BDANet: Multiscale Convolutional Neural Network With Cross-Directional Attention for Building Damage Assessment From Satellite Images

Yu Shen, Sijie Zhu, Taojiannan Yang, Graduate Student Member, IEEE, Chen Chen, Member, IEEE, Delu Pan, Jianyu Chen, Member, IEEE, Liang Xiao, Member, IEEE, and Qian Du, Fellow, IEEE

Abstract—Fast and effective responses are required when a natural disaster (e.g., earthquake and hurricane) strikes. Building damage assessment from satellite imagery is critical before relief effort is deployed. With a pair of predisaster and postdisaster satellite images, building damage assessment aims at predicting the extent of damage to buildings. With the powerful ability of feature representation, deep neural networks have been successfully applied to building damage assessment. Most existing works simply concatenate predisaster and postdisaster images as input of a deep neural network without considering their correlations. In this article, we propose a novel two-stage convolutional neural network for building damage assessment, called BDANet. In the first stage, a U-Net is used to extract the locations of buildings. Then, the network weights from the first stage are shared in the second stage for building damage assessment. In the second stage, a two-branch multiscale U-Net is employed as the backbone, where predisaster and postdisaster images are fed into the network separately. A cross-directional attention module is proposed to explore the correlations between predisaster and postdisaster images. Moreover, CutMix data augmentation is exploited to tackle the challenge of difficult classes. The proposed method achieves state-of-the-art performance on a large-scale dataset—xBD. The code is available at https://github.com/ShaneShen/BDANet-Building-Damage-Assessment.

Index Terms—Building damage assessment, convolutional neural network (CNN), cross-directional attention (CDA), CutMix, multiscale feature fusion (MFF), satellite image.

I. INTRODUCTION

NATURAL disasters, such as earthquakes, floods, and tsunami, can cause serious social and economic devastation. When a natural disaster strikes, accurate and immediate responses are required in humanitarian assistance and disaster response (HADR) for saving thousands of lives [1]–[4]. Before these responses, rescue planning and preparations are conducted based on the damage analysis. With the rapid development of remote sensing technology, high-resolution satellite images are now available for damage assessment. Traditionally, these images of disaster areas are analyzed by experts, which may be time-consuming if the areas are large. Therefore, automatic information extraction from satellite images, such as building segmentation and damage assessment, is imperative under time-critical situations.

Building damage assessment plays a pivotal role in HADR, which aims at predicting the damage level for each pixel based on building segmentation. It can be divided into two categories according to the use of predisaster images. For the methods only using postdisaster images, the damage assessment is considered as a segmentation task [5], [6]. However, without predisaster images, building assessment errors may be inevitable since postdisaster images cannot provide precise contours of intact building objects. With predisaster images, well-shaped buildings can be utilized for damage assessment. Note the problem of damage assessment with paired images shares some similarities with change detection, as both tasks aim to find the changed regions according to two images acquired at different time [7]–[9]. Compared with change detection, building damage assessment is a more challenging task because it is required to classify different damage levels. Moreover, the unchanged objects (undamaged buildings) are required in building damage assessment, which is not considered in change detection.

Recently, deep learning-based methods, such as convolutional neural networks (CNNs), have shown their effectiveness...
image features. Many researchers have leveraged deep neural networks from low level to high level, without selecting handcrafted methods can automatically learn image feature representations in various tasks. With a set of labeled data, deep learning-based networks for building damage assessment and achieved significant progress [10]–[12]. Xu et al. [11] investigated the capability of CNN for building damage detection by identifying damaged and undamaged buildings. Weber and Kane [12] considered building damage assessment as a semantic segmentation task, in which damage levels are assigned to different classes.

With a pair of predisaster and postdisaster images for building damage assessment, a key question is how to effectively model the correlations between these images. Standard practice is to use a two-branch CNN-like architecture with feature fusion schemes. For instance, Hao et al. [13] concatenated the features from predisaster and postdisaster images and fed them into a CNN-based framework. Gupta and Shah [14] developed a framework that uses the difference between predisaster and postdisaster features as input of the deep neural network, which is denoted as RescueNet. More recently, an attention mechanism has been used with deep neural networks for remote sensing images processing [15], [16], which is a strategy of allocating larger weights to informative parts of an image/feature. For instance, the average pooling is applied to extract the channel attention information for remote sensing image segmentation [17], and a nonlocal-based attention module is utilized in the network to explore the spatial relations of satellite images in [13].

Another challenge of building damage assessment from satellite imagery lies in the visual similarity between certain classes (e.g., no damage and minor damage). These classes are considered difficult classes. To illustrate this problem, we use the xBD [18] dataset that is the largest dataset for building damage assessment to date, as an example. Fig. 1 shows a pair of images from this dataset. Based on visual observation, it is difficult to distinguish between classes, such as no damage and minor damage, due to a similar appearance. Table I reports the classification results of ResNet-50 on xBD. From the classification confusion matrix, about 24.2% of minor damage are misclassified as no damage. One effective strategy to cope with difficult classes and improve the model performance is data augmentation [19], [20]. Data augmentation has been widely used as a preprocessing technique to artificially increase the size of the dataset in segmentation tasks [21]. Recently, CutMix [22] is proposed as a new data augmentation technique, which generates a new image by combining two image samples, to enhance the generalization ability of neural networks. CutMix directly cuts and pastes image patches from one image to another, which can be easily used in many tasks.

