics are not much different, and the results of the non-transfer learning model are even better than those of the transfer learning model at the 50th epochs. The highest DF1, precision, and recall with transfer learning are 63.36% at the 400th epoch, 67.96% at the 400th epoch, and 60.40% at the 450th epoch. The highest DF1, precision, and recall without transfer learning is 63.86%, 65.27%, and 62.50%, respectively, all at the 450th epoch.

Based on the interpretation of all the results of these seven metrics, the transfer learning model with pre-trained model is not much better than the non-transfer learning model. One benefit is that it can achieve higher precision than the non-transfer learning model at the beginning for only building localization, which is faster, but the non-transfer learning model can also achieve this high precision after around 100 epochs. The transfer learning model does not show many advantages in damage classification.

The possible reason for that is listed as follows. First, the chosen pre-trained model is for image classification. That is why the pre-trained model has better performance of building localization than damage classification. This pre-trained model may not be the best choice for damage classification. Second, the pre-trained model is trained with ImageNet dataset. This dataset does not contain enough damaged buildings in the images. Hence, the pre-trained model does not show enough good performance as anticipated.

Although several papers adopt this pre-trained model based on experience (Koo et al., 2020), they do not compare its results with the non-transfer learning model. Before the testing, the hypothesis in this study also is that its performance is better, but the test shows that is not the case.

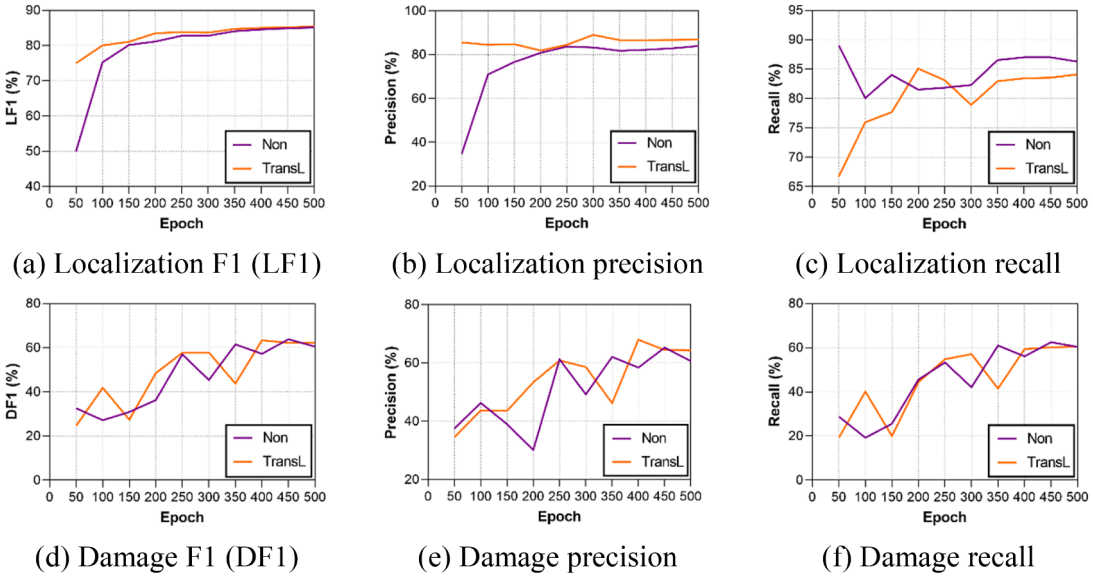


Fig. 16. F1s, precisions and recalls of transfer and non-transfer learning models.

4.4. Comparison of Sigmoid and Hard-Sigmoid activation functions

As the fourth comparative experiment of the testing stage in the workflow, this section compares the performance of models with different activation functions placed in the SE block as mentioned in the third paragraph of Section 3.5.2. Results of the seven metrics are analyzed one by one.

The combination F1 score results are compared first with the optimal model of each situation. As shown in Table 4, the optimal model with Sigmoid function is recorded at the 450th epoch with the highest F1 score 70.17%. The F1 score of the optimal model with Hard-Sigmoid is 69.66%, which is at the 500th epoch. Supplementary Table 4 shows the detailed results with different functions. Hence, the performances of models with Sigmoid and Hard-Sigmoid functions are similar, as shown in Fig. 17 (a). The performance of the Sigmoid model with Sigmoid is 0.51% higher than the Hard-Sigmoid model. Fig. 17 (b) shows that building localization results are much better than damage classification results no matter which activation function is used.

