





BDD-Net+: A Building Damage Detection Framework Based on Modified Coat-Net

Seyd Teymoor Seydi , *Member, IEEE*, Mahdi Hasanlou , *Member, IEEE*, Jocelyn Chanussot , *Fellow, IEEE*, and Pedram Ghamisi , *Senior Member, IEEE*

Abstract—The accurate and fast assessment of damaged buildings following a disaster is critical for planning rescue and reconstruction efforts. The damage assessment by the traditional methods is time-consuming and with limited performance. In this article, we propose an end-to-end deep-learning network named building damage detection network-plus (BDD-Net+). The BDD-Net+ is based on a combination of convolution layers and transformer blocks. The proposed framework takes the advantage of the multiscale residual convolution blocks and self-attention layers. The proposed framework consists of four main steps: data preparation, model training, damage map generation and evaluation, and the use of an explainable artificial intelligence (XAI) framework for understanding and interpretation of the operation model. The experimental results include two representative real-world benchmark datasets (i.e., the Haiti earthquake and the Bata explosion). The obtained results illustrate that BDD-Net+ achieves excellent efficacy in comparison with other state-of-the-art methods. Furthermore, the visualization of the results by XAI shows that BDD-Net+ provides more interpretable and explainable results for damage detection than the other studied methods.

Index Terms—Damage detection, deep learning, earthquake, explainable artificial intelligence (XAI), transformer.

I. INTRODUCTION

EARTHQUAKES are among the most important and destructive natural disasters. An earthquake-induced collapse of buildings causes many fatalities [1]. Therefore, rapid and accurate mapping of damaged buildings after an earthquake is crucial for rescue operations, as it improves the response times of emergency response missions [2]. Satellite-based remote sensing is the most important data source due to its characteristics, such as spatial coverage and spatial resolution, low cost, and availability (both for postevent acquisitions and pre-event archives for change detection) [3]. Remote sensing technologies

are, hence, the primary tool in a wide range of applications in the monitoring of natural hazards, such as flood mapping [4], burned area detection [5], landslide mapping [6], and hurricane monitoring [7].

Building damage detection (BDD) is a vitally important topic of research because of its societal importance concerning potential casualties following a disaster. To this end, many BDD frameworks based on remote sensing images and unmanned aerial vehicle datasets have been proposed in the literature. These methods can be categorized into four main groups, depending on the input data source they utilize, including light detection and ranging (Lidar) data, synthetic aperture radar (SAR) images, 3) very high resolution (VHR) optical imagery, and fusion of multiple modalities.

Lidar data provide the height information of ground objects. Lidar data are not impacted by cloud coverage or illumination conditions. For example, Axel and van Aardt [8] proposed an unsupervised building damage assessment based on the segmentation of the point cloud Lidar dataset. The local surface features are first extracted, then the damaged building is detected based on the estimated rooftop inclination. Additionally, Aixia et al. [9] utilized surface normal algorithms to evaluate the level of building damage based on postearthquake Lidar data. The angle between the surface normal and zenith is utilized to map the damaged building parts. Furthermore, Janalipour and Mohammadzadeh [10] designed a building damage assessment framework based on the postearthquake Lidar dataset. They utilized three texture feature extraction manners, including Haralick's texture feature extraction, Gabor filter, and Laws' mask. Then, the building damage extents were generated using fuzzy inference systems based on the extracted textural features.

SAR images, with phase and amplitude information, may also be used for BDD. The damage mapping based on the SAR data can be applied in two main parts: First, damage detection based on amplitude or intensity. The main idea is that damaged buildings have irregular shapes, which can be captured using the intensity of the SAR imagery [11]. For instance, Chen et al. [12] developed a statistical texture feature G0-para to measure the homogeneity of buildings after a disaster. At first, the statistical texture features of G0-para were estimated by the G0 distribution of SAR images. Then, the ability of G0-para to distinguish between different polarization modes was compared based on the analysis by the receiver operating characteristic curve. Finally, the G0-para and the existing texture features were compared. Second, damage can also be mapped based on the

Manuscript received 7 February 2023; revised 20 March 2023; accepted 6 April 2023. Date of publication 17 April 2023; date of current version 10 May 2023. (Corresponding author: Mahdi Hasanlou.)

Seyd Teymoor Seydi and Mahdi Hasanlou are with the School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran 1439957131, Iran (e-mail: seydi.teymoor@ut.ac.ir; hasanlou@ut.ac.ir).

Jocelyn Chanussot is with the CNRS, Grenoble INP, GIPSA-Lab, University Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn.chanussot@grenoble-inp.fr).

Pedram Ghamisi is with the Helmholtz-Zentrum Dresden-Rossendorf, Machine Learning Group, Helmholtz Institute Freiberg for Resource Technology, D-09599 Freiberg, Germany, and also with the Institute of Advanced Research in Artificial Intelligence, 1030 Vienna, Austria (e-mail: pedram.ghamisi@iarai.ac.at).

Digital Object Identifier 10.1109/JSTARS.2023.3267847

phase information. To this end, the coherence map based on pre/postearthquake is generated and the normalized difference (ND) of the interferometric coherence map is created. Finally, the damage map is obtained based on the thresholding of the calculated ND map [13].

The BDD is most commonly conducted by using optical VHR remote sensing imagery. Unlike Lidar and SAR data, optical VHR datasets are simple to interpret and process. For instance, Wu et al. [14] proposed a damage mapping method based on the U-Net algorithm and attention mechanism. They utilized pre/postdisaster optical VHR remote sensing imagery for BDD. Abdi and Jabari [15] developed a fusion structure to map building damages based on combining off-nadir and orthophoto VHR datasets. The fusion structure of this framework is based on deep transfer learning. Qing et al. [16] designed a change detection-based BDD framework based on CNN and superpixel. This framework is applied in three steps:

- 1) building extraction based on extra feature enhancement bands;
- 2) damage detection based on change detection method with pre-event superpixel constraint strategy;
- 3) a quantitative assessment of the damage using a damage index.

Zheng et al. [17] designed an object-based semantic change detection framework for BDD. For end-to-end building damage assessment, the deep object localization network and deep damage classification network were also merged into one semantic change detection network. Ge et al. [18] presented an incremental learning framework for classifying collapsed buildings. For this purpose, they used end-to-end gradient boosting networks with an assemble-decision strategy as an incremental learning framework. Furthermore, using cycle-consistent generative adversarial networks, the pre-event disaster dataset is transformed into the same style as a postdisaster dataset. Chen et al. [19] developed a transformer-based framework for BDD using bitemporal datasets. The nonlocal deep features are extracted from bitemporal images by transformer encode, and then are integrated by fuse module. Finally, multilevel features are aggregated for final prediction by a lightweight dual-task decoder. Shen et al. [20] proposed a two-stage CNN for BDD. They utilized the U-Net model employed to extract buildings. Then, the weight of network is shared into next stage for BDD. In the second stage, a dual-branch multiscale U-Net is used as a backbone, feeding bitemporal datasets separately. In order to explore the correlations between bitemporal images, a cross-directional attention module was proposed.

Multiple modalities of acquisition may also be used jointly to improve the results of BDD. For example, Li et al. [21] investigated a damage detection framework based on the fusion of bitemporal Lidar and optical VHR datasets. This method is applied in two stages: First, three-dimensional (3-D) building model reconstruction based on the pre-event dataset, and second, estimation of the rooftop patch-oriented 3-D for determining potential damage. Adriano et al. [22] proposed a framework for rapid damage detection methods based on the fusion of multisource datasets. They used bitemporal Sentinel-1, Sentinel-2,

and ALOS-2 datasets for damage assessment. In addition, the open street map (OSM) layer was utilized to determine the built-up area. An ensemble classifier procedure is used to produce the building damage map.

The above-mentioned methods provide acceptable results for BDD but suffer from some limitations.

- 1) BDD methods focused on change detection using bitemporal datasets fail when predisaster data are not available. Also, the extracted changes may be originated from external factors (i.e., registration error, atmospheric conditions, and noise).
- 2) Designing a generic data fusion framework for multimodal approaches appears to be challenging.
- 3) Most methods are based on the conventional classification algorithms, while advance deep-learning methods can significantly improve BDD results.

Considering these challenges, we propose a new framework. To this end, we propose an efficient framework for BDD based on a single postevent remote sensing dataset that results in improved accuracy and reduced error rates. The main contributions of this article are as follows.

- 1) BDD is based on a modified Coat-Net algorithm for the first time.
- 2) Proposed framework uses only the postevent optical VHR dataset without any additional processing.
- 3) Combining the multiscale convolution layers and separable convolution layers with a transformer encoder.
- 4) Utilizing a gradient-weighted class activation mapping (Grad-CAM) algorithm for the understanding and interpretation of the operating model in the BDD [explainable artificial intelligence (XAI)].

II. CASE STUDY AND SATELLITE IMAGES

The performances of BDD-Net+ are evaluated with two real disaster datasets.

A. Haiti Earthquake

An earthquake with a magnitude of 7.0 hit the western part of Haiti, approximately 25 km south of Port-au-Prince, on January 12, 2010, at 4:53 P.M. Fig. 1 shows the study area for Haiti Earthquake and the ground truth for it.

In this study area, we separated the training and test areas (see Fig. 1). The training sample data contain 645 polygons with 302 and 343 polygons as nondamaged and damaged buildings, respectively. Moreover, the size of sample data in the test area is 1440 polygons with 943 polygons and nondamaged buildings, while the remaining 497 polygons are associated with the damaged class.