Motivated by the above observations, we introduce a two-stage CNN-based framework, named BDANet, for building damage assessment. First, a single U-Net [23] is used for building segmentation (Stage 1). U-Net is an encoder–decoder network architecture that has been widely used in segmentation tasks, e.g., in [24]. The results of building segmentation are used to guide the locations of building damages in the second stage. Then, a two-branch multiscale CNN (U-Net structure) is applied for damage assessment (Stage 2). By using the network weights from building segmentation, the network in Stage 2 can be trained more efficiently. Due to the scale variance of building objects, we introduce a multiscale feature fusion (MFF) module in the encoder layer, which enables the network to aggregate features from multiple scales. In addition, a cross-directional attention (CDA) module is proposed to explore the correlations between features from predisaster and postdisaster images. By leveraging channel-wise and spatialwise correlations, the network can fuse and enhance the features from predisaster and postdisaster images. Moreover, to tackle difficult classes, CutMix is employed for data augmentation. Specifically, we only apply CutMix to difficult classes to make the network pay more attention to those classes, thereby learning more robust representations for their separation. In the experiments, we show that this strategy yields superior classification performance over simply adopting CutMix for all classes.

The contributions of this article are summarized as follows.

1) We present a two-stage CNN-based learning framework for building segmentation (Stage 1) and damage assessment (Stage 2). The building segmentation helps the network to locate the building objects. The proposed framework achieves state-of-the-art performance on a
large-scale building damage assessment benchmark—xBD [18].

2) To reduce the influence of scale variance of buildings, an MFF module is introduced to learn features from multiple scales.

3) A CDA module is proposed to effectively aggregate the features from predisaster and postdisaster images. With the attention module, informative channel and spatial information can be extracted to enhance the feature representation.

4) We also unveil the challenge of difficult classes in building damage assessment and explore an effective data augmentation strategy, CutMix [22], to cope with this challenge.

The remainder of this article is organized as follows. Section II gives an overview of related works. Section III presents the details of the proposed BDANet for building damage assessment. Section IV provides comprehensive experimental results and parameter analysis. Finally, the conclusion is drawn in Section V.

II. RELATED WORK

A. Building Damage Assessment

Building damage-based analysis, including detection, segmentation, and assessment, are vital topics in HADR. Recently, many algorithms have been developed for building damage analysis [25]–[27]. For instance, in [26], to detect the damaged building objects after an earthquake, change features, including texture, color, statistical similarity, and correlation descriptors are all analyzed before applying a support vector machine (SVM) classifier. Zhu et al. [28] proposed an instance segmentation network, named MSNet, for building damage detection. In [29], a multiple kernel learning framework is proposed. To improve the performance of building damage detection, the CNN feature and 3-D cloud point feature are integrated into the framework. In [30], an unsupervised anomaly detection method is developed to automatically detect the location of damaged buildings.

Different from building damage detection and segmentation, building damage assessment aims at identifying the damage level of building objects after natural disasters. Depending on whether to use predisaster images, building damage assessment can be divided into two categories. One category only uses postdisaster images, and the other uses paired predisaster and postdisaster images for damage assessment. Many previous works focus on using only postdisaster images. For example, in [31], SVM is used to assess the building damage after the earthquake with QuickBird satellite images. In [32], the object-based analysis is integrated with machine learning methods to preserve the boundary and detailed information of remote sensing satellite images in damaged buildings. However, without predisaster images, building assessment errors can be easily made because postdisaster images cannot provide precise boundaries/contours of complete building objects. Therefore, researchers resort to using a pair of satellite images (predisaster and postdisaster) for building damage assessment. Cooner et al. [33] extracted textural and structural features from both predisaster and postdisaster images, which improved the performance of earthquake damage assessment. Due to the limited amount of labeled data, some works simplify the damage assessment task as binary classification, which assigns damage and no damage labels to building objects. For example, Xu et al. [11] compared the performance of different CNN models in detecting damaged buildings in the Haiti earthquake.

Recently, researchers start to differentiate more building damage levels for better disaster analysis. Different from the binary classification, the extent of damage is considered, such as minor damage and major damage. In [6], the level of building damage after the earthquake is evaluated by a framework integrating CNN and ordinal regression. In [34], to reduce the uncertainty in predicting building damage levels, building objects in the center of an image patch are considered, while the surrounding buildings are occluded by negating the pixels. Hence, the model can focus on the buildings in the center of an image patch.

B. U-Net

U-Net is a CNN-based network that was first proposed for biomedical image segmentation [23]. Compared with traditional CNN-based architectures, U-Net introduced downsampling and upsampling to aggregate features from multiple scales. By adopting the skip connection, features from low and high levels are integrated to enhance the learning ability of the network [35]. Due to these advantages, U-Nets have been widely used in image segmentation [36] and classification [37].

U-Net-based architectures have also been applied in remote sensing image processing. For instance, in [38], U-Net is introduced for very-high-resolution remote sensing image segmentation. In [39], U-Net is integrated with multiscale strategy for building instance extraction. In [40], U-Net is used as a feature extraction method for building segmentation, which provides promising results. Moreover, U-Net-based architectures have been applied in a variety of tasks, such as road extraction [41] and change detection [42]. However, the U-Net architecture is simple and cannot be directly used in building damage assessment.