Fig. 18 shows all results at the localization and classification steps. The highest LF1 with Sigmoid function is 85.10% at the 500th epoch, and the highest LF1 with Hard-Sigmoid function is 84.05% at the 400th epoch, as shown in Supplementary Table 4. Fig. 18 (a) shows that the Sigmoid model performs better than Hard-Sigmoid model from the 300th epoch. The results of another metric, localization precision, are shown in Fig. 18 (b). Its highest is 83.96% at the 500th epoch with Sigmoid and 83.69% at the 450th epoch with Hard-Sigmoid. Fig. 18 (a) and (b) show that the trends of the two lines are similar. Despite the fact that the performance of Hard-Sigmoid model is better at the beginning epochs, the performance of Sigmoid model increases quicker and better than that of Hard-Sigmoid model at last. Fig. 18 (d) shows the localization recall results. The highest recalls are 87.02% with Sigmoid and 86.36% with Hard-Sigmoid, both at the 400th epoch. The recall of the Sigmoid model is 0.66% higher than that of the Hard-Sigmoid model.

Fig. 18 (d) to (f) show the results at the damage classification step. The highest DF1 is 63.86% at the 450th epoch with Sigmoid, which is 0.36% higher than that with Hard-Sigmoid (63.50% at the 500th epoch). The highest damage precisions are 65.27% at the 450th epoch with Sigmoid and 68.87% at the 500th epoch with Hard-Sigmoid shown as Fig. 18 (e). The highest damage recalls are 62.50% at the 450th epoch with Sigmoid and 60.72% at the 400th epoch with Hard-Sigmoid. Therefore, the damage classification results show that the performances of these two models are similar.

Based on the analysis, the Hard-Sigmoid model only has better results based on the damage recall metric, and the other five results show that the performance of the Sigmoid model is better. Therefore, it can be concluded that the model with Sigmoid performs slightly better than the model with Hard-Sigmoid.

5. Discussion

Four types of experiments have been conducted for different comparison tasks. The first experiment tested the performance of five model options, including four modified models (SE, SE-PRE, SE-POST, SE-Identity) and one original model. In the second experiment, two models were tested for the performance comparison of different training input sizes. The next experiment was implemented for the comparison of models with and without transfer learning. The last experiment was implemented for comparison of different activation functions in the SE block. The results of these four experiments are discussed as follows.

Table 4
Combination F1 scores with Sigmoid and Hard-Sigmoid activation functions.

Activation function	Sigmoid	Hard-Sigmoid
F1 Score	70.17%	69.66%
Epoch	450	500

First, the outcome of the experiment with five options (refer Section 4.1) shows that the SE-PRE model has the best performance across all metrics except localization recall among all these five models. One possible reason is that SE-PRE gives channel attention at the very beginning. Hence, the model could judge which feature needs more attention during the training and which feature is less important. The worst two models are SE-Post and SE-Identity because they perform worst among all metrics, and their performances are much worse than the performances of the other two modified models and the original No-SE model. Therefore, although several papers state that a model would perform better after adding channel attention (Li, Tian et al., 2020; Li, Wang et al., 2020), the finding from this study is that different added places of SE in the original basic block have different effects on the model performances of building damage classification, sometimes improves the accuracy and sometimes decrease the accuracy.

In the second experiment, the results show that the standard SE model (Fig. 8 (a)) with the input resolution of 512 × 512 by random cropping during the training can give better performance than that with 256 × 256. This might be because a 512 × 512 image contains more features than a 256 × 256 image. Hence, more information can be retained during the training with 512 × 512 size. Similarly, higher resolution images could have better results for models because they contain more features and information than lower resolution images. However, the larger size 512 × 512 spends more time and memory for training, so which size to be adopted depends on the time and hardware constraints of a task.