B. Bata Explosion

An explosion occurred at the Nkuantoma gendarmerie armory and military barracks in Bata, Equatorial Guinea's economic center, on the afternoon of March 7, 2021. Fig. 2 shows the

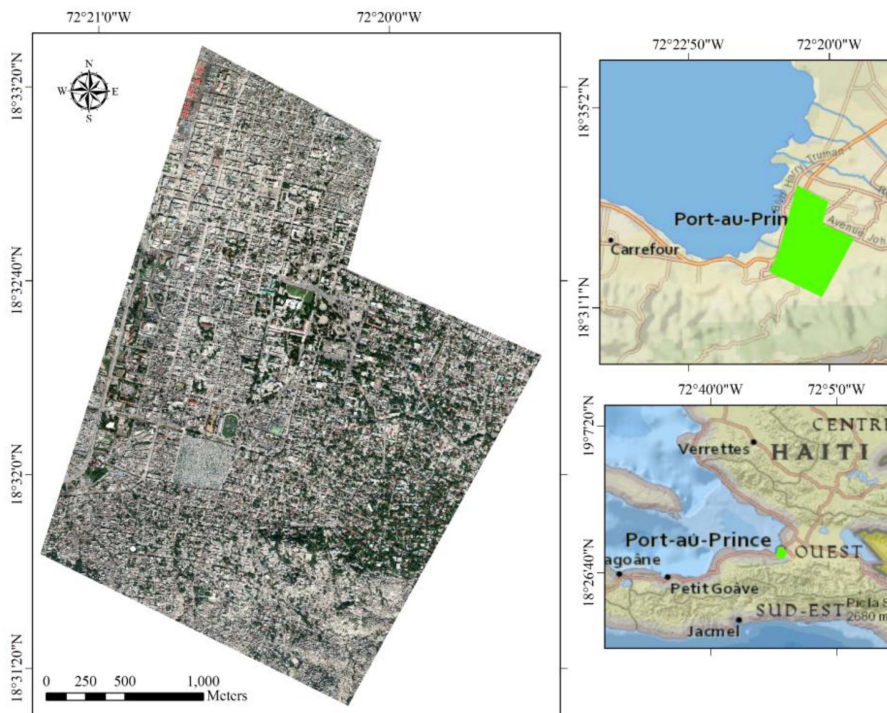


Fig. 1. Location of study for Haiti Earthquake.

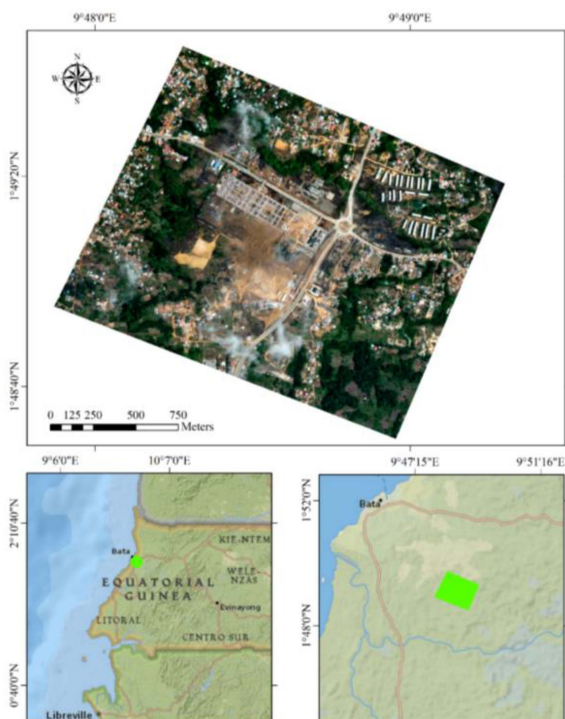


Fig. 2. Location of the second study area in Bata, Equatorial Guinea.

location of the study area for Bata Explosion. This dataset contains 706 building polygons among which 338 and 368 polygons belong to nondamaged and damaged, respectively.

TABLE I
DATASET DESCRIPTIONS FOR BOTH STUDY AREAS

Dataset	Haiti Earthquake	Bata-Explosion
Sensor Name	World-View-II	World-View-III
Spatial Resolution	0.5	0.5
Spectral Bands	3	4
Acquisition Date	January 16, 2011	March 9, 2011
Damaged Buildings	799	368
Non Damaged Buildings	1286	338

The bold values indicate the highest and best output and performance compared to the implemented algorithms.

Furthermore, the model is trained by 282 building polygons, while it is evaluated by 424 polygons.

C. Datasets

Both datasets were captured by optical VHR sensors (i.e., Worldview series sensors). The complete description of these datasets is presented in Table I. The red (R), green (G), and blue (B) spectral bands are used for the first dataset, while the near-infrared is also included in the second dataset.

D. Data Inventory

It is worth noting that the used datasets are benchmark datasets and have been employed in lots of research [15], [17], [23]. The ground truth of the Haiti-Earthquake dataset is available on the article presented in [18] and the website.¹ In addition, the ground

¹[Online]. Available: <https://dataverse.harvard.edu>

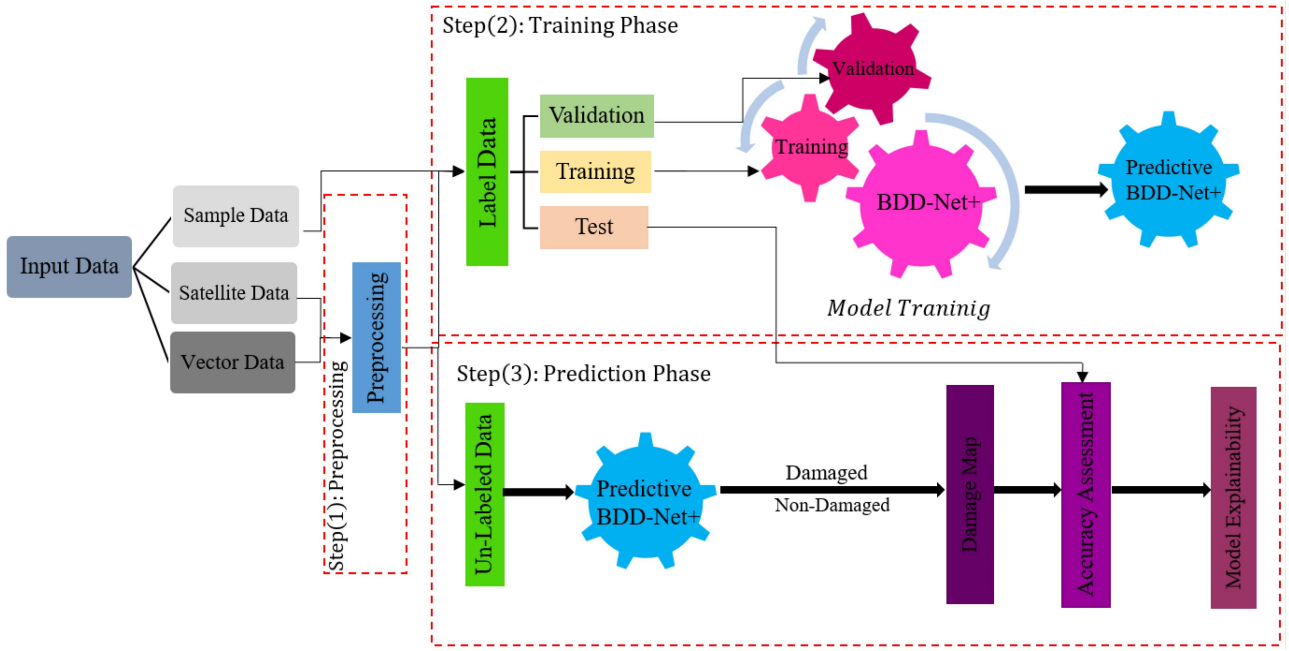


Fig. 3. General overview of the proposed BDD-Net+.

truth of the Bata-Explosion dataset is available on an open public website.²

E. Accuracy Assessment

To evaluate the performance of BDD-Net+, we employed the most popular quality assessment metrics, including the overall accuracy (OA), Kappa coefficient (KC), $F1$ -score, Recall, Precision, and Intersection Over Union (IOU). Furthermore, we analyzed the results of damage detection based on the visual inspection by comparing them with the ground truth map.

III. METHODOLOGY

A four-step framework is proposed for building damage mapping using a single postevent VHR dataset:

- 1) preprocessing;
- 2) training of BDD-Net+;
- 3) prediction by the predictive BDD-Net+ model;
- 4) model interpretation.

Fig. 3 presents the general overview of the proposed framework for BDD.

A. Preprocessing

The preprocessing step consists of extracting the footprint of candidates building from the optical VHR dataset and building the corresponding vector map. There are many ways for building footprint extraction (e.g., utilizing pretrained building segmentation models or an OSM). In this study, the footprint polygons of the buildings were manually delineated. To this end, the footprint polygons are overlaid on the raster VHR images and the buildings are extracted by masking the raster dataset.

B. BDD-Net+

Convolution layers have translation equivariance. This is an advantage that helps to have a strong inductive bias and critically improves the model generalization for unseen datasets when a limited dataset is available for training. A convolution step for input (x) at position (i) can be written as follows:

$$y_i = \sum_{j \in \mathcal{L}(i)} w \odot x_i \quad (1)$$

where y_i is the output of the convolution, and $\mathcal{L}(i)$ refers to the local receptive field.