C. Attention-Based Models

Attention-based models [43], [44] have been widely used in computer vision tasks, which learn to put more emphasis on important information of images/features. Recently, many attention-based models have been developed and achieved great success in remote sensing image processing tasks, such as image classification and segmentation [39], [45]. In [17], a channel-attention mechanism is integrated with a fully convolutional network (FCN) for high-resolution aerial image segmentation. In [46], a foreground-aware network is developed, which incorporates both spatial and channel attention models to improve foreground modeling for geospatial object segmentation. In [47], an attention embedding module is designed to embed attention from high-level layers into low-level ones.
Fig. 2. Overview of the proposed framework (BDANet). The U-Net structure is used in both stages. It consists of two stages: (a) Stage 1: building segmentation and (b) Stage 2: damage assessment.

to enrich the semantic information for satellite image segmentation. In [48], a self-attention mechanism that considers both spatial-dipartite context aggregation information and the relative position encoding information is proposed for remote sensing scene classification.

Attention-based models are barely applied in building damage assessment. In [13], a nonlocal attention model [44] is adopted to capture long-range spatial information of predisaster and postdisaster images. However, the computational cost of the nonlocal mechanism is very high because the attention map is calculated based on high-resolution features.

D. Data Augmentation Strategies

Data augmentation has been widely used as a preprocessing technique to artificially increase the size of the dataset in computer vision tasks [21], [49], [50]. Data augmentation methods usually employ image transformations, such as image flip, rotation, and crop, to improve the model generalization performance. Manually annotating remote sensing images is difficult and time-consuming. As a result, labeled remote sensing data are usually limited. To facilitate the training of deep neural networks, data augmentation is critical in deep learning-based remote sensing image processing.

Recently, various data augmentation techniques have emerged and shown their efficiency. Cutout [51] randomly masks out square regions of input to force the network to pay attention to diverse image regions, thereby improving the feature representation capability of CNNs. Some works focus on synthesizing image samples to expand the size of training data. Cubuk et al. [20] proposed a search algorithm for image augmentation, which enables the network to find the best validation accuracy. In [52], Mixup is introduced to regularize the neural network by using a linear combination of pairs of samples and labels. Different from the Mixup strategy that merges two images with the linear summation, CutMix [22] directly cuts and pastes image patches from one image to another, which can be used in segmentation tasks.

At present, most deep learning-based methods still follow the architectures of change detection and seldom consider the completeness of building objects. Furthermore, due to the similarities between certain classes, it is difficult for the network to distinguish their differences. To deal with these problems, a two-stage framework is proposed for building damage assessment. The MMF and CDA models are proposed to enhance the feature representation. Moreover, the CutMix strategy is employed to alleviate the difficulty of classifying hard classes.

III. PROPOSED METHOD

Fig. 2 presents an overview of the proposed BDANet, which consists of two stages: building segmentation (Stage 1) and damage assessment (Stage 2). In Stage 1, a single U-Net branch (i.e., the upper one) is used for building segmentation. This U-Net branch uses only predisaster images as input and produces segmentation masks of building objects. In Stage 2, the predisaster and postdisaster images are fed into two
network branches separately. To further enhance the feature representations, the MFF module and the CDA module are introduced. We describe the details of the proposed damage assessment framework and its components in the following.

A. Two-Stage Framework

Although the building damage assessment task shares some similarities with the change detection task, there are several key differences. First, building damage assessment not only requires detecting the changes but also needs to classify their damage levels. Second, the undamaged buildings also need to be detected, which are usually unnecessary in change detection. Change detection-based networks mainly follow a three-branch architecture, which contains a dual-branch for image pairs learning and a fusion branch for change detection [53], [54]. These networks are designed for learning the differences of image pairs, while unchanged objects (e.g., undamaged buildings) are ignored. Due to these problems, directly using networks for change detection cannot satisfy the needs of building damage assessment. Therefore, in this article, a two-stage learning framework is proposed for building damage assessment.

Stage 1 (Building Segmentation): As shown in Fig. 2(a), to locate the building objects, only predisaster images are used as input to the network in this stage. The network is separately trained for building object segmentation in this stage. Compared with postdisaster images, building objects in predisaster images are intact and well-shaped. The network is designed based on the architecture of U-Net [23]. In the encoder part, a ResNet [55] is used as the backbone to extract deep features. By using the strided convolution, the input image is first downsampled into half of the original resolution. With the network becoming deeper, feature maps are projected into a higher dimension for high-level feature extraction. In the decoder part, the resolution of features is gradually recovered to the original input size by upsampling. To capture the details of target objects, skip connections are utilized in the network [see Fig. 2(a)]. Before each upsampling operation, the low-level feature maps from the encoder layer are concatenated with the high-level feature maps. The output of the network is a binary segmentation map indicating the building distribution in a satellite image. Similar to some segmentation tasks, the cross-entropy loss is used in this stage, which is formulated as

\[ L = - \sum_{i} y^{i} \log \hat{y}^{i} + (1 - y^{i}) \log (1 - \hat{y}^{i}) \]  

(1)

where \( L \) quantifies the overall loss by comparing the target label \( y^{i} \) with predicted label \( \hat{y}^{i} \) and \( N \) represents the number of training pixels.

The objective of Stage 1 is locating the building objects. As many buildings are damaged in postdisaster images, it is difficult for the network in Stage 2 to perform building segmentation and damage assessment at the same time. Therefore, to guide the assessment process of Stage 2, a single U-Net using only predisaster images is first applied to identify the locations of buildings.

Stage 2 (Damage Assessment): In Stage 2, both predisaster and postdisaster images are fed into a two-branch network. The backbone network for damage assessment is the same as that for building segmentation. The convolutional kernels are shared in both branches. Therefore, the parameter weights obtained from Stage 1 are used here for weights initialization.

