



UNet-GCViT: a UNet-based framework with global context vision transformer blocks for building damage detection

Masoomeh Gomroki, Mahdi Hasanlou, Jocelyn Chanussot & Danfeng Hong

To cite this article: Masoomeh Gomroki, Mahdi Hasanlou, Jocelyn Chanussot & Danfeng Hong (24 Jan 2025): UNet-GCViT: a UNet-based framework with global context vision transformer blocks for building damage detection, International Journal of Remote Sensing, DOI: [10.1080/01431161.2025.2454531](https://doi.org/10.1080/01431161.2025.2454531)

To link to this article: <https://doi.org/10.1080/01431161.2025.2454531>



Published online: 24 Jan 2025.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



UNet-GCViT: a UNet-based framework with global context vision transformer blocks for building damage detection

Masoomeh Gomroki ^a, Mahdi Hasanlou ^a, Jocelyn Chanussot ^{b,c}
and Danfeng Hong ^c

^aSchool of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran;
^bCNRS, Grenoble INP, GIPSA-Lab, University Grenoble Alpes, Grenoble, France; ^cAerospace Information
Research Institute, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Natural and man-made disasters take place around the world and cause significant financial and human losses. An accurate and fast post-disaster building damage mapping could play a crucial role in rapid rescue planning and operations. Remote sensing satellite images are the main source of building damage map generation. Usually, both pre-disaster and post-disaster satellite images are used for the generation of building damage maps, which encounter some challenges such as registration errors, noise, and atmospheric conditions. This study proposed a new method for building damage detection based only on post images with the UNet architecture network and Global Context Vision Transformer blocks. The deep learning network proposed in this research is automatic without any further processing. The proposed method comprises four main steps: (1) pre-processing, (2) network training, (3) building damage map generation, and (4) accuracy assessment for the final damage map. This network is applied to three different natural and man-made disaster datasets. The first dataset is the post-satellite image of the 2023 Turkey earthquake, the second one is the post-satellite image of the 2021 Bata explosion and the last one is the post-satellite image of the 2011 haiti earthquake. Results of the final building damage map indicate that the proposed method is highly effective with OA above 96%, which is superior to the other deep learning methods.

ARTICLE HISTORY

Received 14 May 2024
Accepted 8 January 2025

KEYWORDS

Building damage detection;
GCViT blocks; unet;
turkey earthquake;
bata explosion;
haiti earthquake;
natural disaster;
man-made disaster

1. Introduction

The disasters cause extensive damage to buildings, as one of the most important elements in cities (Hacıfendioğlu, Başağa, and Demir 2021). Determining the position of damaged buildings in the early hours of a disaster is the foremost undertaking for emergency relief (Dell'acqua and Gamba 2012; Eguchi et al. 2010; Nex et al. 2019). To address this issue, remote sensing (RS) data have been utilized as efficient and fast tools for extracting the required information about damaged buildings after such sudden events (Nex et al. 2019). Recent advances in RS data capturing and computer vision

techniques have raised researchers' interest in investigating building damage detection (Cotrufo et al. 2018; Seydi et al. 2023a).

As reported in the state-of-the-art research in this field, light detection and ranging (LiDAR) data and point clouds (Hong et al. 2023; Talreja et al. 2021), synthetic-aperture radar (SAR) images (Bolorani et al. 2021; ElGharbawi and Zazoura 2021; Wang et al. 2022), very-high-resolution (VHR) satellite images (D'Addabbo, Pasquariello, and Amodio 2021; Mohammadian and Ghaderi 2023; Yeom, Jung, and Kim 2017), unmanned aerial vehicle (UAV) images (Khankeshizadeh et al. 2024, Naito et al. 2020), and multimodal data such as VHR together with LiDAR (Li et al. 2008), and various types of satellite images (Adriano et al. 2019) have been exploited to assess the severity of building damages after the natural and man-made event. Thanks to easy access, simple display, low costs, and the possibility of obtaining both pre- and post-event satellite data, such data have been employed over and above other RS data for building damage map generation purposes (Seydi et al. 2023a).

Developing damage detection methods with high-accuracy results in acceptable speed is a serious challenge, particularly in recent years in which disasters such as earthquakes have hit different urban areas with their complexities and different levels of damage. In recent years, Deep Learning (DL) methods have been extensively employed in various applications, such as two- and three-dimensional (2D and 3D) change detection (Gomroki, Hasanlou, and Chanussot 2023, 2023; Lian et al. 2020; Yadav, Nascetti, and Ban 2022), urban planning (Chen et al. 2018), and disaster monitoring (Zheng et al. 2021).

Overall, the procedures adopted for building damage mapping using satellite images have been split into two main categories: (1) building damage detection through two-event (namely, pre-event and post-event) data, and (2) building damage detection via single-event (i.e. post-event) data. The methods in the first category are more popular. As an illustration, Zheng et al. (2021) used an object-based method in combination with DL for building damage detection. Wu et al. (2021) used UNet architecture with the attention mechanism on VHR-related pre-event and post-event satellite images for building damage detection. In addition, Chen et al. (2022) developed a method based on the encoder-decoder and transformer networks for building damage detection using pre-event and post-event data. They extract non-local deep features from both pre-event and post-event images by the transformer encoder, and then the fuse module integrates the given features. Finally, the dual-task decoder aggregates the hierarchical features for mapping damaged buildings. In addition, Gupta and Shah (2021) introduced a RescueNet network, a combination of ResNet50 and the Atrous Spatial Pyramid Pooling (ASPP) module to extract multiscale features that could simultaneously measure the severity of damages by segmenting buildings. Shen et al. (2021) developed a dual-branch Unet architecture that detects building damages using pre-event and post-event images separately. Ji et al. (2020) employed the pre-trained Visual Geometry Group (VGG) network for building damage detection. Additionally, Merlin and Jiji (2019) used a multistep method for building damage mapping through pre-event and post-event images. In their study, the buildings are first detected via thresholding and colour invariance, and then the changes are identified by image differencing. Afterwards, the spectral and spatial features are extracted and selected using feature-ranking techniques, and they are finally

classified by the mean ratio for feature selection (MRFS) algorithm. Also, Ji et al. (2019) detected damaged buildings on VHR pre-event and post-event datasets by the combination of the random forest (RF) classification with the convolutional neural network (CNN) architecture to have higher accuracy than the CNN architecture alone.