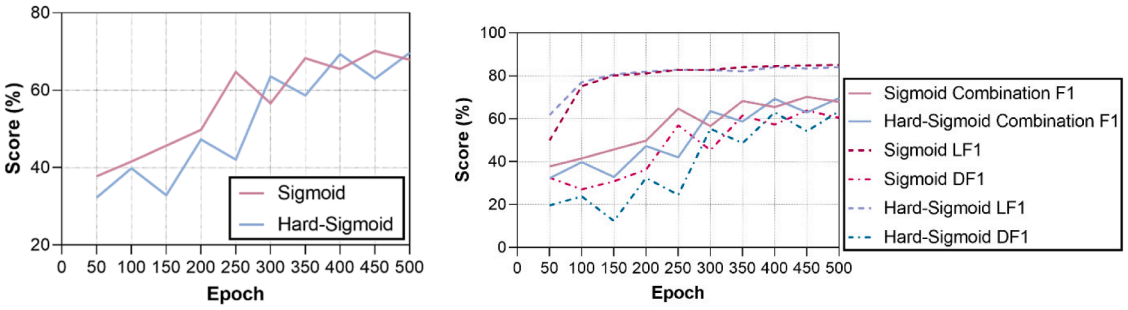
In the third comparison, the results of the models with and without transfer learning are compared. The model with transfer learning of pre-trained ImageNet weights does not perform better than model training from scratch. This may be because the pre-trained weights are not very suitable for this study since ImageNet does not contain damaged building labels. Moreover, the pre-trained weight is only used for image classification, whose task is easier for this study, including both image segmentation and classification.

Fourth, the difference in the performance of Sigmoid and Hard-Sigmoid is not obvious. Although the training time of the model with Hard-Sigmoid should be shorter than that with Sigmoid in theory, the time used in this experiment is similar.

The model performs better for building localization than damage classification. The technical reason might be that the building localization step is easier than damage classification. It only needs to segment two classes, buildings, and non-buildings, and the difference of features between these two classes are obvious such as outlines or colors. However, the feature difference of each damage level is small and complex between minor, major, and total damage. Moreover, the roof shapes of some totally damaged buildings did not change or break after disasters, especially after earthquakes. Hence, it increases the difficulty for models to assess the damage levels of buildings.

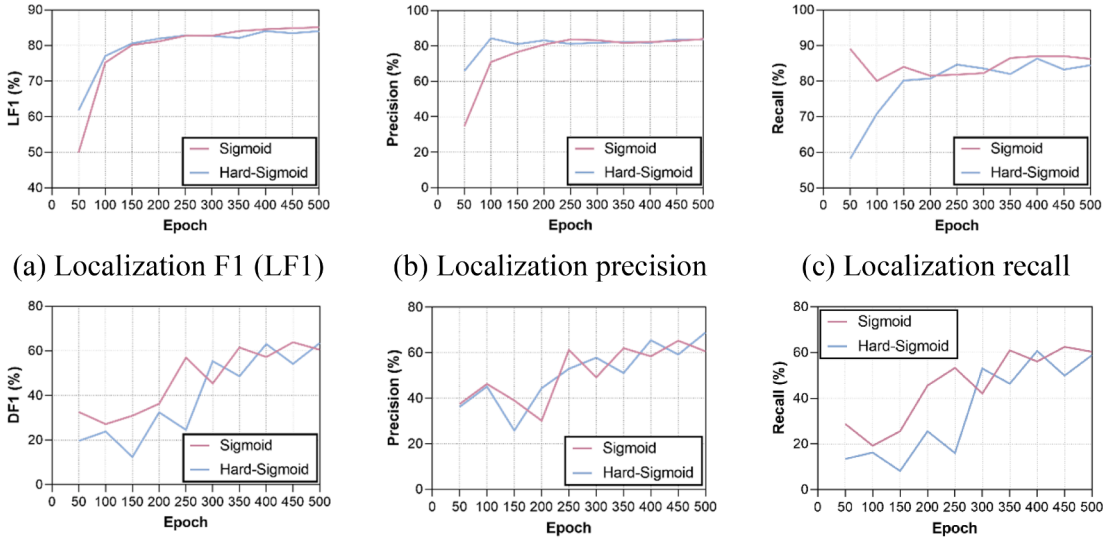
While the main target of this paper is building damage classification, the observation shows that the SE-PRE also offers a high accuracy (82.07% of LF1 score) for building localization or building footprint segmentation. Offering acceptable outcomes for both building damage classification and localization will increase the practical implication of the SE-PRE model to guide disaster managers and practitioners for emergency actions by finding the most damaged buildings and their locations simultaneously.