As shown in Fig. 2(b), the feature maps from two branches are concatenated in the deconv4 layer. Then, the output of damage assessment is generated after a convolutional layer. The cross-entropy loss is also applied in Stage 2. Let the output of building segmentation be \( P_{b} \in \mathbb{R}^{2 \times H \times W} \), where \( H \) and \( W \) denote the height and width of the output, respectively. Then, the segmentation results can be expressed as

\[ P_{b} = \arg \max (P_{b}) \]  

(2)

where \( P_{b} \in \{0, 1\}^{1 \times H \times W} \) with 1 representing for buildings and 0 for background.

Let the output of damage assessment be \( P_{d} \in \mathbb{R}^{C \times H \times W} \), where \( C \) denotes the number of damage levels. To guide the building locations, the final output \( P \) is represented as

\[ P = \arg \max (P_{b} \cdot P_{d}) \]  

(3)

where \( \cdot \) denotes the elementwise multiplication.

B. MFF Module

As the scale of buildings varies in images captured from various disasters and regions, learning a robust representation of building objects under various scales is important. To this end, an MFF module is introduced in the encoder of the network. The input image is processed with three different scales: \( 1 \times \) input resolution (original size), \( 0.5 \times \) input resolution, and \( 0.25 \times \) input resolution. Features from different scales are extracted by a convolutional block and then embedded in the encoder of the network, as shown in Fig. 2(b). Specifically, the MFF contains three streams. The first stream uses images with the original resolution. In this stream, a convolutional layer with stride 2 is used to generate feature maps with half of the input size. This stream is the same as the first layer of ResNet, which is used as the backbone of the proposed framework. The \( 0.5 \times \) resolution images are used as input of the second stream. The size of feature maps of this stream keeps the same with \( 0.5 \times \) resolution images. Then, feature maps from \( 1 \times \) and \( 0.5 \times \) streams are concatenated for further processing. To keep a similar setting of the number of channels as the ResNet, a dimension reduction is conducted on the concatenated feature map. The second stream, the feature map of \( 0.25 \times \) resolution is concatenated with the second convolutional block of ResNet. In this way, features from three different image scales are embedded in the network, which enables the network to learn robust features.

C. CDA Module

To further explore the correlations between predisaster and postdisaster features, a CDA module is developed. Inspired by the squeeze and excitation (SE) block [56], the proposed CDA module focuses on recalibrating features from the channel and spatial dimensions. Moreover, the channel and spatial
information from predisaster and postdisaster features is aggregated together in a cross manner and then embedded in the network, respectively. The details are depicted in Fig. 3. Let \( U_{\text{pre}} \in \mathbb{R}^{E \times h \times w} \) and \( U_{\text{post}} \in \mathbb{R}^{E \times h \times w} \) be the feature maps obtained from the two branches of U-Net, respectively, where \( E \) denotes the number of channels of feature maps, and \( h \) and \( w \) are the height and width of feature maps, respectively. Then, the channel information can be extracted by

\[
I_{\text{cha}} = \sigma\left( P_g\left( [U_{\text{pre}}, U_{\text{post}}] \right) \right) \tag{4}
\]

where \([U_{\text{pre}}, U_{\text{post}}]\) denotes the concatenation of feature maps, \( P_g(\cdot) \) represents the global average pooling, \( \sigma(\cdot) \) is the sigmoid function, and \( I_{\text{cha}} \) is a feature vector of \( E \) channels after dimension reduction from \( 2E \). Then, the new features from two branches can be formulated as

\[
U_{\text{cha}}^{\text{pre}} = I_{\text{cha}} \ast U_{\text{post}} + U_{\text{pre}}
\]

\[
U_{\text{cha}}^{\text{post}} = I_{\text{cha}} \ast U_{\text{pre}} + U_{\text{post}} \tag{5}
\]

where \( \ast \) denotes the channelwise multiplication between the input feature maps and vector \( I_{\text{cha}} \). The output of channel feature fusion is then used in the spatial feature fusion. We concatenate \( U_{\text{cha}}^{\text{pre}} \) and \( U_{\text{cha}}^{\text{post}} \) and feed them into a 1 \( \times \) 1 convolution \( P_{\text{conv}} \) as

\[
I_{\text{spa}} = \sigma\left( P_{\text{conv}}\left( [U_{\text{cha}}^{\text{pre}}, U_{\text{cha}}^{\text{post}}] \right) \right) \tag{6}
\]

where \( P_{\text{conv}}\left([U_{\text{cha}}^{\text{pre}}, U_{\text{cha}}^{\text{post}}]\right) \in \mathbb{R}^{1 \times H \times W} \). Then, the output features are

\[
U_{\text{spa}}^{\text{pre}} = I_{\text{spa}} \cdot U_{\text{cha}}^{\text{post}} + U_{\text{pre}}
\]

\[
U_{\text{spa}}^{\text{post}} = I_{\text{spa}} \cdot U_{\text{cha}}^{\text{pre}} + U_{\text{post}} \tag{7}
\]

where \( I_{\text{spa}} \cdot U_{\text{cha}}^{\text{pre}} \) and \( I_{\text{spa}} \cdot U_{\text{cha}}^{\text{post}} \) denote the spatial elementwise multiplication. As a result, channel and spatial information from predisaster and postdisaster branches is effectively aggregated in the CDA module. It is worth noting that the proposed CDA module consists of only simple convolution and matrix operations, which is easy to implement and integrate with existing CNN architectures.

Therefore, the overall architecture of BDANet is established. The detailed parameters are listed in Table II.