The methods under the second category make use of only post-event data for building damage detection as one of their advantages. Moreover, these algorithms have no need for additional processing, such as image-to-image registration for this purpose. For example, Ünlü and Kiriş (2022) utilized a two-phase deep learning (DL) method. In the first phase, the K-means algorithm was exploited for image segmentation and then, the damaged and non-damaged labels were placed on the data. During the second phase, the VGG16 network was recruited to divide the labelled data into less damaged, damaged, and non-damaged classes. Moreover, Li et al. (2019) evaluated the severity of damage with an autoencoder network using the VGG16 network on post-event images.

Even though the methods in the first category have achieved acceptable results in building damage detection, sometimes finding pre-event data may be challenging. In addition, different conditions during data acquisition before and after a disaster could affect the result of algorithms and bring about some errors (Janalipour and Mohammadzadeh 2019; Khankeshizadeh et al. 2024; Seydi et al. 2023a). Most of the above-mentioned methods have generally benefitted from both pre- and post-event data for building damage detection, but sometimes the pre-event data have not been available. Furthermore, extracting the changes from two-time data typically encountered some challenges such as registration error, atmospheric condition, noise, etc (Janalipour and Mohammadzadeh 2019; Khankeshizadeh et al. 2024; Seydi et al. 2023b).

Considering the capabilities of DL methods and the challenges in building damage detection, a network based on the UNet architecture using the global context vision transformer (GCViT) architecture blocks is presented in this study. The proposed method utilizes the three different datasets of post-event satellite images and achieves high accuracy. This method compared to the other method that only uses post-event data has a higher speed in generating damaged building maps also it has lower parameters and is GPU-friendly.

The main contributions of this study were as follows: (1) The GCViT architecture blocks are employed in the form of a UNet, which helped reduce its parameters compared to the original GCViT architecture and increased its speed and efficiency; in other words, the network was lightened. (2) The UNet-GCViT framework was applied for the first time in this study for building damage detection. (3) The performance of the GCViT architecture blocks in the form of a UNet was compared with other networks with high-speed performance such as FastViT, indicating the generation of accurate and fast damage maps by the proposed method. (4) Only the post-event images are used in the proposed method without any auxiliary data. (5) The proposed method is applied to various natural and man-made disasters such as earthquakes and explosions.

The material and dataset are explained in the second section. The proposed method is discussed in the third section and the experimental results and the comparison are expressed in the fourth section. The discussion is provided in the fifth section and the conclusion is presented in the last section.

2. Materials and datasets

2.1. 2023 Turkey earthquake (natural disaster)

The performance of the UNet-GCViT framework was evaluated with the Turkey earthquake (with a magnitude of 7.8 on the Richter scale) dataset, which occurred on 6 February 2023, at 4:17 am (1:17 UTC) in the southern and central parts of Turkey southern and western Syria. The study area was the city of Nurdagi in Turkey, suffering much post-earthquake damage. Figure 1 illustrates the study area and the corresponding ground truth. To generate the ground truth, the post-event data and corresponding Google Earth images are used. The polygons related to the damaged and non-damaged buildings are drawn in Global Mapper manually and then layered with different numbers to generate the ground truth. The number of polygons of the non-damaged and damaged building is equal to 1833 and 400, respectively. The description of the dataset is provided in Table 1. The satellite images of the Turkey earthquake in this area have been made available by Maxar Technologies Inc. at (<http://www.maxar.com/open-data>).

2.2. 2021 bata explosion (man-made disaster)

On 7 March 2021, in the afternoon, a series of explosions took place at the Nkuantoma armoury and military barracks in Bata, Equatorial Guinea's economic centre. Figure 2

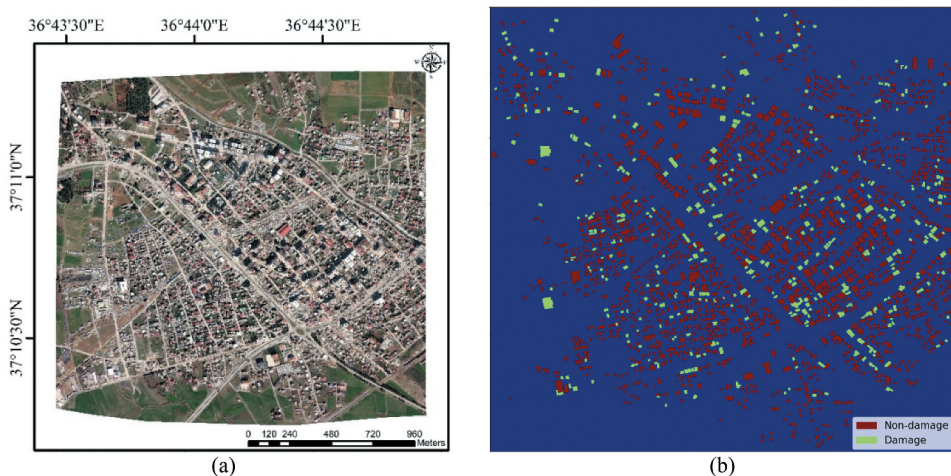


Figure 1. Turkey earthquake dataset: (a) RGB post image and (b) ground truth.