The transformers were originally developed for a sequential dataset. It is proven that the transformer-based models [i.e., vision transformer (ViT)] have higher model capacity than CNN models [24]. The transformer-based deep-learning model employs a self-attention layer that has a global receptive field. One of the most important differences between convolution and self-attention layers is the size of the receptive field [25]. The self-attention layers have a global receptive field that provides more contextual information. Furthermore, the self-attention layers have an input-adaptive weighting mechanism. Thus, the transformer-based models have a high model capacity for large datasets. It is worth noting that there is a tradeoff between the size of the receptive field and the computational complexity. The self-attention mechanism is defined as follows [24]:

$$y_i = \sum_{j \in \mathcal{G}} \frac{e^{x_i^T x_j}}{\sum_{k \in \mathcal{G}} e^{x_i^T x_k}} x_j \quad (2)$$

where \mathcal{G} denotes the global spatial space, and $\sum_{k \in \mathcal{G}} e^{x_i^T x_k}$ refers to the dynamic attention weights.

The BDD-Net+ is inspired by the Coat-Net algorithm that combines BDD convolution and transformer layers. The main idea of

²[Online]. Available: <https://www.unitar.org>

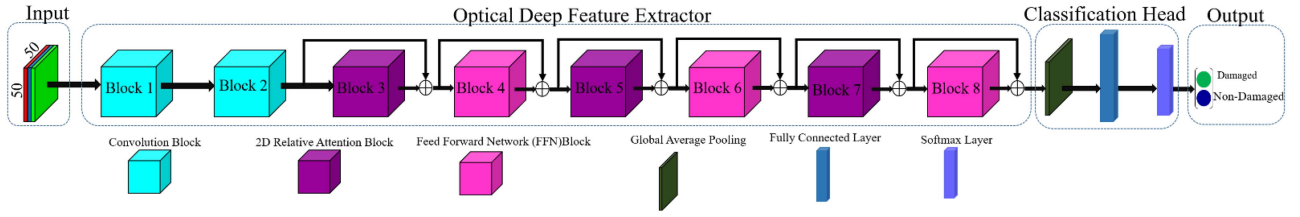


Fig. 4. Proposed BDD-Net+ framework.

BDD-Net+ is to improve the performance of BDD by taking the advantage of both convolution and self-attention layers. Thus, BDD-Net+ aims at taking the advantage of the convolution block for better generalization and the self-attention layer for increasing the model capacity [24]. Simple solutions to combine the self-attention and convolution layers include summing a global static convolution kernel with an adaptive attention matrix, which is shown as follows:

$$y_i = \sum_{j \in \mathcal{G}} \frac{e^{(x_i^T x_j + w_{i-j})}}{\sum_{k \in \mathcal{G}} e^{(x_i^T x_k + w_{i-k})}} x_j. \quad (3)$$

It should be noted that this form actually corresponds to a special case of a self-attention mechanism, called relative self-attention, which only focuses on relative position or distance.

The direct combination of attention and convolution layers increases the computational complexity significantly. It is, hence, proposed by Dai et al. [24] to reduce the spatial size of the feature map and to use global relative attention. The overview of the proposed BDD-Net+ is presented in Fig. 4. This architecture has been made up of two convolution blocks, three 2-D-relative attention blocks, and three feedforward-network (FFN) modules, with a global average pooling, a fully connected layer, and a *Softmax* layer in the classification head. The proposed framework has two main differences from the original Coat-Net that are included utilizing asymmetric convolutional structure with kernel sizes and utilizing depthwise separable convolution for increasing effeteness of network and reducing models' parameters.

1) *Convolution Block*: The convolution block is the first part of the proposed framework, aiming at deep feature extraction. The convolution layers extract meaningful high-level features from the input dataset [26]. Here, we use three types of convolution layers that are included: a standard convolution layer with kernel size (3×3) , an asymmetric convolutional structure with kernel sizes (3×1) and (1×3) , and a depthwise separable convolution. The depthwise separable convolution layer includes a depthwise convolution layer with a kernel size of 3×3 and pointwise kernel convolution (kernel size 1×1), respectively. Fig. 5 shows the corresponding structure.

2) *Residual Multihead Self-Attention Module*: The self-attention layer is a key component of the transformer that is widely used in many applications in signal, image, or language processing. It employs a self-attention module with 2-D relative position encoding [27]. The structure of the residual multihead

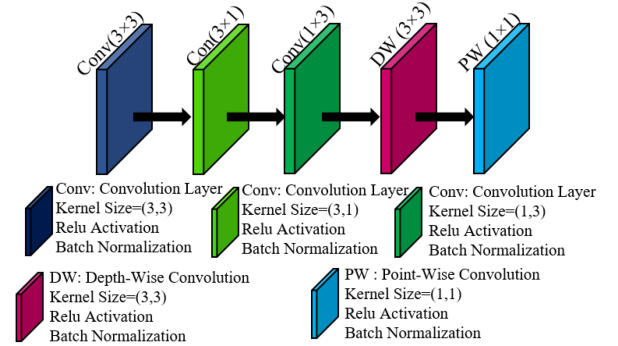


Fig. 5. Structure of the convolution block.

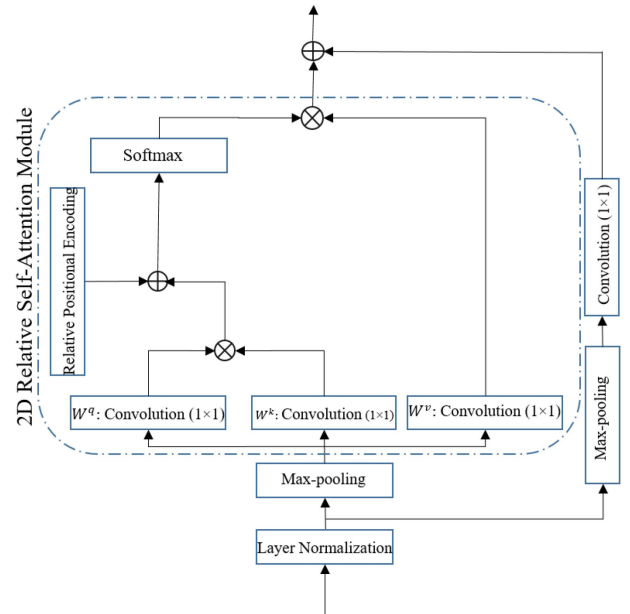


Fig. 6. Structure of the residual multihead self-attention module.

attention module with relative positional encoding is shown in Fig. 6. The layer normalization first processes the input feature map. Then, a pooling layer is used to reduce the spatial dimension of the feature map. Next, the feature map is fed to the relative self-attention module. Finally, a convolution layer is used for further exploration before adding the feature map to the output of the attention module. For the feature map with the

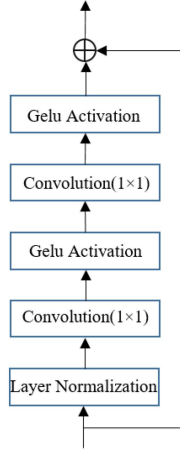


Fig. 7. Structure of the residual FFN block.

size of $(H \times W)$, the relative positional encoding has a learnable parameter (P) with the size of $(2H-1) \times (2W-1)$.

3) *Residual FFN Module*: The structure of the FFN module is similar to the multilayer perceptron head in the ViT algorithm. This block is composed of the following layers:

- 1) layer normalization;
- 2) convolution (1×1) ;
- 3) Gaussian error linear unit (GELU) activation function;
- 4) convolution (1×1) ;
- 5) GELU activation function;
- 6) sum with the input.

Fig. 7 features this residual FFN structure.

C. Model Training

The model parameters are obtained iteratively by employing an optimizer. Through backpropagation, the model parameters are tuned at every step to minimize errors when assessing the output of the model with the true value. To this end, the model is trained by training sample data, then evaluated by calculating the loss function on the validation dataset. Finally, the performance of the model is evaluated using the testing dataset. The performance of the model is compared with other state-of-the-art transformer-based models, including Swin stands for shifted window (Swin-Transformer) [28], pooling-based Vision Transformer (PiT) [29], compact convolutional transformer (CCT) [30], and the original Coat-Net [24]. The Swin Transformer creates hierarchical feature maps through the merging of image patches in deeper layers. The PiT model is a state-of-the-art transformer-based framework that considers the spatial dimension conversion on the transformer-based architecture. The CCT uses a convolutional tokenizer to produce richer tokens and preserve local information [30].

D. BDD-Net+ Model Explainability

Deep-learning-based models usually provide highly promising results. However, since the internal functioning of these models is unclear, these models have often been viewed as “black box” methods [31]. Grad-CAM is one of the most common

visual explanation methods for deep-learning models. It is used to show the predictions of the models more visually. To begin with, the gradient of the score y^c (before softmax) for each class is calculated concerning the feature maps (f^k) of a particular layer as follows [31]:

$$g_c(f^k) = \frac{\partial y^c}{\partial f^k} \quad (4)$$

where k is the channel index. Next, these gradients are global average pooled to estimate the importance of weight (a_k) f^k for the class c in each channel [31]

$$a_k^c = \frac{1}{w_f \times h_f} \sum_{i=1}^{w_f} \sum_{j=1}^{h_f} \frac{\partial y^c}{\partial f_{i,j}^k} \quad (5)$$

where w_f and h_f are the width and height of feature maps, respectively. The final Grad-CAM heat map ($H_{\text{Grad-Cam}}^c$) is a weighted sum of the feature maps, followed by a rectified linear unit (ReLU) activation function

$$H_{\text{Grad-Cam}}^c = \text{ReLU} \left(\sum_k a_k^c \cdot f^k \right). \quad (6)$$

IV. EXPERIMENTAL RESULTS

The supervised deep-learning models investigated in this study have different hyperparameters that require to be set. A trial-and-error procedure is followed to set these hyperparameters. The values of hyperparameters are patch size: 50×50 , batch-size: 300, number of iterations: 1500, weight initializer: Glorot initialization [32], optimizer: Adam³ optimizer, loss function: Tversky loss function [33], and learning rate 10^{-3} .