### Table II: Parameters of the BDANet

<table>
<thead>
<tr>
<th>Layer</th>
<th>Feature Size</th>
<th>Kernel Size</th>
<th>Layer</th>
<th>Feature Size</th>
<th>Kernel Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input-pre</td>
<td>512 ( \times ) 512 ( \times ) 3</td>
<td>-</td>
<td>Input-pre</td>
<td>512 ( \times ) 512 ( \times ) 3</td>
<td>-</td>
</tr>
<tr>
<td>conv1-1</td>
<td>256 ( \times ) 256 ( \times ) 64</td>
<td>5 ( \times ) 5</td>
<td>conv2-1</td>
<td>256 ( \times ) 256 ( \times ) 64</td>
<td>-</td>
</tr>
<tr>
<td>conv1-2</td>
<td>128 ( \times ) 128 ( \times ) 256</td>
<td>3 ( \times ) 3</td>
<td>conv2-2</td>
<td>128 ( \times ) 128 ( \times ) 256</td>
<td>-</td>
</tr>
<tr>
<td>conv1-3</td>
<td>64 ( \times ) 64 ( \times ) 512</td>
<td>3 ( \times ) 3</td>
<td>conv2-3</td>
<td>64 ( \times ) 64 ( \times ) 512</td>
<td>-</td>
</tr>
<tr>
<td>conv1-4</td>
<td>32 ( \times ) 32 ( \times ) 1024</td>
<td>3 ( \times ) 3</td>
<td>conv2-4</td>
<td>32 ( \times ) 32 ( \times ) 1024</td>
<td>-</td>
</tr>
<tr>
<td>conv1-5</td>
<td>16 ( \times ) 16 ( \times ) 2048</td>
<td>3 ( \times ) 3</td>
<td>conv2-5</td>
<td>16 ( \times ) 16 ( \times ) 2048</td>
<td>-</td>
</tr>
<tr>
<td>dconv1-1</td>
<td>64 ( \times ) 64 ( \times ) 256</td>
<td>3 ( \times ) 3</td>
<td>dconv2-1</td>
<td>64 ( \times ) 64 ( \times ) 512</td>
<td>-</td>
</tr>
<tr>
<td>dconv1-2</td>
<td>128 ( \times ) 128 ( \times ) 96</td>
<td>3 ( \times ) 3</td>
<td>dconv2-2</td>
<td>128 ( \times ) 128 ( \times ) 192</td>
<td>-</td>
</tr>
<tr>
<td>dconv1-3</td>
<td>256 ( \times ) 256 ( \times ) 32</td>
<td>3 ( \times ) 3</td>
<td>dconv2-3</td>
<td>256 ( \times ) 256 ( \times ) 32</td>
<td>-</td>
</tr>
<tr>
<td>Output</td>
<td>512 ( \times ) 512 ( \times ) 1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Authorized licensed use limited to: Qian Du. Downloaded on January 06,2022 at 00:43:11 UTC from IEEE Xplore. Restrictions apply.
where $M \in \mathbb{R}^{1 \times W \times H}$ denotes a binary mask indicating where to cut out and fill in from two image samples, and $(\hat{X}_{\text{pre}}, \hat{X}_{\text{post}}, \hat{Y})$ represents the generated new sample. With the increased sample size using CutMix, the network is forced to pay more attention to these difficult classes and learn more robust representations.

IV. EXPERIMENTS

A. Dataset and Evaluation

To evaluate the effectiveness of our proposed method, the xBD dataset [18] is used in the experiments. The xBD dataset is the only publicly available large-scale dataset of satellite images for building segmentation and damage assessment at present. It was sourced from the DigitalGlobe Open Program, which was released for sudden major disaster events. Specifically, the Open Data Program was chosen for its availability of high-resolution imagery from many disparate regions of the world. In this dataset, 19 different disasters (such as hurricanes, floods, wildfire, and earthquakes) are selected at various locations with more than 800,000 building annotations. The dataset consists of image pairs (predisaster and postdisaster) of size $1024 \times 1024$ pixels. Each image contains three spectral bands (red, green, and blue) and has a resolution of 0.8 m/pixel. The damage assessment contains four levels, including no damage, minor damage, major damage, and destroyed. The training and testing sets are listed in Table III. It is worth mentioning that the data are imbalanced, and the damage level is highly skewed toward no damage. The number of each damage level’s polygons is reported in Table IV.

The experimental results are evaluated by using the F1 score ($F_{1_b}$) for building segmentation and the harmonic mean of classwise damage classification F1 ($F_{1_d}$) for building damage assessment, which are defined as

$$F_{1_b} = \frac{2TP}{2TP + FP + FN}$$

$$F_{1_d} = \frac{n}{\sum_{i=1}^{n} 1/F_{1_c}}$$

respectively, where TP, FP, and FN are the number of true-positive, false-positive, and false-negative pixels of segmentation results, respectively, $1/F_{1_c}$ denotes the F1 score of each damage level for damage assessment, and $C_i$ denotes the damage level. F1 score is a challenging evaluation index as it heavily penalizes the overfitting to over-represented classes. The overall score $F_1$ [18] comprehensively evaluates the performance of building segmentation and damage assessment, which is formulated by a weighted average of $F_{1_b}$ and $F_{1_d}$

$$F_1 = 0.3 \times F_{1_b} + 0.7 \times F_{1_d}. \quad (10)$$

B. Implementation Details

We implement the proposed framework using Pytorch. Experiments are conducted on a computer with an Intel i9-9920X CPU and two NVIDIA TITAN-V GPUs. Images are cropped to a size of $512 \times 512$ pixels for training. The prediction for the testing set is performed on the original size of images ($1024 \times 1024$). Apart from CutMix, basic data augmentation is conducted in the training phase, such as flip and rotation. Due to GPU memory limitation, we use ResNet-50 as the backbone for all compared networks with pretrained weights loaded from the PyTorch library. The number of channels in each convolutional block is [64, 256, 512, 1024, 2048]. In the decoder, the number of channels in each convolutional block is [512, 256, 96, 32]. The number of convolutional channels is the same as Stage 1 and Stage 2. The cross-entropy loss is used for both building segmentation and damage assessment. The optimization method is Adam. In the building segmentation stage, the learning rate is 0.00015, and the number of the epoch is 120. In the damage assessment stage, the learning rate is 0.0002, and the number of the epoch is 25.