Table 1. Information on the dataset used in this study.

Dataset	Turkey Earthquake	Bata Explosion	Haiti Earthquake
satellite	World View II	World View II	World View II
Spectral bands	3	3	3
Special resolution (m)	0.5	0.5	0.5
Acquisition Date	11 February 2023	9 March 2021	16 January 2011
Damage Building(polygons)	400	368	497
Non-damage Buildings(polygons)	1833	338	943

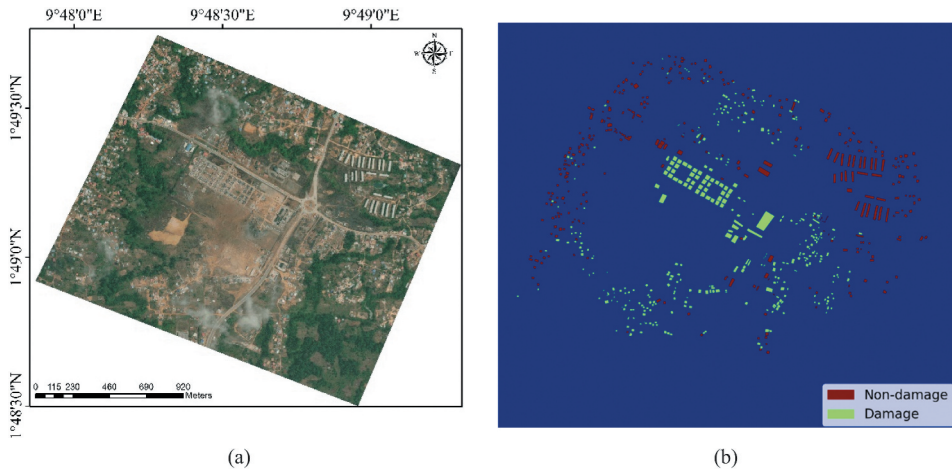


Figure 2. Bata explosion dataset: (a) RGB post image and (b) ground truth.

displays the WorldViewII image of the Bata explosion region and the corresponding ground truth. To generate the ground truth, we operate like the 2023 Turkey earthquake dataset. In this dataset, there are 338 non-damaged polygons of buildings and 368 damaged ones. Additional information about this dataset is given in [Table 1](#).

2.3. 2010 Haiti earthquake (natural disaster)

On 12 January 2010, at 16:53 UTC, a 7.0 magnitude earthquake struck the western part of Haiti, approximately 25 km south of Port-au-Prince. [Figure 3](#) displays the WorldViewII image of this region and the corresponding ground truth. To generate the ground truth, we operate like two previous datasets. In this dataset, there are 943 non-damaged polygons of buildings and 497 damaged ones. Additional information about this dataset is given in [Table 1](#).

3. Proposed method

The proposed method in this study had three major steps (1) data pre-processing, (2) the UNet-GCViT framework training for building damage detection, and (3) prediction using the trained network. [Figure 4](#) depicts the general flowchart of the proposed method introduced in the present study.

In a pre-processing step, the images are normalized and then the train and validation data are separated from the test ones. In addition, all data are augmented using the rotation method with angles of $+90^\circ$, $+180^\circ$ and $+270^\circ$. The network used in this study has a UNet architecture, in which the GCViT blocks are considered as its encoder part.

3.1. GCViT architecture blocks

The GCViT is a novel architecture based on vision transformer (ViT), which was first presented by Hatamizadeh et al. (2023). In this method, the powerful tools of the global

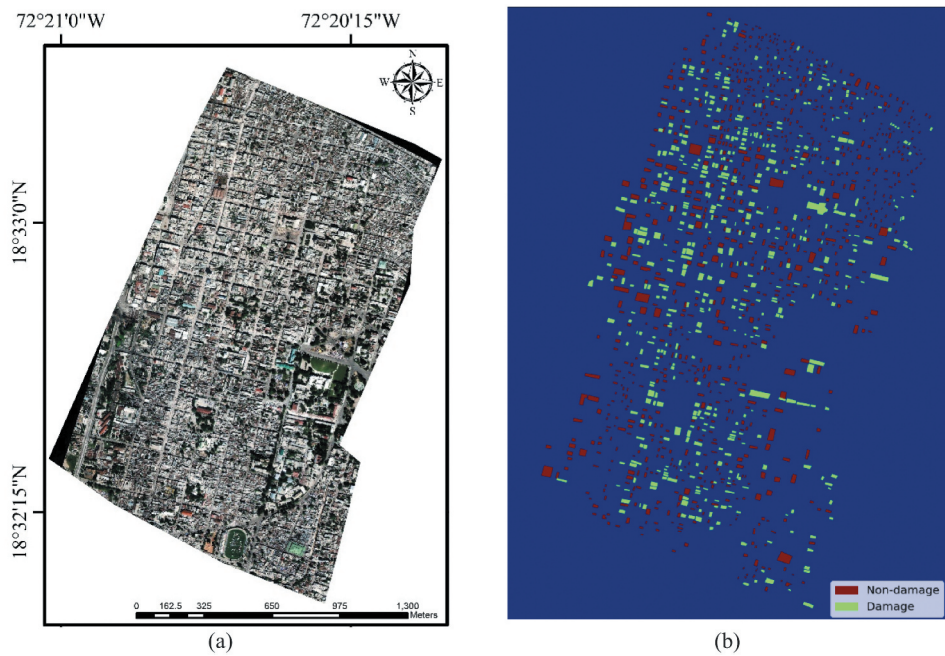


Figure 3. Haiti earthquake dataset: (a) RGB post image and (b) ground truth.