A. Visual Analysis

The results of BDD for the Haiti Earthquake are shown in Fig. 8. Based on the BDD results, we observed perfect classifications for all three transformer-based deep-learning methods for the nondamaged class. Generally, all methods have provided promising results in the BDD while differing in more detail. In contrast to the other four frameworks, the BDD-Net+’s damage detection result is consistent and satisfying, as shown in Fig. 8(e).

Fig. 8(b) shows the result of the PiT algorithm, which has many damaged buildings classified wrongly into the nondamaged class. Furthermore, the original Coat-Net led to many miss-detected buildings [see Fig. 8(d)]. BDD-Net+ leads to the best performances in terms of both correct detection of damaged buildings and the false alarm rate [see Fig. 8(e)]. In addition, the zoom areas of the results of BDD are shown in Fig. 9. As seen, the proposed model has considerable consistency in the different regions compared with other models.

Fig. 10 illustrates the result of BDD for Bata Explosion. It can be seen that all models provided promising results in the BDD. However, the PiT framework provided lower false detection than the other four methods, but it has many miss detection pixels [see Fig. 10(b)]. Some buildings that are not detected by the

³Adaptive Moment Estimation (Adam)

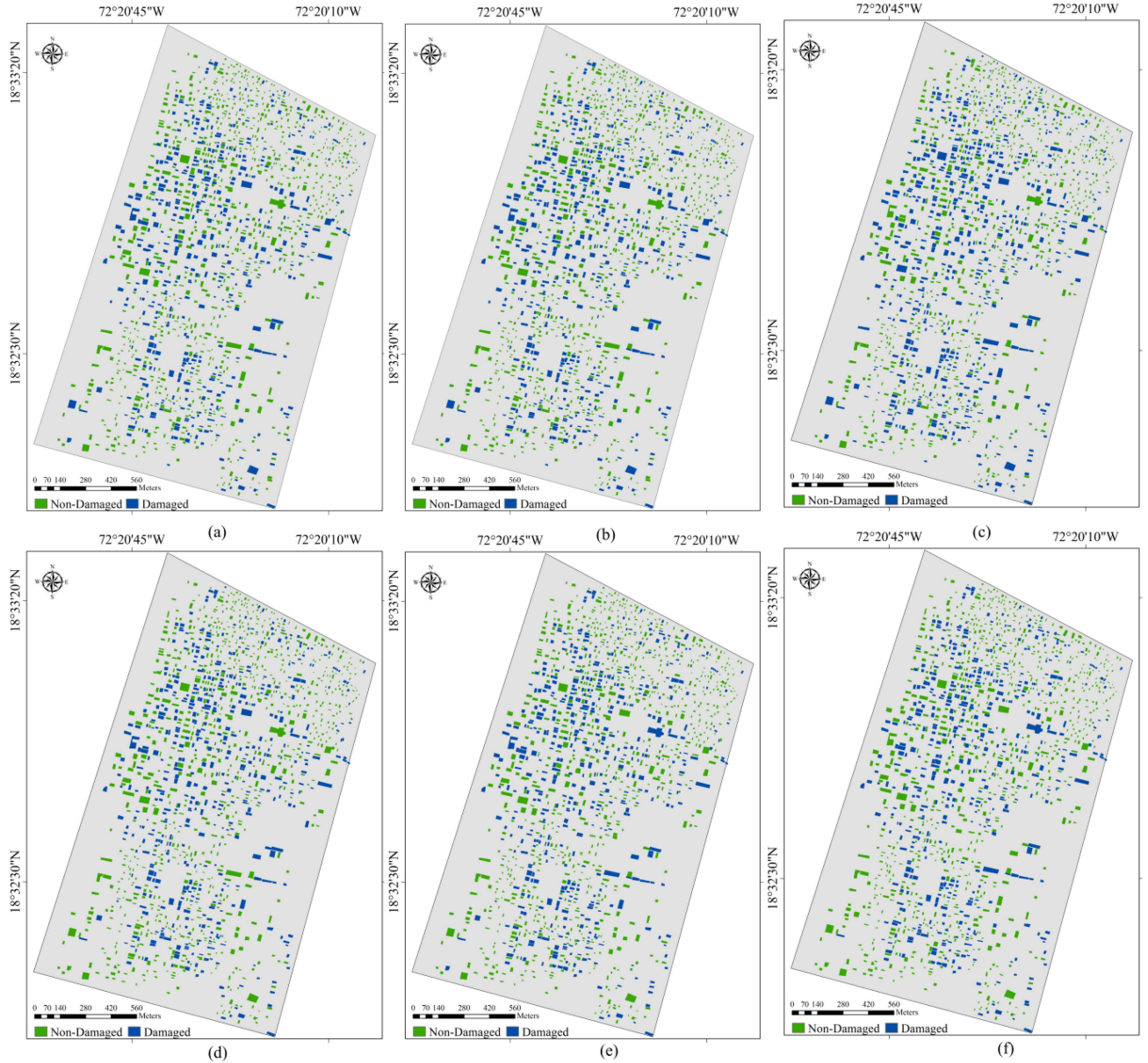


Fig. 8. Result of BDD for the Haiti-Earthquake dataset. (a) Swin-Transformer. (b) PiT. (c) CCT. (d) Coat-Net. (e) BDD-Net+. (f) Ground truth.

TABLE II
COMPARISON OF THE NUMERICAL RESULT OF BDD FOR THE HAITI EARTHQUAKE

Metric	OA (%)	Precision (%)	Recall (%)	F1-Score (%)	KC	IOU
Swin-Transformer	87.43	80.15	84.50	82.27	0.725	0.699
PiT	92.01	89.79	86.72	88.23	0.822	0.789
CCT	83.89	71.97	87.32	78.91	0.661	0.652
Coat-Net	92.36	94.08	83.10	88.24	0.826	0.789
BDD-Net+	94.02	91.68	90.94	91.13	0.868	0.840

The bold values indicate the highest and best output and performance compared to the implemented algorithms.

original Coat-Net model [see Fig. 10(d)] are accurately detected by BDD-Net+ [see Fig. 10(e)]. Furthermore, the zoom areas of the result of BDD by different models are shown in Fig. 11.

B. Numerical Analysis

The quantitative analysis of the Haiti Earthquake is reported in Table II. Again, BDD-Net+ outperforms the original Coat-Net.



Fig. 9. Comparison of the results of BDD algorithms or Haiti Earthquake.

Based on the numerical results, the BDD-Net+ considerably improved the result of BDD in most metrics. For instance, the OA was reported as 87%, 92%, 84%, 92%, and 94% when the Swin-Transformer, PiT, CCT, Coat-Net, and BDD-Net+ were applied, respectively. This shows that BDD-Net+ improved the result more, which is about 7% points more than Swin-Transformer, 2% points more than PiT, 10% points more than CCT, and 2% points more than the Coat-Net algorithm. Furthermore, the efficiency of the BDD-Net+ is more evident in other indices. It

is worth noting that Coat-Net has provided a performance better than BDD-Net+ while having missed the effectiveness by other metrics.

The quantitative analysis for Bata Explosion is shown in Table III. In accordance with these results, all five models achieved an accuracy of 80% or higher.

BDD-Net+ clearly outperforms Swin-Transformer, PiT, CCT, and the original Coat-Net with a 6%, 2%, 5%, and 4% points increase in OA and $F1$ -score, respectively. However, PiT

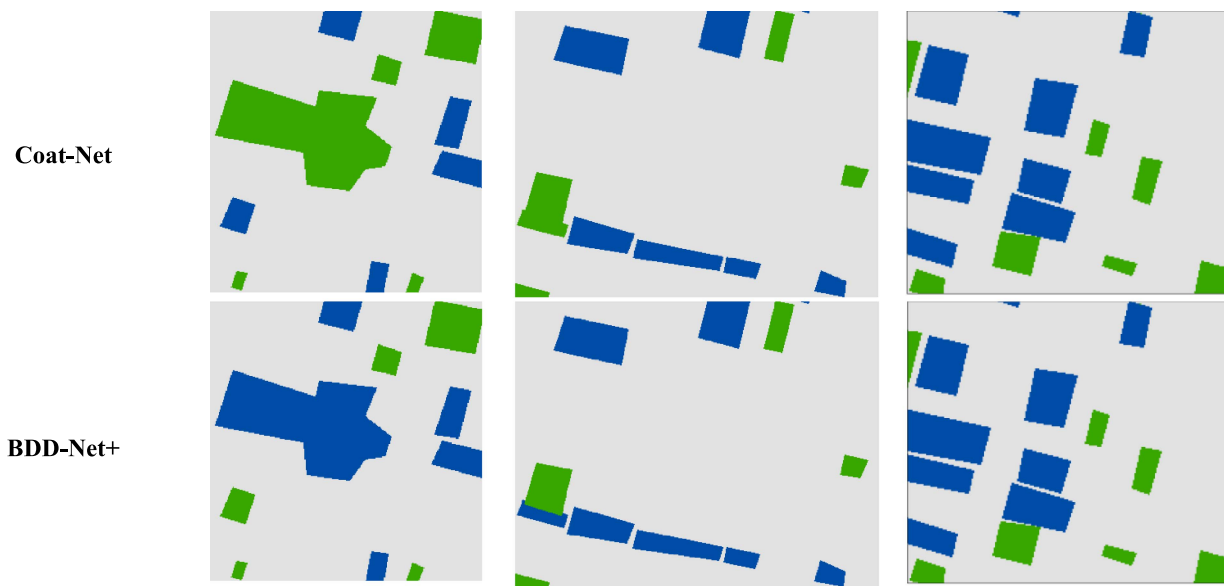


Fig. 9. (Continued.)