C. Results Analysis

To validate the effectiveness of the proposed framework, we employ several state-of-the-art methods for comparison on xBD.

1) Two change detection-based networks, including WNet [53] and U-Net++ [57], are evaluated for building damage assessment.

2) The RescueNet [14] uses a dilated ResNet-50 as the backbone network. The difference of features from predisaster and postdisaster images is utilized for damage assessment.

3) In Weber’s method [12], after feature extraction with ResNet-50, features from predisaster and postdisaster images are concatenated for feature fusion. Then, the segmentation is performed for damage assessment.
4) Several state-of-the-art semantic segmentation networks, including FCN [58], SegNet [59], and DeepLabv3 [60], are also employed for comparison. We set the two-stage framework unchanged and only replace the U-Net-based architecture in Stage 2 with these mentioned networks. According to Table V, both methods provided low scores compared with other networks. It is reasonable.

Apart from our proposed two-stage framework, a network without the MFF module, the CDA module, and CutMix is used as a clean baseline for comparison, which is denoted as the vanilla network. The damage assessment results of different approaches are reported in Table V. The proposed framework using the vanilla network alone outperforms the existing methods, e.g., [14] and [12], in terms of all three metrics ($F_1s$, $F_1b$, and $F_1d$). Unlike the RescueNet and Weber’s method, the decoder and skip connections are used in our framework. Therefore, our method provides a better performance in the building segmentation stage (i.e., an improvement of over 2% in $F_1b$). Moreover, our method achieves a significant improvement (about 12%, $0.493 \rightarrow 0.616$) in minor damage level compared with RescueNet [14].

Different from our proposed framework, the change detection-based networks (WNet [53] and U-Net++ [57]) perform building segmentation and damage assessment in one network. According to Table V, low scores of $F_1$ are provided by both methods. It is reasonable because it is difficult for networks to locate buildings when postdamage images are also fed into the networks. From Table V, WNet provided the lowest score on $F_1b$ and the no damage level. As FCN uses simple upsampling to predict the damage levels and hardly considers features from shadow layers, the score in $F_1s$ is not satisfactory compared with other methods under the same framework. In contrast, SegNet and DeepLabv3 have achieved promising results using our proposed framework. Compared with the vanilla network, our overall approach incorporating the proposed components further improves the accuracy for damage assessment (0.782 versus 0.757 in $F_1d$), and the improvement is consistent for all the damage levels. By considering both the building segmentation and damage assessment performance (i.e., the overall score $F_1s$), our proposed two-stage framework outperforms the existing methods [12], [14] by a considerable margin. The classification confusion matrix of proposed framework is reported in Table VI. Compared with classification accuracy in Table I, much improvements are achieved in the minor damage and major damage level, indicating the effectiveness of our proposed method.

Some visual results of building segmentation are presented in Fig. 5. By decomposing the framework into two stages, the network can focus on building segmentation in Stage 1, achieving good segmentation results compared with the ground-truth segmentation maps. Therefore, the results can be used to effectively guide the building locations for damage assessment in Stage 2.

Fig. 6 shows a visual comparison of damage assessment results of different networks. For U-Net++, it is obvious that this method cannot provide precise outlines of building objects, and thus, many segmentation mistakes can be found. For FCN, more misclassification can be found from its
classification map compared with other methods. We can also observe that much misclassification comes from the levels of minor and major damages because both two levels are difficult classes. Compared with FCN and SegNet, DeepLabv3 provides a better visual result. Overall, our proposed BDANet achieves the best performance with fewer damage assessment errors.

D. Ablation Study

To investigate the contribution of the MFF module, the CDA module, and CutMix data augmentation, we carry out ablation studies on these components. Table VII reports the results of the vanilla network with these individual components. It can be seen that CDA and MFF bring about 0.6% and 0.5% improvements in $F_1$, respectively. Moreover, the CutMix strategy on difficult classes boosts the overall performance ($F_1$) by 1.3%, which is a considerable gain. In particular, the accuracy of minor damage class is improved by 2.7% (0.578 versus 0.605) with CutMix.

1) Analysis of MFF: As illustrated in Fig. 2, three feature streams are used to extract the features from different image resolutions. Therefore, this ablation study investigates the effectiveness of the three streams. Note that, when
only the stream with the original image resolution (1×) is used, the framework is consistent with the vanilla network. Table VIII reports the results of different streams applied in the network, where the 0.5× and 0.25× scales yield about 0.4% and 0.2% performance gains in \( F_{1s} \), respectively. When three streams are used together in the network, an improvement of 0.5% is obtained, indicating that learning features from different image scales is useful for improving the overall performance of damage assessment.

The feature maps from different streams are also depicted in Fig. 7. Specifically, we visualize several feature maps extracted from the first convolutional layer of the three streams. We can observe that local and detailed information can be extracted from the 1× stream, while more global information is captured from streams with lower resolution (0.5× and 0.25×).