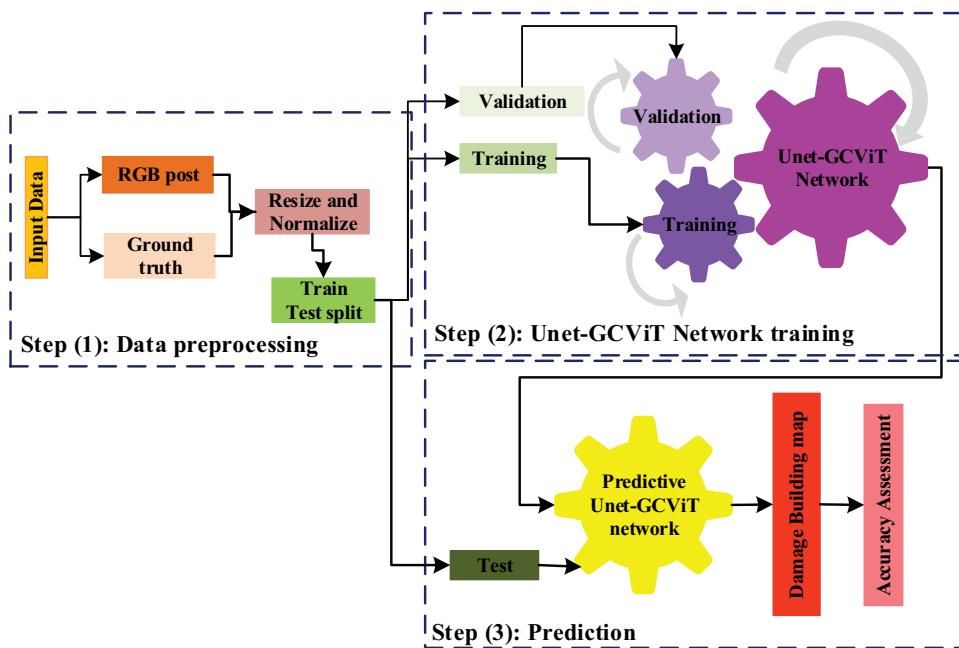


Figure 4. The flowchart of the proposed method.

context self-attention module and standard local self-attention are thus combined to enable the network to perform both short-range and long-range spatial interactions, without costly and time-consuming additional processing, such as shifting local windows or computing attention masks. In that network, the problem of inductive bias in ViTs was also solved (Hatamizadeh et al. 2023). The proposed method uses GCViT blocks in Unet architecture, this architecture has higher speed in training and generating the final damage map. Moreover, this architecture has fewer parameters than the main GCViT network so it is GPU-friendly. The GCViT architecture blocks include three main parts: (1) downsampling operator, (2) global query generator, and (3) global self-attention module, as discussed below.

3.1.1. Downsampling operator

The structure of the downsampling operator blocks is a modified Fused-MBConv includes a pooling layer with a kernel size of three and a stride of two. The Fused-MBConv block in the GCViT network is a modified form of the EfficientNetV2 blocks (Tan and Le 2021), and the modifications compared to the main blocks of the EfficientNetV2 network are as follows:

$$\begin{aligned}\hat{x} &= DW - Conv_{3 \times 3}(x), \\ \hat{x} &= GELU(\hat{x}), \\ \hat{x} &= SE(\hat{x}), \\ x &= Conv_{1 \times 1}(\hat{x}) + x,\end{aligned}\tag{1}$$

where SE , $GELU$ and $DW - Conv_{3 \times 3}$ denotes Squeeze and Excitation block (Hu, Shen, and Sun 2018), Gaussian Error Linear Unit (Hendrycks and Gimpel 2016), and 3×3 depth-wise convolution, respectively. In this architecture, the Fused-MBConv blocks act like the inductive bias.

3.1.2. Global query generator

The global query generator has been designed to: (1) convert the input feature image into the current step, (2) extract the features by repeating $\log_2 \frac{H}{h}$ times the converted Fused-MBConv blocks and then concatenate them with down-sampling for dimension matching of the window with size h (as the channel dimensions), and (3) reshape the output and repeat it $\left(\frac{H}{h}\right)^2$ times as the number of local tokens until they can attend to the global contextual information (Hatamizadeh et al. 2023). Figure 5 displays the schematic diagram of the global query generator.

3.1.3. Global self-attention module

The global self-attention module is defined as follows:

$$Attention(q_g, k, v) = Softmax\left(\frac{q_g k}{\sqrt{d}} + b\right)v\tag{2}$$

in which, q_g , k and v are considered Global query tokens, keys, and values, respectively. d and b represent the scale factor and the learnable relative position bias, respectively. Assuming that the variations between $[-p + 1, p - 1]$ are along the horizontal and vertical axes make positive changes, so b is a sampling from the grid

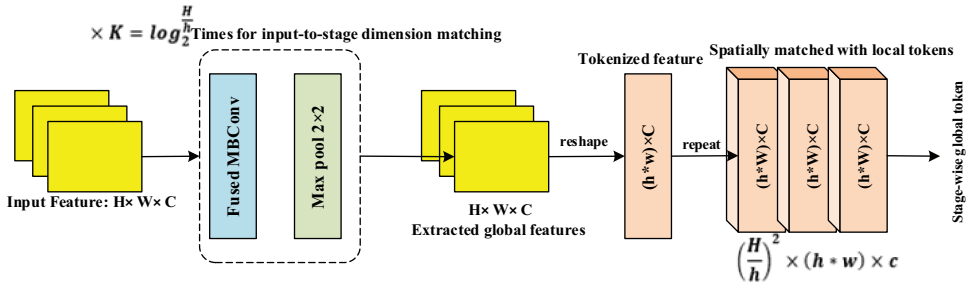


Figure 5. Global query generator diagram.