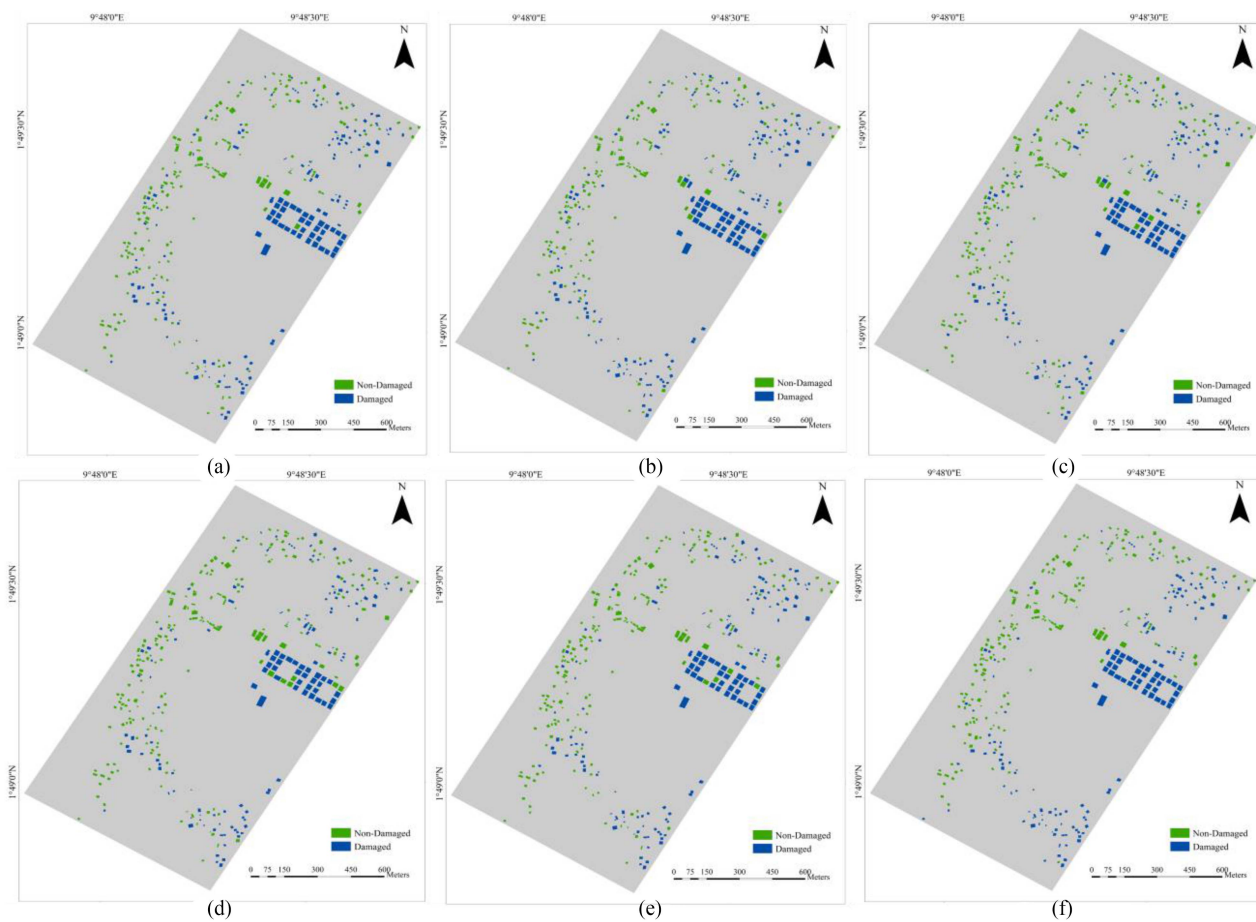


Fig. 10. Result of BDD for Bata Explosion. (a) Swin-Transformer. (b) PiT. (c) CCT. (d) Coat-Net. (e) BDD-Net+. (f) Ground truth.

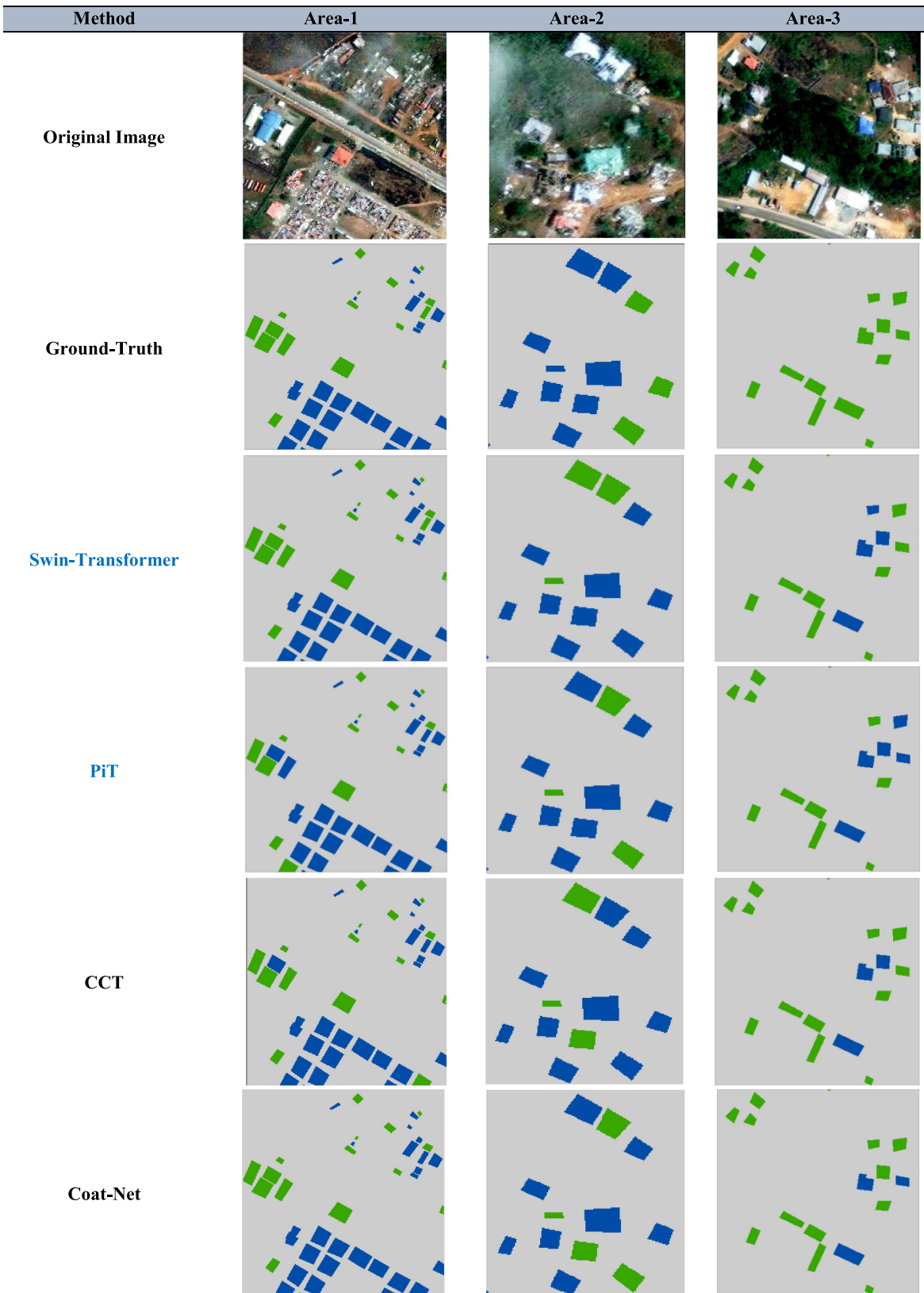


Fig. 11. Comparison of the results BDD by different models by Zoom regions in Bata Explosion.

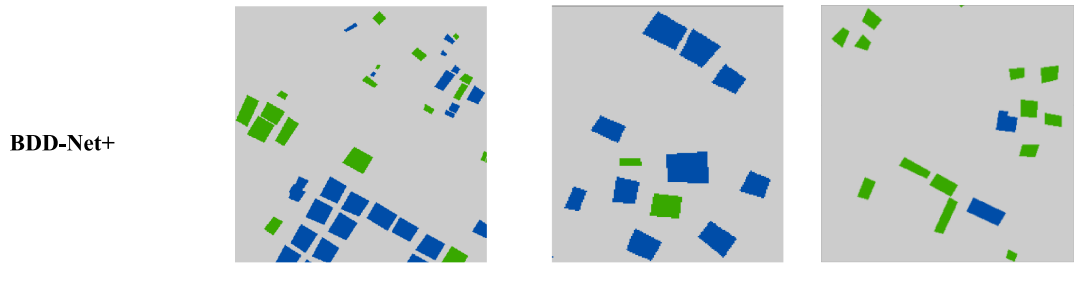


Fig. 11. (Continued.)

TABLE III
COMPARISON OF THE NUMERICAL RESULT OF BDD FOR BATA EXPLOSION

Metric	OA (%)	Precision (%)	Recall (%)	F1-Score (%)	KC	IOU
Swin-Transformer	83.10	82.25	81.29	81.76	0.660	0.691
PiT	85.55	80.72	90.64	85.39	0.712	0.745
CCT	81.73	81.32	78.94	80.11	0.632	0.668
Coat-Net	82.29	85.81	74.27	76.62	0.641	0.661
BDD-Net+	86.92	88.20	83.04	85.54	0.736	0.747

The bold values indicate the highest and best output and performance compared to the implemented algorithms.

has achieved a Recall a bit more than BDD-Net+ (about 7% points) but it has a weaker performance in terms of Precision (about 8% points) and KC. In general, BDD-Net+ obtained the best results for Bata explosion, which again were slightly higher than PiT but significantly better than the original Coat-Net and CCT.