2) **Analysis of CDA:** To explore the correlations between predisaster and postdisaster features, three CDA modules are used in the decoder of the network, which is represented as dconv1, dconv2, and dconv3 in Fig. 2. This experiment investigates the influence of the number of CDA modules and the embedding position of CDA in the network. Table IX shows that only slight performance gain is achieved when CDA is added in the convolutional layer of a small scale (dconv1). In contrast, a better improvement is obtained when CDA is added in the layer with a larger scale (dconv3). This is reasonable because more detailed information can be extracted when the feature maps are on a larger scale. When all three CDA modules are used in the network, around 0.6% improvement is achieved compared with the vanilla network.

As discussed in Section III-C, the CDA module is developed from the SE module [56]. Therefore, in this experiment, we further evaluate the effectiveness of CDA compared with the SE module. In the experiment, CDA is replaced with the channel (cha) and spatial (spa) attention modules of SE. As reported in Table X, only 0.1% improvement is achieved when the channel attention module is applied in the
3) Analysis of CutMix: For the xBD dataset, minor damage and major damage are the most difficult classes based on the results in Table I. Therefore, in our proposed method, CutMix data augmentation is mainly focused on these two damage levels to generate more training samples for them. We further conduct experiments to investigate the influence of CutMix on different damage levels. Table XI presents the overall scores when CutMix is applied on different damage levels. It is obvious that there is barely any improvement when CutMix is used on the entire dataset (i.e., considering all classes). In contrast, a significant improvement is achieved when only minor and major damage levels are considered in the data augmentation. This class-specific data augmentation strategy facilitates the network to learn better representations for those classes.

E. Computational Cost and Training Efficiency

We also compare the parameter size, computational cost (FLOP) and inference speed of the proposed overall framework, and the vanilla network with MFF and CDA modules. The CutMix strategy is used in the data preprocessing stage, and thus, it does not consume additional computational resources in the network. As shown in Table XII, the MFF module adds a little more computational cost since more convolutional operations are applied. Although the CDA module has a complicated architecture, it uses few convolutional operations. Therefore, it includes less than 1M additional parameters. This results in only a slight increase in computational cost (107.6 versus 92.9 GFLOPs). Moreover, our proposed framework only increases less than 2M parameters and about 60G FLOPs compared with the vanilla network. We also calculate the average inference time (s) for each image.

---

**Table X**

<table>
<thead>
<tr>
<th>Module</th>
<th>No Damage</th>
<th>Minor</th>
<th>Major</th>
<th>Destroyed</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE cha [56]</td>
<td>0.923</td>
<td>0.579</td>
<td>0.761</td>
<td>0.871</td>
<td>0.790</td>
</tr>
<tr>
<td>SE spa [56]</td>
<td>0.923</td>
<td>0.580</td>
<td>0.762</td>
<td>0.871</td>
<td>0.792</td>
</tr>
<tr>
<td>SE cha + spa [56]</td>
<td>0.923</td>
<td>0.586</td>
<td>0.763</td>
<td>0.871</td>
<td>0.793</td>
</tr>
<tr>
<td>CDA</td>
<td>0.923</td>
<td>0.592</td>
<td>0.769</td>
<td>0.871</td>
<td>0.795</td>
</tr>
</tbody>
</table>

**Table XI**

<table>
<thead>
<tr>
<th>Method</th>
<th>No damage</th>
<th>Minor</th>
<th>Major</th>
<th>Destroyed</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (vanilla)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.789</td>
</tr>
<tr>
<td>vanilla + CutMix</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.790</td>
</tr>
<tr>
<td>vanilla + CutMix</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.797</td>
</tr>
<tr>
<td>vanilla + CutMix</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>0.802</strong></td>
</tr>
<tr>
<td>vanilla + CutMix</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.795</td>
</tr>
</tbody>
</table>

**Table XII**

<table>
<thead>
<tr>
<th>Method</th>
<th>Params (M)</th>
<th>FLOPs (G)</th>
<th>Inference time (s/rgb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (vanilla)</td>
<td>32.5</td>
<td>92.9</td>
<td>0.174</td>
</tr>
<tr>
<td>vanilla + MFF</td>
<td>33.8</td>
<td>141.7</td>
<td>0.186</td>
</tr>
<tr>
<td>vanilla + CDA</td>
<td>33.1</td>
<td>107.6</td>
<td>0.181</td>
</tr>
<tr>
<td>Ours (BDANet)</td>
<td>34.4</td>
<td>155.4</td>
<td>0.190</td>
</tr>
</tbody>
</table>
method also yields state-of-the-art building damage assessment performance on the xBD dataset.

TABLE XIII
COMPARISON OF DIFFERENT METHODS ON THE LEVIR-CD DATASET

<table>
<thead>
<tr>
<th>Method</th>
<th>precision</th>
<th>recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNet</td>
<td>0.879</td>
<td>0.905</td>
<td>0.934</td>
</tr>
<tr>
<td>U-Net++ [57]</td>
<td>0.890</td>
<td>0.996</td>
<td>0.940</td>
</tr>
<tr>
<td>BDANet (Stage 2)</td>
<td>0.895</td>
<td>0.995</td>
<td>0.942</td>
</tr>
</tbody>
</table>

(1024 × 1024 pixels) in the testing phase. The additional runtime (0.016 s) is negligible.