This paper contributes to the body of literature by addressing a challenging task of building damage classification utilizing the novel approach of SE-PRE. Compared to the previous work, the present paper shows an improvement in the classification tasks by classifying into not only collapse or not, but also four damage levels and providing a new modified model. Several current papers just apply models to the building damage classification without modification of the structures of models. For example, Yang et al. (2021) classified the damages keeping the original structures of the convolutional neural network (CNN).



(a) Combination F1 Score (b) Combination F1 scores, LF1s, and DF1s

Fig. 17. F1 scores of models with different SE block activation functions.



(a) Localization F1 (LF1) (b) Localization precision (c) Localization recall

(d) Damage F1 (DF1) (e) Damage precision (f) Damage recall

Fig. 18. F1s, precisions and recalls with different SE block activation functions.

6. Conclusion

The aim of this paper was to provide a quick deep learning method for post-disaster building multi-level damage classification with optical satellite images. This has been achieved by improving the performance of the deep learning method with SE added dual HRNet model and applying it on building damage classification with a total of 8,664 images from xBD and our newly created datasets. Four novel options of models with adding SE channel attention to different places of basic residual blocks in HRNet have been used to compare the original model without SE. These four are called standard SE, SE-PRE, SE-Post and SE-Identity. Four types of experiments applying seven metrics (refer Section 3.4) were implemented on different model modifications to measure the effect on the model performance.

The results show that the deep learning model proposed in this paper can classify building damage levels, which are hard or impossible to be achieved by human eyes. Four options of SE channel attention added models are tested, and results show that SE-PRE has the best performance. A larger input size can have better results but use much more computing time. Transfer learning with pre-trained ImageNet dataset does not have advantages because the dataset does not contain several damaged building images. The block with Sigmoid function has slightly better performance than that with Hard-Sigmoid. This paper also creates a building damage level dataset based on the official damage assessment document.

Some limitations exist in the dataset. One limitation is that the difference between pre- and post-event images is large. Some post-event images are taken several months after the event happened. The imaging angle and the brightness are different between these two images. This limitation would not affect the comparison outcome of the experiment but cannot be avoided because this is very tough for remote sensing technology to take two images with the exact same air, sunlight, and imaging angle conditions on two different days. In the future, scholars can replicate the suggested method on new datasets with the development of remote sensing to avoid this limitation. Second, some pre- or post-event images contain a large area of clouds covering buildings, so the model will wrongly learn the white cloud area as buildings based on the ground truth map. Hence, future work would focus on reducing the limitations of data.

CRediT authorship contribution statement

Chang Liu: Conceptualization, Methodology, Software, Data curation, Formal analysis, Writing – original draft, Visualization. **Samad M. E. Sepasgozar:** Conceptualization, Methodology, Writing – review & editing. **Qi Zhang:** Visualization, Writing – review & editing. **Linlin Ge:** Resources, Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Funding: This work was supported by SmartSat Cooperative Research Centre (grant number P2.30s).

The authors would also be grateful to the Operational Satellite Applications Programme, Joint Research Centre and World Bank for the 2010 Haiti building damage assessment shapefiles.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2022.117268.

References

Berman, M., Triki, A. R., & Blaschko, M. B. (2018). The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4413–4421).

Courbariaux, M., Bengio, Y., & David, J.-P. (2015). Binaryconnect: Training deep neural networks with binary weights during propagations. *Advances in Neural Information Processing Systems, 28*, 3123–3131. https://proceedings.neurips.cc/paper/2015/file/3e15cc11f979ed25912dff5b0669f2cd-Paper.pdf.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition, 2009*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

DesRoches, R., Comerio, M., Eberhard, M., Mooney, W., & Rix, G. J. (2011). Overview of the 2010 Haiti earthquake. *Earthquake Spectra, 27*(1_suppl1), 1–21. https://doi.org/10.1193/1.3630129

DIUx. (2019). *xView2 Scoring*. In https://github.com/DIUx-xView/xView2_scoring/blob/master/xview2_metrics.py.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*. https://doi.org/10.48550/arXiv.2010.11929.