C. Model Visualization

The model explainability is the latest step of the proposed method. To this end, we visualize the latest layer before the global average pooling layer by the Grad-CAM algorithm. We consider the performance of the model in two classes (nondamaged and damaged) in a Grad-CAM-based manner.

The results of model visualization for nondamaged buildings are shown in Fig. 12. The nondamaged regions have red color in the results of visualization. As can be seen in the visualization of the model, the red areas indicate the points where the BDD-Net+ has focused to predict the outcome. Since the nondamaged building appears with a smooth texture, the model has considered all building surfaces for classification.

Fig. 13 demonstrates the result of the Grad-CAM algorithm for damaged buildings. Based on the results, the rough surface roofs have high intensity (red color in Fig. 13, third column), which shows that the focus of BDD-Net+ is on rough surface areas. As a result, the damaged areas have highly rough surfaces and the model focuses mainly on the collapsed areas for identification of the damaged buildings.

D. Ablation Analysis

Ablation studies provide insight into the relative contribution of different architectural components to the performance of a deep-learning framework. To achieve this, the BDD-Net+ is analyzed using three scenarios:

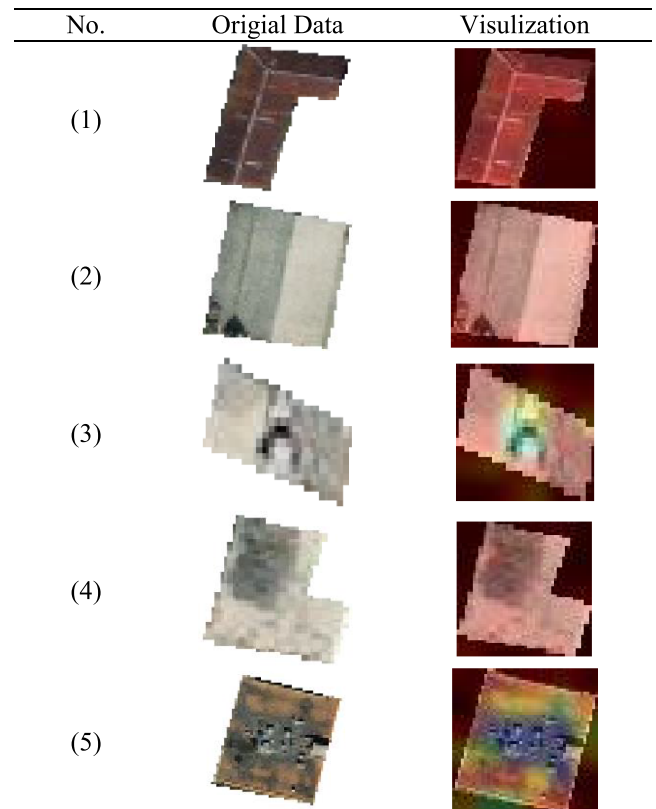


Fig. 12. Comparison of Grad-CAM results for the nondamaged class.

- 1) BDD-Net+ without asymmetric convolution structure and depth separable convolution (S#1);
- 2) BDD-Net+ without all convolution blocks (S#2);
- 3) BDD-Net+ without transformer layers (S#3).

TABLE IV
BDD-NET+ ABLATION ANALYSIS FOR BATA EXPLOSION

Scenario	OA (%)	Precision (%)	Recall (%)	F1-Score (%)	KC	IOU
S#1	82.83	65.40	82.40	81.70	0.654	0.691
S#2	82.29	85.81	74.27	76.62	0.641	0.661
S#3	85.01	79.29	91.81	85.09	0.701	0.740
BDD-Net+	86.92	88.20	83.04	85.54	0.736	0.747

The bold values indicate the highest and best output and performance compared to the implemented algorithms.

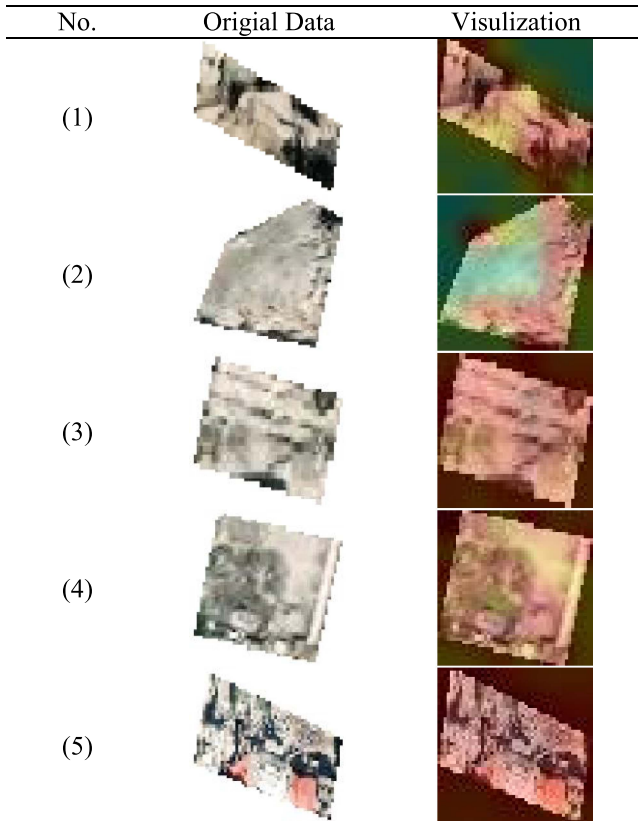


Fig. 13. Comparison of Grad-CAM results for the damaged class.

Table IV presents the ablation study in three scenarios. The numerical results show that the second scenario has the highest influence on the performance of the BDD-Net+. Furthermore, the transformer layers improved the accuracy of the model.

V. DISCUSSION

To effectively respond to earthquake emergencies, it is critical to accurately identify the areas affected by an earthquake after it has occurred. To this end, this research proposed a deep-learning framework by the combination of the multiscale residual convolution blocks and self-attention layers. The efficacy of the proposed framework is evaluated by two real-world VHR datasets with different hazards (i.e., earthquake and explosion). The results of BDD were compared with the other two state-of-the-art deep-learning-based methods. Concerning the BDD results, the three deep-learning-based methods have provided promising results in both datasets.

The normalized confusion matrix of the BDD for the Haiti-Earthquake dataset is shown in Fig. 14. As can be seen, the performance of all models in Haiti Earthquake is better than Bata Explosion (see Fig. 15). The BDD-Net+ has provided promising results in terms of both nondamaged and damaged classes. It is worth noting that there is a tradeoff between nondamaged or damaged classes. An ideal situation would be to be able to detect both nondamaged and damaged buildings in the most accurate manner possible. However, some methods only focused on one nondamaged or damaged class (i.e., Coat-Net and PiT) as they provided high accuracy on one class while noticeably missing the performance in another class. This issue can be seen in Fig. 14(b)–(d) for the Haiti-Earthquake dataset.

The normalized confusion matrix of the Bata-Explosion dataset is shown in Fig. 15. Based on these results, the Coat-Net and BDD-Net+ have similar results for the nondamaged class, while those are different for the case of damaged classes. A bit of improvement has been obtained with the PiT in nondamage detection compared with BDD-Net+, while significant deficiencies have been observed in the detection of damaged buildings. In general, both the PiT model and the BDD-Net+ model perform similarly when it comes to detecting the nondamage class. However, when it comes to detecting the damaged class, the BDD-Net+ model outperforms the PiT model.

It is worth noting that the number of sample datasets in the Haiti-Earthquake dataset is more than in the Bata-Explosion sample dataset. This issue may influence the BDD results and lead to greater results for Haiti Earthquake than the Bata-Explosion dataset.

The generalization ability is one of the most important deep-learning-based methods. In this regard, the Haiti-Earthquake dataset has a separate test area. We also evaluated the performance of the model in a different area. The presented result in Table II, Table III, Figs. 8, and 10 demonstrates that BDD-Net+ has a high generalization in comparison with other methods.

The sample data size is another criterion of supervised learning methods that BDD-Net+ has trained about 640 and 282 samples for Haiti-Earthquake and Bata-Explosion datasets. The result of BDD by BDD-Net+ based on this amount of sample dataset is valuable. The semantic segmentation model (i.e., U-Net models) requires a high amount of sample datasets. The collection of a high amount of sample datasets for such an application is challenging.