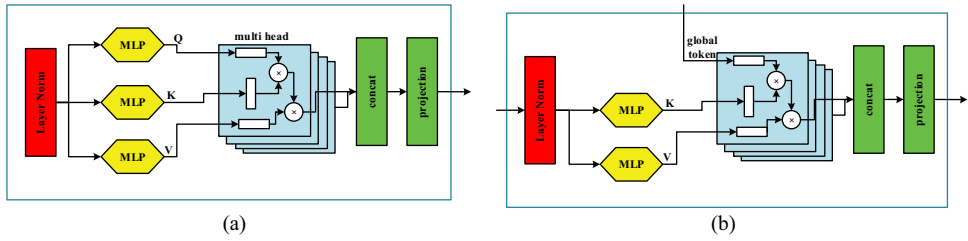


Figure 6. Comparison between local and global attention blocks; (a) local MSA, (b) global MSA.

$\hat{b} \in \mathbb{R}^{(I_1 - J_1) \times (J_1 - J_2)}$ (Hatamizadeh et al. 2023). Figure 6 shows the difference between the global and local attention blocks. The global attention block does not calculate the query vector and reuses the global query computed through the global token generation (Hatamizadeh et al. 2023).

3.2. UNet-GCViT framework training for building damage detection

The GCViT architecture has been pre-trained by the ImageNet dataset (Hatamizadeh et al. 2023). Since only the encoder part of UNet was pre-trained, the semi-transfer learning technique was employed in this study (Gomroki, Hasanlou, and Reinartz 2023). The improved UNet blocks (Gomroki, Hasanlou, and Chanussot 2023) were further considered as the decoder part. The method proposed in this study for building damage detection has acceptable speed with high accuracy. In addition, it is completely automatic and does not require additional processing, as well as it only uses post-event data for detecting building damage.

In this study, the GCViT architecture blocks of the XTiny version are employed, whose architecture details are given in Figure 7. As seen, it has four stages and one stem which are concatenated with Conv2D Transpose blocks in different sizes, creating the UNet-GCViT framework.

3.3. State-of-the-art methods

The performance of utilizing GCViT blocks in the UNet architecture is compared with the other DL networks. As mentioned earlier, one of the contributions of this study is to