In addition, we evaluate the training efficiency of using weights from Stage 1. As discussed in Section III, the network weights from Stage 1 are used as initial weights in the network of Stage 2 for fine-tuning. Fig. 8 plots the validation accuracy and loss versus the number of epochs. It is obvious that a higher validation accuracy and a smaller loss value are achieved at the first epoch when adopting the weights from Stage 1 as initialization. With the number of epochs increasing, the curve of using weights from Stage 1 converges faster. Fig. 8 shows that at least 5 epochs are saved when using the weights from Stage 1.

F. Evaluation on Building Change Detection Dataset

As more datasets on building damage assessment are difficult to acquire, a building change detection dataset LEVIR-CD [61] is used for evaluation of our proposed framework. These images are captured in several cities of TX, USA, from 2002 to 2018. The dataset collected 637 high-resolution images (0.5 m/pixel) with a size of 1024 × 1024 via Google Earth. Among these images, 509 image pairs are used for training, and 128 image pairs are used for testing.

Some change detection-based methods, including WNet [53] and U-Net++ [57], are used in comparison with BDANet. For the change detection task, only the network in Stage 2 is used for training. The results of different methods on LEVIR-CD are reported in Table XIII. From the table, we can observe that the proposed method achieved the best performance on the xBD dataset.

Some change detection-based methods, including WNet [53] and U-Net++ [57], are used in comparison with BDANet. For the change detection task, only the network in Stage 2 is used for training. The results of different methods on LEVIR-CD are reported in Table XIII. From the table, we can observe that the proposed method achieved the best performance on the xBD dataset.

V. CONCLUSION

In this article, a two-stage CNN-based learning framework, named BDANet, is proposed for building damage assessment from satellite images. In Stage 1, a U-Net architecture is used for building segmentation, and the output is used for guiding the building locations in damage assessment. In Stage 2, the MFF and CDA modules are used for further improving the assessment performance. MFF is used to learn features from various scales of images, which enables the network to build a robust feature representation. The CDA module is used to explore the correlations between predisaster and postdisaster features and enhance the feature representation capability. Moreover, a data augmentation strategy CutMix is employed to mitigate the challenge of difficult classes. Experimental results show that significant improvements can be achieved when CutMix is applied on difficult damage levels.

REFERENCES

J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, “Multi-scale spatial and


Yu Shen received the B.E. degree in computer science and technology from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 2015, where he is pursuing the Ph.D. degree in computer science with the School of Computer Science and Engineering. His research interests include the areas of remote sensing image processing, computer vision, and machine learning.
Sijie Zhu received the M.S. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2018 and the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2015. He is pursuing the Ph.D. degree with the University of North Carolina at Charlotte, Charlotte, NC, USA.

His research interests lie in visual representation learning and computer vision applications.

Taojiaman Yang (Graduate Student Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2017. He is pursuing the Ph.D. degree with the University of North Carolina at Charlotte, Charlotte, NC, USA.

His research interests lie in visual representation learning and computer vision applications.

Chen Chen (Member, IEEE) received the B.E. degree in automation from Beijing Forestry University, Beijing, China, in 2009, the M.S. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2012, and the Ph.D. degree from the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA, in 2016.

He held a Post-Doctoral Research Associate position at the Center for Research in Computer Vision (CRCV), University of Central Florida, Orlando, FL, USA, from July 2016 to June 2018. He is an Assistant Professor with the Department of Electrical and Computer Engineering at the University of North Carolina at Charlotte, Charlotte, NC, USA. His research interests include signal and image processing, computer vision, and deep learning.

Dr. Chen is an Associate Editor of Journal of Real-Time Image Processing and IEEE Journal on Miniaturization for Air and Space Systems.

Delu Pan received the B.S. degree in physics from the Nanjing University of Science and Technology, Nanjing, China, in 1968. He is a Senior Scientist (expert of marine remote sensing) with the Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, China. He is also an Academician of the Chinese Academy of Engineering, Beijing, China, and a Professor with Zhejiang University, Hangzhou.

Mr. Pan is the Vice President of the Chinese Society of Oceanography and has been a member of the International Ocean Color Coordinating Group since 2000.

Jianyu Chen (Member, IEEE) received the M.S. degree in environmental science and the Ph.D. degree in earth science from Zhejiang University, Hangzhou, China, in 1998 and 2004, respectively.

He is a Professor of remote sensing with the State Key Laboratory of Satellite Ocean Environment Dynamics, Second Institute of Oceanography, Ministry of Natural Resources, Hangzhou, China. He has been a Post-Doctoral Researcher with the Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, China. He has also been a Visiting Professor with the Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, Canada. His research interests include object-based image analysis, remote sensing of the coastal environment, and geographic information system.

Liang Xiao (Member, IEEE) received the B.S. degree in applied mathematics and the Ph.D. degree in computer science from the Nanjing University of Science and Technology (NJUST), Nanjing, China, in 1999 and 2004, respectively.

Since 2013, he has been the Deputy Director of the Jiangsu Key Laboratory of Spectral Imaging Intelligent Perception, China. Since 2014, he has been the second Director of the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, NJUST, where he is a Professor with the School of Computer Science and Engineering. His fields of interests include remote sensing image processing, image modeling, computer vision, machine learning, and pattern recognition.

Qian Du (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Maryland Baltimore County, Baltimore, MD, USA, in 2000.

She is the Bobby Shackouls Professor with the Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS, USA. Her research interests include hyperspectral remote sensing image analysis, pattern recognition, and machine learning.

Dr. Du served as a Co-Chair for the Data Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) from 2009 to 2013 and the Chair for the Remote Sensing and Mapping Technical Committee of the International Association for Pattern Recognition (IAPR) from 2010 to 2014. She served as the Editor-in-Chief of the IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing from 2016 to 2020. She is a Fellow of SPIE.