The XAI can explain to that model how to predict the result based on the input dataset. This advantage helps us find out behind false positive and false negative polygons what has happened. Fig. 16 demonstrates some false negative building

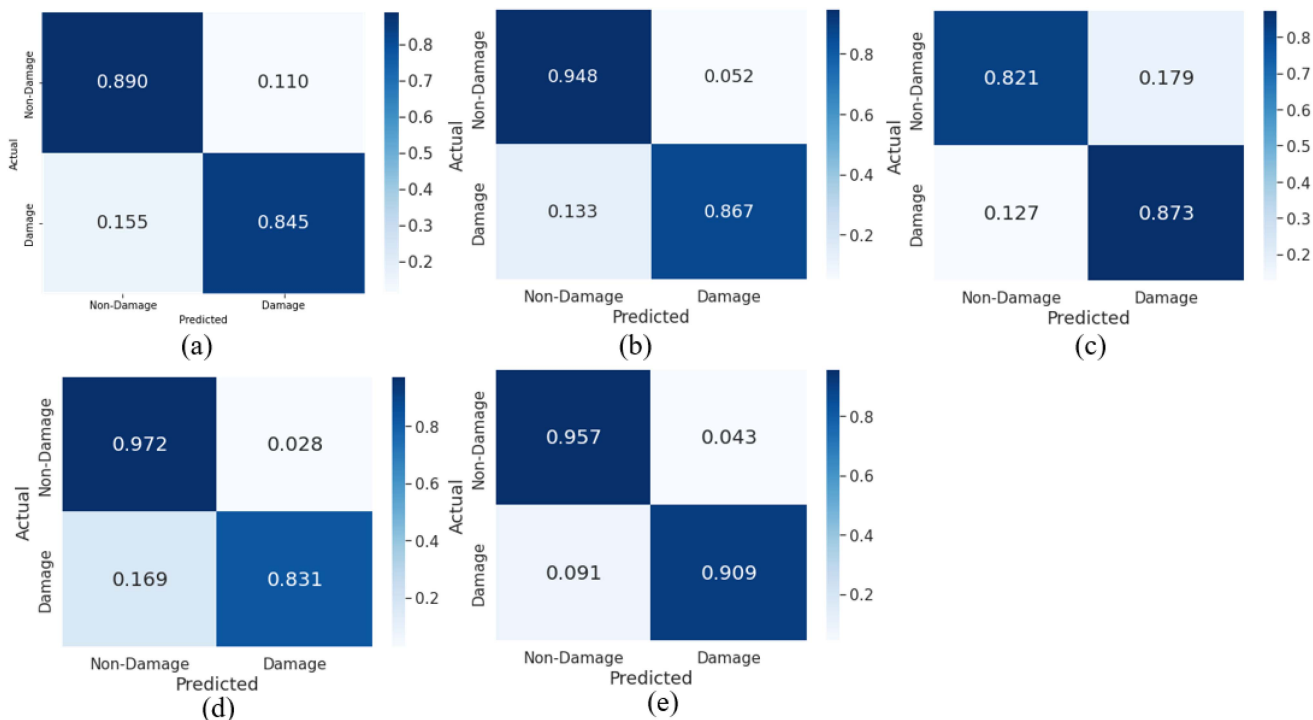


Fig. 14. Normalized confusion matrix comparison for Haiti Earthquake. (a) Swin-Transformer. (b) PiT. (c) CCT. (d) Coat-Net. (e) BDD-Net+.

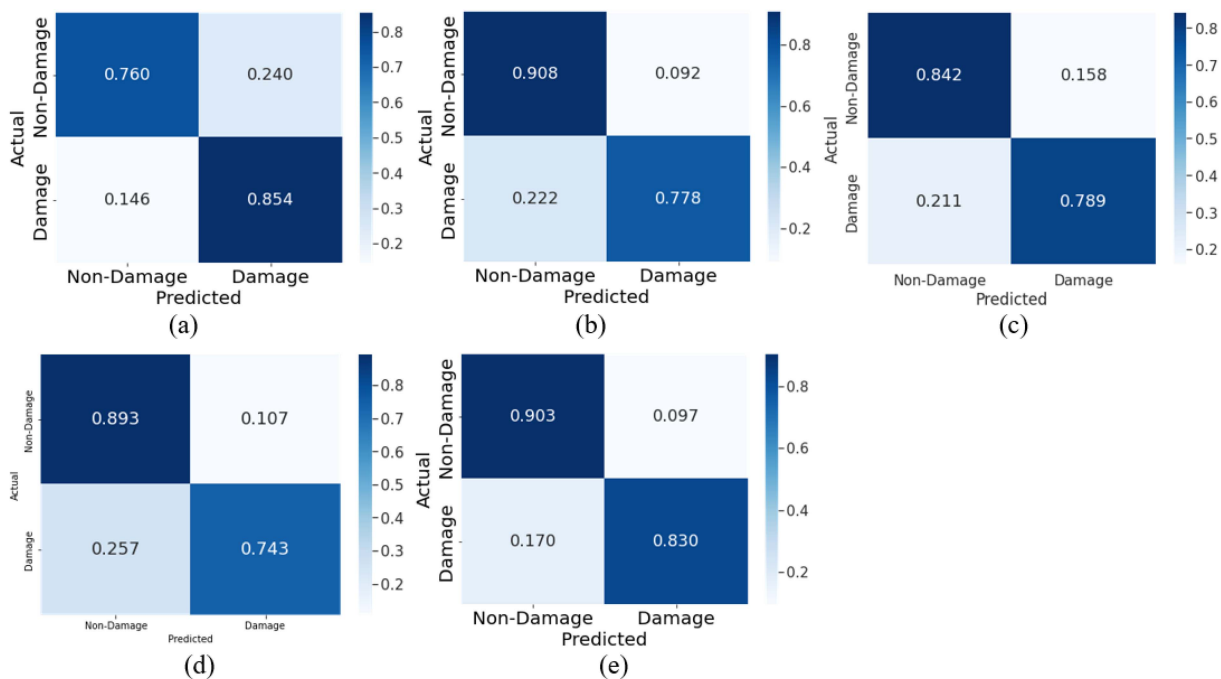


Fig. 15. Normalized confusion matrix comparison for Bata explosion. (a) Swin-Transformer. (b) PiT. (c) CCT. (d) Coat-Net. (e) BDD-Net+.

polygons. The model was predicted as a nondamaged class, while those being damaged buildings. As seen, the BDD-Net+ focused on smooth texture areas (red areas) because these buildings have highly smooth areas in comparison with the damaged

region. Concerning, the result of XAI, it is better to model trained with such samples.

Similarly, we explored the false positive samples to understand what the main reason for miss detection was. Fig. 17

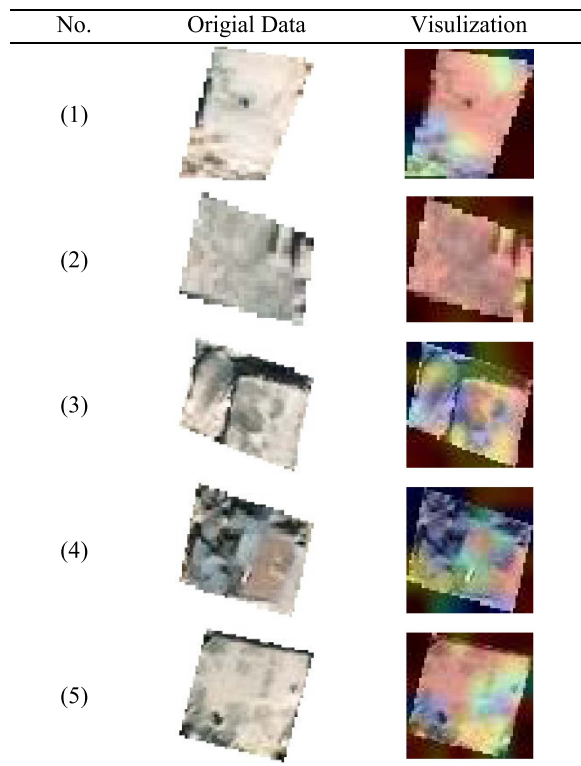


Fig. 16. Comparison of Grad-CAM results for false negative samples.

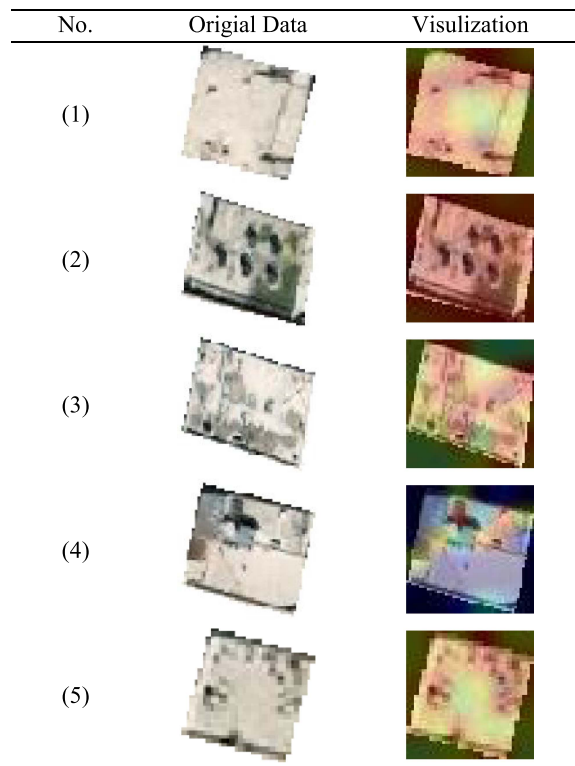


Fig. 17. Comparison of Grad-CAM results for false positive samples.

TABLE V
COMPARISON OF THE NUMBER OF PARAMETERS OF DEEP-LEARNING MODEL

Model	Number of Parameters
Swin-Transformer	2 185 666
PiT	645 250
CCT	773 955
Coat-Net	14 437 846
BDD-Net+	208 527

illustrates some false positive samples. Based on this, some roof structure has caused the BDD-Net+ to focus on these regions for prediction. Thus, the roof structure has a key role in the model prediction.

This study only utilized the postearthquake dataset for building damage assessment. However, their many methods focused on bitemporal pre/postevent datasets that suffer high computational costs (due to processing bitemporal images). The computational cost of the proposed framework is lower than other similar methods for BDD because of using only postevent datasets. In addition, the building vector map is available for the whole world. Thus, we used the building vector map to prevent processing nonbuilding pixels that reduce the mapping processing times significantly.

The commotional of the cost of deep-learning methods is a more important factor. To this end, we evaluated the number of parameters of all methods. Table V presents that the number of parameters of deep-learning models that BDD-Net+ has lower than parameters comparison with other models. This subject causes the proposed framework to be quickly trained.