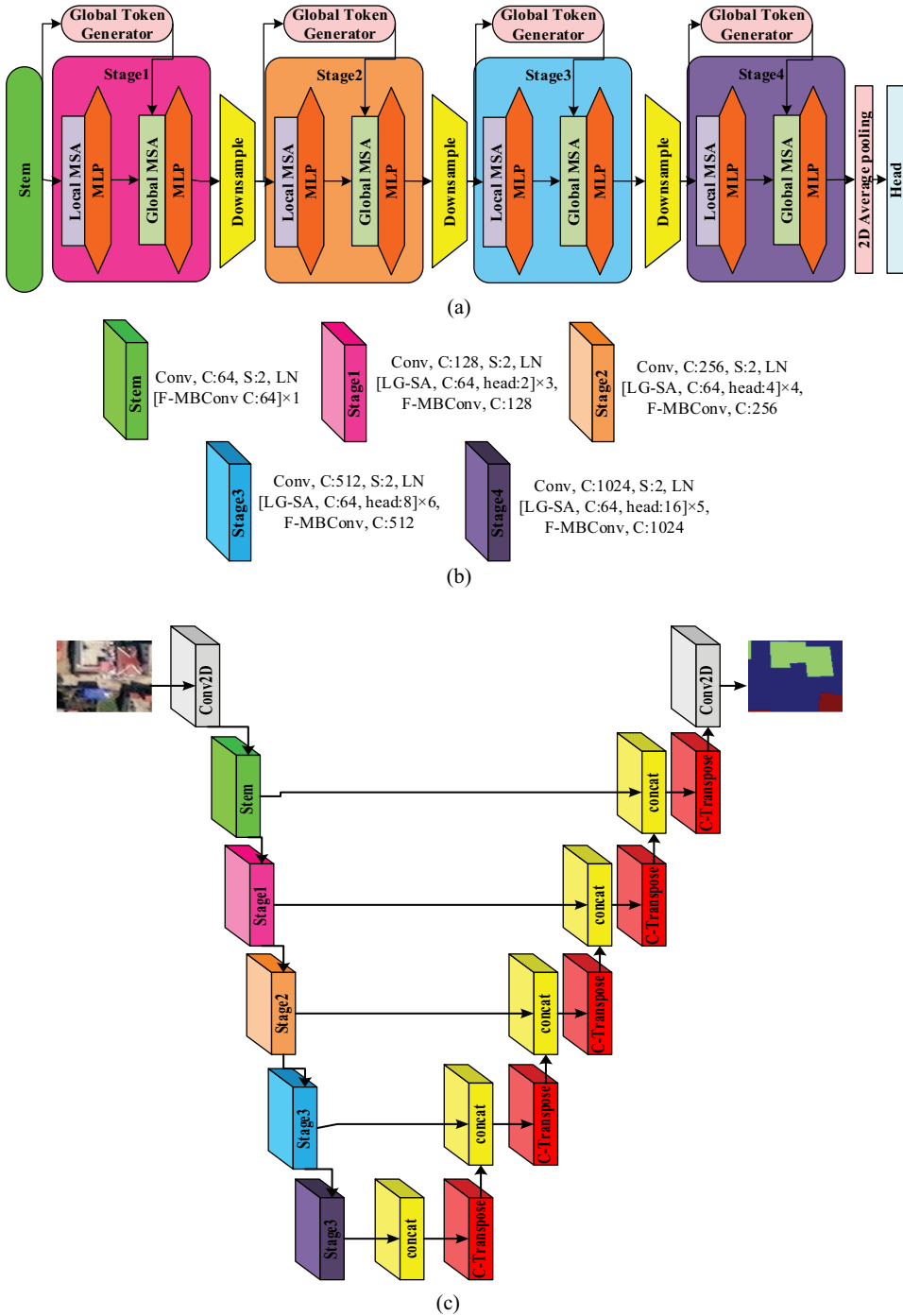


Figure 7. The architecture of unet-GCViT; (a) architecture of GCViT network blocks. At each stage, a query generator extracts global query tokens that capture long-range information by interacting with local key and value representation, alternating blocks of local and global context self-attention layers are used (b) the structure of GCViT XTiny network which contains one stem and four stages (represented with a different color) and LG-SA and Conv denotes local, global self-attention and

develop a network to reach high accuracy for building damage detection with suitable speed. From this perspective, the performance of the proposed network blocks is compared with other networks' blocks, such as FastViT (Vasu et al. 2023), FasterNet (Chen et al. 2023), ConvNextV2 (Woo et al. 2023), TinyViT (Wu et al. 2022) and EfficientformerV2 (Li et al. 2022) as an encoder of the Unet architecture.

4. Experimental results

The performance of the proposed framework is compared with other networks, considering experimental parameters, evaluation metrics, and comparison of experimental results.

4.1. Experimental parameter settings

All DL networks in this study are implemented using TensorFlow 2.10.0 and Python 3.8.8 through a system with an NVIDIA GeForce GTX 1050Ti GPU, a 3.5 GHz CPU and 32 GB RAM. The data in this study are divided into patches with the dimensions of 128×128 pixels. Among the dataset, 60% is considered as the train-validation data and 40% as the test ones. Moreover, in the Turkey earthquake dataset, 1166 patches are counted as the test data and 1749 patches are available as the train-validation ones before augmentation. Following augmentation using the rotation method with the angles of $+90^\circ$ and $+180^\circ$, this value reached 5247 patches. In the Bata explosion, 434 and 651 patches are considered as test and train-validation, respectively. After augmentation, the train-validation patches reached 2604 patches. In the Haiti earthquake, 437 and 655 patches were considered as test and train, respectively. After augmentation, the train-validation patches reached 2620 patches. In the proposed framework, Adam optimizer was utilized with the learning rate and learning decay of $1e-4$ and $1e-6$, respectively. The GCViT architecture blocks have been pre-trained by the ImageNet dataset. The network is thus implemented for 100 epochs and the focal loss function is applied for multiple classifications, defined according to the following equation (Liu, Chen, and Chen 2018):

$$L_{focalloss} = - \sum_{i=1}^c \alpha_i (1 - y_i)^\gamma t_i \log(y_i) \quad (3)$$

which, c , t_i , and y_i represent the number of classes, the real probability distribution, and the probability distribution of the prediction, respectively. As well, γ and α_i are the hyperparameters of the focal loss function, in which γ is considered equal to 2 and α_i is the weight of the classes. The performance of SGD, AdamW, and Adam optimizers was computed and Adam had a proper result. The learning rate of the Adam optimizer tested for values of $\{10^{-2}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, and 10^{-3} had better performance.

3×3 convolutional layer, respectively, (c) the structure of the proposed method using GCViT network blocks as an encoder. The encoder is concatenated with the decoder at five different resolutions (stem, stage1, stage2, stage3, and stage4), C-Transpose is transpose convolution and concat is concatenation.

4.2. Evaluation metrics

The evaluation metrics investigated in this study are precision, F1-score, intersection over union (IOU), overall accuracy (OA), and the kappa coefficient (KC). Furthermore, TP indicates the true positive, FP shows the false positive, and FN and TN represent the false negative and the true negative, respectively. The evaluation metrics are expressed using the equation below:

$$\begin{aligned}
 \text{precision} &= \frac{TP}{TP + FP} \\
 \text{F1-score} &= \frac{2 \times TP}{(2 \times TP) + FP + FN} \\
 \text{IOU} &= \frac{TP}{TP + FP + FN} \\
 \text{Overall Accuracy (OA)} &= \frac{TP + TN}{TP + FN + TN + FP} \\
 \text{Kappa Coefficient (KC)} &= \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP)(FP + TN) + (TP + FN)(FN + TN)}
 \end{aligned} \tag{4}$$

4.3. Comparison of experimental results

The evaluation metrics explained in the previous section are examined to evaluate the performance of the networks. The number of parameters of original networks is compared with those when network blocks are employed in our Unet architecture in Table 2. The main GCViT architecture has 19.9 M parameters, which are reduced to 12.4 M after its blocks are employed in the UNet architecture i.e. the GCViT is lightened and therefore, its training speed typically increases. One of the main objectives of this study is post-disaster building damage map generation with the highest possible speed so the given network is favourable.