Dumoulin, V., & Visin, F. (2016). A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*. https://doi.org/10.48550/arXiv.1603.07285.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., … Gaston, M. (2019). Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 10–17).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning, 37*, 448–456. https://proceedings.mlr.press/v37/ioffe15.html.

Ji, M., Liu, L., & Buchroithner, M. (2018). Identifying collapsed buildings using post-earthquake satellite imagery and convolutional neural networks: A case study of the 2010 Haiti earthquake. *Remote Sensing, 10*(11), 1689. https://doi.org/10.3390/rs10111689

Kolar, Z., Chen, H., & Luo, X. (2018). Transfer learning and deep convolutional neural networks for safety guardrail detection in 2D images. *Automation in Construction, 89*, 58–70. https://doi.org/10.1016/j.autcon.2018.01.003

Koo, J., Seo, J., Yoon, K., & Jeon, T. (2020). *Dual-HRNet for building localization and damage classification* [Unpublished manuscript]. https://github.com/DIUx-xView/xView2_fifth_place/blob/master/figures/xView2_White_Paper_SI_Analytics.pdf.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Krupiński, M., Lewiński, S., & Malinowski, R. (2019). One class SVM for building detection on Sentinel-2 images. *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments, 2019*(11176), 1117635. https://doi.org/10.1117/12.2535547

Li, L., Tian, T., Li, H., & Wang, L. (2020). SE-HRNet: A deep high-resolution network with attention for remote sensing scene classification. In *IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium* (pp. 533–536). https://doi.org/10.1109/IGARSS39084.2020.9324633

Li, Y., Wang, C., Cao, Y., Liu, B., Luo, Y., & Zhang, H. (2020). A-HRNet: Attention based high resolution network for human pose estimation. *Second International Conference on Transdisciplinary AI (TransAI), 2020*, 75–79. https://doi.org/10.1109/TransAI49837.2020.00016

Liu, C., Sepasgozar, S. M., Shirowzhan, S., & Mohammadi, G. (2021). Applications of object detection in modular construction based on a comparative evaluation of deep learning algorithms. *Construction Innovation, 22*(1), 141–159. https://doi.org/10.1108/CI-02-2020-0017

Majd, R. D., Momeni, M., & Moallem, P. (2019). Transferable object-based framework based on deep convolutional neural networks for building extraction. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12*(8), 2627–2635. https://doi.org/10.1109/JSTARS.2019.2924582

Maxar. (2010). *Open Data Program: Haiti Earthquake*. https://www.maxar.com/open-data/haiti-earthquake.

Paszke, A., Gross, S., Chintala, S., & Chanan, G. (2021). *Pytorch-BatchNorm2d*. Facebook. Retrieved June 30, 2021 from https://pytorch.org/docs/stable/generated/torch.nn.BatchNorm2d.html.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. https://doi.org/10.48550/arXiv.1409.1556.

Su, J., Bai, Y., Wang, X., Lu, D., Zhao, B., Yang, H., … Koshimura, S. (2020). Technical solution discussion for key challenges of operational convolutional neural network-based building-damage assessment from satellite imagery: perspective from benchmark xBD dataset. *Remote Sensing, 12*(22), 3808. https://doi.org/10.3390/rs12223808

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5693–5703).

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., & Wang, J. (2019). High-resolution representations for labeling pixels and regions. *arXiv preprint*. https://doi.org/10.48550/arXiv.1904.04514.

Tanjung, J., Sanada, Y., Nugroho, F., & Wardi, S. (2020). Seismic analysis of damaged buildings based on postearthquake investigation of the 2018 Palu Earthquake. *International Journal of GEOMATE, 18*(70), 116–122. https://doi.org/10.21660/2020.70.9490

UNITAR/UNOSAT/EC/JRC/WB. (2010). *Port-au-Prince Atlas of Building Damage Assessment*. http://www.unitar.org/unosat/node/44/1425.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., … Wang, X. (2020). Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(10), 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686

Wheeler, B. J., & Karimi, H. A. (2020). Deep learning-enabled semantic inference of individual building damage magnitude from satellite images. *Algorithms, 13*(8), 195. https://doi.org/10.3390/a13080195

Yang, W., Zhang, X., & Luo, P. (2021). Transferability of convolutional neural network models for identifying damaged buildings due to earthquake. *Remote Sensing, 13*(3), 504. https://doi.org/10.3390/rs13030504