VI. CONCLUSION

We proposed a novel framework for BDD using a single postevent VHR dataset. To this end, a novel effective BDD has been proposed, which combines the transformer and convolution layers. The performances of BDD are positively evaluated on two real-world datasets. The results of BDD show that the proposed BDD-Net+ has very high performance in the detection of damaged buildings. Furthermore, BDD-Net+ provides robust results even with an unbalanced dataset. Moreover, we used the Grad-CAM algorithm for model explainability. The result of the model visualization shows that BDD-Net+ focuses on the collapsed areas of the damaged building. In addition, the model focused on all smooth parts of the building for classifying the nondamaged buildings. In a conclusion, BDD-Net+ is highly effective and generates informative deep features for BDD purposes. Furthermore, the Grad-CAM algorithm can be considered an informative tool for exploring the performance of the model. It helps to find out the weakness of the model and enhance the performance of the model according to it.

REFERENCES

- [1] K. Hacrefendioğlu, H. B. Başağa, and G. Demir, "Automatic detection of earthquake-induced ground failure effects through faster R-CNN deep learning-based object detection using satellite images," *Natural Hazards*, vol. 105, no. 1, pp. 383–403, 2021.
- [2] X. Ye et al., "Building-based damage detection from postquake image using multiple-feature analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 4, pp. 499–503, Apr. 2017.
- [3] S. T. Seydi and M. Hasanlou, "A new structure for binary and multiple hyperspectral change detection based on spectral unmixing and convolutional neural network," *Measurement*, vol. 186, 2021, Art. no. 110137, doi: [10.1016/j.measurement.2021.110137](https://doi.org/10.1016/j.measurement.2021.110137).

- [4] G. Mandlbürger, M. Kölle, H. Nübel, and U. Soergel, "BathyNet: A deep neural network for water depth mapping from multispectral aerial images," *PFG–J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 89, no. 2, pp. 71–89, 2021.
- [5] L. Knopp, "Development of a burned area processor based on sentinel-2 data using deep learning," *PFG – J. Photogramm., Remote Sens. Geoinf. Sci.*, vol. 89, pp. 357–358, 2021.
- [6] B. Aslam, A. Zafar, and U. Khalil, "Comparative analysis of multiple conventional neural networks for landslide susceptibility mapping," *Natural Hazards*, vol. 115, no. 1, pp. 673–707, 2023.
- [7] S. K. Kim and J. K. Hammit, "Hurricane risk perceptions and housing market responses: The pricing effects of risk-perception factors and hurricane characteristics," *Natural Hazards*, vol. 114, pp. 3743–3761, 2022.
- [8] C. Axel and J. A. N. van Aardt, "Building damage assessment using airborne lidar," *J. Appl. Remote Sens.*, vol. 11, no. 4, 2017, Art. no. 046024.
- [9] D. Aixia, M. Zongjin, H. Shusong, and W. Xiaoqing, "Building damage extraction from post-earthquake airborne LiDAR data," *Acta Geologica Sinica*, vol. 90, no. 4, pp. 1481–1489, 2016.
- [10] M. Janalipour and A. Mohammadzadeh, "Evaluation of effectiveness of three fuzzy systems and three texture extraction methods for building damage detection from post-event LiDAR data," *Int. J. Digit. Earth*, vol. 11, no. 12, pp. 1241–1268, 2018.
- [11] E. Ferrentino, F. Nunziata, C. Bignami, L. Graziani, A. Maramai, and M. Migliaccio, "Multi-polarization C-band SAR imagery to quantify damage levels due to the Central Italy earthquake," *Int. J. Remote Sens.*, vol. 42, no. 15, pp. 5969–5984, 2021.
- [12] Q. Chen, H. Yang, L. Li, and X. Liu, "A novel statistical texture feature for SAR building damage assessment in different polarization modes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 1, pp. 154–165, Dec. 2019, doi: [10.1109/JSTARS.2019.2954292](https://doi.org/10.1109/JSTARS.2019.2954292).
- [13] M. Hasanlou, R. Shah-Hosseini, S. T. Seydi, S. Karimzadeh, and M. Matsuoka, "Earthquake damage region detection by multitemporal coherence map analysis of radar and multispectral imagery," *Remote Sens.*, vol. 13, no. 6, 2021, Art. no. 1195.
- [14] C. Wu et al., "Building damage detection using U-Net with attention mechanism from pre- and post-disaster remote sensing datasets," *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 905.
- [15] G. Abdi and S. Jabari, "A multi-feature fusion using deep transfer learning for earthquake building damage detection," *Can. J. Remote Sens.*, vol. 47, no. 2, pp. 337–352, 2021.
- [16] Y. Qing et al., "Operational earthquake-induced building damage assessment using CNN-based direct remote sensing change detection on superpixel level," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, 2022, Art. no. 102899.
- [17] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, 2021, Art. no. 112636.
- [18] J. Ge, H. Tang, N. Yang, and Y. Hu, "Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 105–128, 2023.
- [19] H. Chen, E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley, "Dual-tasks Siamese transformer framework for building damage assessment," in *Proc. IGARSS-IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1600–1603.
- [20] Y. Shen et al., "BDANet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, May 2021, Art. no. 5402114, doi: [10.1109/TGRS.2021.3080580](https://doi.org/10.1109/TGRS.2021.3080580).
- [21] M. Li et al., "Post-earthquake assessment of building damage degree using LiDAR data and imagery," *Sci. China Ser. E, Technol. Sci.*, vol. 51, no. 2, pp. 133–143, 2009.
- [22] B. Adriano, J. Xia, G. Baier, N. Yokoya, and S. Koshimura, "Multi-source data fusion based on ensemble learning for rapid building damage mapping during the 2018 Sulawesi earthquake and tsunami in Palu, Indonesia," *Remote Sens.*, vol. 11, no. 7, 2019, Art. no. 886.
- [23] F. Eslamizade, H. Rastiveis, N. K. Zahraee, A. Jouybari, and A. Shams, "Decision-level fusion of satellite imagery and LiDAR data for post-earthquake damage map generation in Haiti," *Arabian J. Geosci.*, vol. 14, no. 12, 2021, Art. no. 1120.
- [24] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoatNet: Marrying convolution and attention for all data sizes," Sep. 15, 2021, *arXiv:2106.04803*.
- [25] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
- [26] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, no. 1, Sep. 2022, Art. no. 5412012, doi: [10.1109/TGRS.2022.3207551](https://doi.org/10.1109/TGRS.2022.3207551).
- [27] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10013–10021.
- [28] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.
- [29] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11916–11925.
- [30] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," Jun. 7, 2022, *arXiv:2104.05704*.
- [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.
- [32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
- [33] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Tversky as a loss function for highly unbalanced image segmentation using 3D fully convolutional deep networks," Jun. 18, 2018, *arXiv:1803.11078*.



Mr. Seydi is a regular Reviewer in about five international remote sensing journals.



thermal, optical, and SAR remote sensing for urban and agro-environmental applications.

Seyd Teymoor Seydi (Member, IEEE) received the B.Eng. degree in surveying and geomatics engineering from the University of Shahid Rajaei, Tehran, Iran, in 2015, and the M.Eng. degree in remote sensing from the University of Tehran, Tehran, in 2018.

He has authored or coauthored more than 35 peer-reviewed journal and conference papers. His research interests include multitemporal multispectral/hyperspectral and SAR remote sensing processing and classification, and advance deep learning algorithms.

Mahdi Hasanlou (Member, IEEE) received the B.Sc. degree in surveying and geomatics engineering, and the M.Sc. and Ph.D. degrees in remote sensing from the University of Tehran, Tehran, Iran, in 2003, 2006, and 2013, respectively.

Since 2013, he has been an Assistant Professor with the School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, where he is currently the Head of the Remote Sensing Laboratory and the Remote Sensing And Photogrammetry Group. His research focuses on hyperspectral,



Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree in electrical engineering from the Universit  de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He was a Visiting Scholar with Stanford University, Stanford, CA, USA, KTH, Stockholm, Sweden, and NUS, Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland, and the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. In 2015–2017, he was a Visiting Professor with the University of California, Los Angeles, Los Angeles, CA, USA. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Prof. Chanussot holds the AXA Chair in remote sensing with the Chinese Academy of Sciences, Aerospace Information research Institute. He is the Founding President of IEEE Geoscience and Remote Sensing French Chapter (2007–2010), which received the 2010 IEEE GRS-S Chapter Excellence Award. He was the recipient of multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017–2019). He was the General Chair of the first IEEE GRSSWorkshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing. He was the Chair (2009–2011) and Co-Chair of the GRS Data Fusion Technical Committee (2005–2008). He was a Member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and PROCEEDINGS OF THE IEEE. He was the Editor-In-Chief of IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015). In 2014, he was a Guest Editor for the IEEE Signal Processing Magazine. He is a Member of the Institut Universitaire de France (2012–2017) and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters).



Pedram Ghamisi (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Iceland, Reykjavik, Iceland, in 2015.

He is the Head of the Machine Learning Group with Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany, and a Visiting Professor and Group Leader of AI4RS with the Institute of Advanced Research in Artificial Intelligence (IARAI), Vienna, Austria. He is a Co-Founder of VasoGnosis, Inc., with two branches in San Jose and Milwaukee, WI, USA. His research interests include interdisciplinary research on machine (deep) learning, image and signal processing, and multisensor data fusion.

Dr. Ghamisi was the recipient of the IEEE Mikio Takagi Prize for winning the Student Paper Competition at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, First Prize of the Data Fusion Contest organized by the IEEE IADF in 2017, Best Reviewer Prize of IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2017, and IEEE Geoscience and Remote Sensing Society 2020 Highest-Impact Paper Award. He is an Associate Editor for IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and IEEE GEOSCIENCE AND REMOTE SENSING LETTERS. He was the Co-Chair of the IEEE Image Analysis and Data Fusion Committee (IEEE IADF) between 2019 and 2021.