In the Turkey earthquake dataset, comparing the performances of the networks outlined in Table 2 reveals that the proposed method reached the highest OA (96.71) and the highest KC (0.85). Considering high training speed among the main objectives of this study, FasterNetV2 has the highest training speed (namely, 2 h, 48 m, and 20 s), but fails to achieve proper results and has a very low KC compared to other networks. The performance of FasterNet V2 and ConvNeXt are poor and they could not handle the problem of building damage detection. The main objective of the FastViT network which is based on a transformer architecture has been to achieve the highest speed in training (Vasu et al.

Table 2. Quantitative evaluation of the results of the Turkey earthquake.

Method	Accuracy (%)	Precision (%)	F1-Score (%)	IOU (%)	Kappa Coefficient (KC)	Time of training (h min sec)	Parameters (Million)	
							main network	Unet Blocks
EfficientformerV2L	95.88	99.66	99.36	98.74	0.80	4h 36min 40sec	26.1	6.9
TinyViT	96.19	99.73	99.41	98.83	0.83	4h 35min 0sec	5.3	7.4
ConvNeXt V2	75.03	79.88	85.89	71.99	0.13	5h 0min 0sec	3.7	12.8
FasterNet V2	75.43	79.31	89.99	72.22	0.14	2h 48min 20sec	7.6	11.4
FastViT	96.52	99.78	99.45	98.65	0.83	4h 16min 40sec	11.6	16.3
Proposed method	96.71	99.79	99.48	98.98	0.85	4h 35min 5sec	19.9	12.4

Table 3. Quantitative evaluation of the results of the bata explosion.

Method	Accuracy (%)	Precision (%)	F1-Score (%)	IOU (%)	Kappa Coefficient (KC)	Time of training (h min sec)	Parameters (Million)	
							main network	Unet Blocks
EfficientformerV2L	97.01	99.90	99.71	98.44	0.77	2h 18min 20sec	26.1	6.9
TinyViT	97.98	99.89	99.72	98.43	0.76	2h 16min 20sec	5.3	7.4
FastViT	97.13	99.98	99.70	98.40	0.75	2h 8min 20sec	11.6	16.3
Proposed method	98.74	99.92	99.85	99.72	0.81	2h 17min 40sec	19.9	12.4

Table 4. Quantitative evaluation of the results of the hati earthquake.

Method	Accuracy (%)	Precision (%)	F1-Score (%)	IOU (%)	Kappa Coefficient (KC)	Time of training (h min sec)	Parameters (Million)	
							main network	Unet Blocks
EfficientformerV2L	97.08	99.70	99.25	98.51	0.69	2h 18min 20sec	26.1	6.9
TinyViT	96.75	99.49	99.15	98.33	0.67	2h 20min 0sec	5.3	7.4
FastViT	96.76	99.61	99.19	98.39	0.66	2h 8min 20sec	11.6	16.3
Proposed method	98.75	99.91	99.31	98.73	0.72	2h 17min 55sec	19.9	12.4

2023). The speed of our proposed method is comparable with that of FastViT, but it reaches a higher score in all Table 2 metrics. The proposed method compared to the other method has a higher speed in generating damaged building maps also it has lower parameters and is GPU-friendly so its computational cost is lower than other methods.

To shed light on the performance of the proposed method for building damage detection in natural and man-made disasters, it is applied to the 2021 Bata explosion and the 2010 haiti earthquake datasets. According to Table 3, the OA and KC of this network are, respectively, 98.74% and 0.81 for the 2021 Bata explosion dataset which shows its proper performance in comparison to other methods.

According to Table 4, the proposed network can reach an overall accuracy (OA) of 98.75% and kappa coefficient (KC) of 0.72 for the 2010 haiti earthquake dataset, indicating its better performance than other networks. In terms of other parameters such as precision, intersection over union (IoU) and the F1-score, the proposed network reaches 99.91%, 98.73%, and 99.31%, respectively. Compared to the FastViT network which is utilized commonly in high-speed tasks, the proposed network speed is comparable and shows better performance. Although the TinyViT network is lightweight and has fewer parameters due to its structure, its performance is not as proper as the proposed method because it cannot handle the complexities of building damage detection.

4.4. Ablation study

In the ablation study, the existence and non-existence of some parts of the proposed method and their effects on the results are discussed. Since the proposed method is based on GCViT blocks in the encoder, Table 5 indicates the influence of the Global query token as an ablation study in the Turkey earthquake dataset. The Global query token plays a major role in the architecture of the GCViT blocks so it is considered for ablation study. The S#1 model is designed without the Global query token and the S#2 model is designed with the Global query token (proposed method).

Table 5. Ablation study of the proposed method for the Turkey earthquake dataset.

Method	OA (%)	KC	Precision (%)	IOU (%)	F1-score (%)
S#1	96.01	0.83	99.75	98.88	99.34
S#2	96.71	0.85	99.79	98.98	99.48

As outlined in Table 5 the overall performance of the given method decreases when the Global query token is eliminated.

5. Discussion

The performance of the proposed network for building damage map detection is investigated using qualitative and quantitative evaluation as well as visual analysis. Figure 8 shows the performance of the proposed network with the other ones listed in Table 2 for the 2023 Turkey earthquake. In this study, the buildings that are partially damaged or completely destroyed are considered as the class of damaged buildings, which are marked with green, and the non-damaged ones are shown in crimson in Figure 8.

According to Figure 8, the results of networks employed by the FasterNet and ConvexNetV2 blocks are very poor. According to the number of the damaged buildings compared to the non-damaged ones, the dataset is unbalanced but the proposed network could handle it. In addition, since only post-event data are used, it is difficult to detect the boundaries of the damaged buildings, particularly those that had been completely destroyed, but the proposed method could correctly detect them.

To scrutinize the performance comparison of the proposed method with other networks (Figure 9), three areas of the 2023 Turkey earthquake dataset investigated in this study are selected. In Area 1 the damaged and non-damaged buildings are densely located next to each other. As observed the proposed method can correctly detect the boundaries of the buildings and make a distinction between the damaged and non-damaged ones. In contrast, the FastViT method has wrongly detected the non-damaged buildings in some areas, and the TinyViT and EfficientFormer methods have failed to correctly detect the boundaries of the buildings, so the polygons have not been correctly reconstructed by these two methods. In Area 2, the damaged and non-damaged buildings have relatively complex shapes and boundaries, but the proposed network can correctly detect them and separate the damaged buildings from the non-damaged ones. As incompetency, additional non-damaged buildings are observed in the final map generated by the FastViT, TinyViT, and EfficientFormer methods, which are wrongly detected. In Area 3, the damaged and non-damaged buildings are located next to each other and the proposed network and FastViT could correctly distinguish their boundaries and shapes, but TinyViT and EfficientFormer do not perform well and the polygons of the damaged buildings are not correctly detected. According to Figures 8 and 9, the network proposed in this study could generate a map of the damaged buildings with high accuracy, and the polygons and boundaries of the buildings are correctly detected.

The second dataset used in this study is the 2021 Bata explosion. As shown in Figure 10, The proposed method can separate the damaged buildings from the non-damaged ones. Furthermore, one of the challenges in this method is the lack of pre-event data so it is not easy to detect the boundary of the buildings that are completely damaged, but the

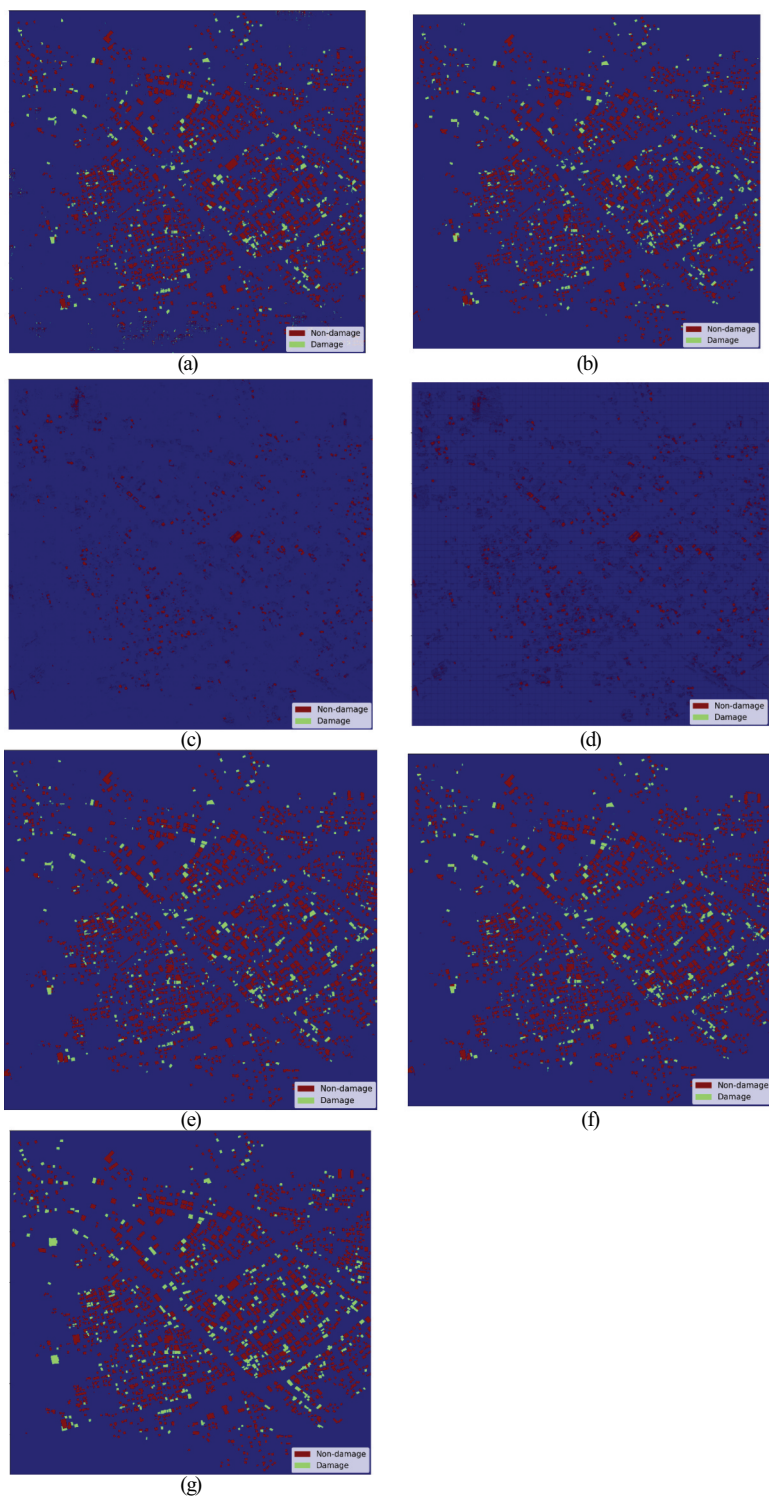


Figure 8. Comparison between the various networks: (a) EfficientformerV2L, (b) TinyViT, (c) ConvNeXt V2, (d) FasterNet T2, (e) FastViT, (f) proposed method, and (g) ground truth.

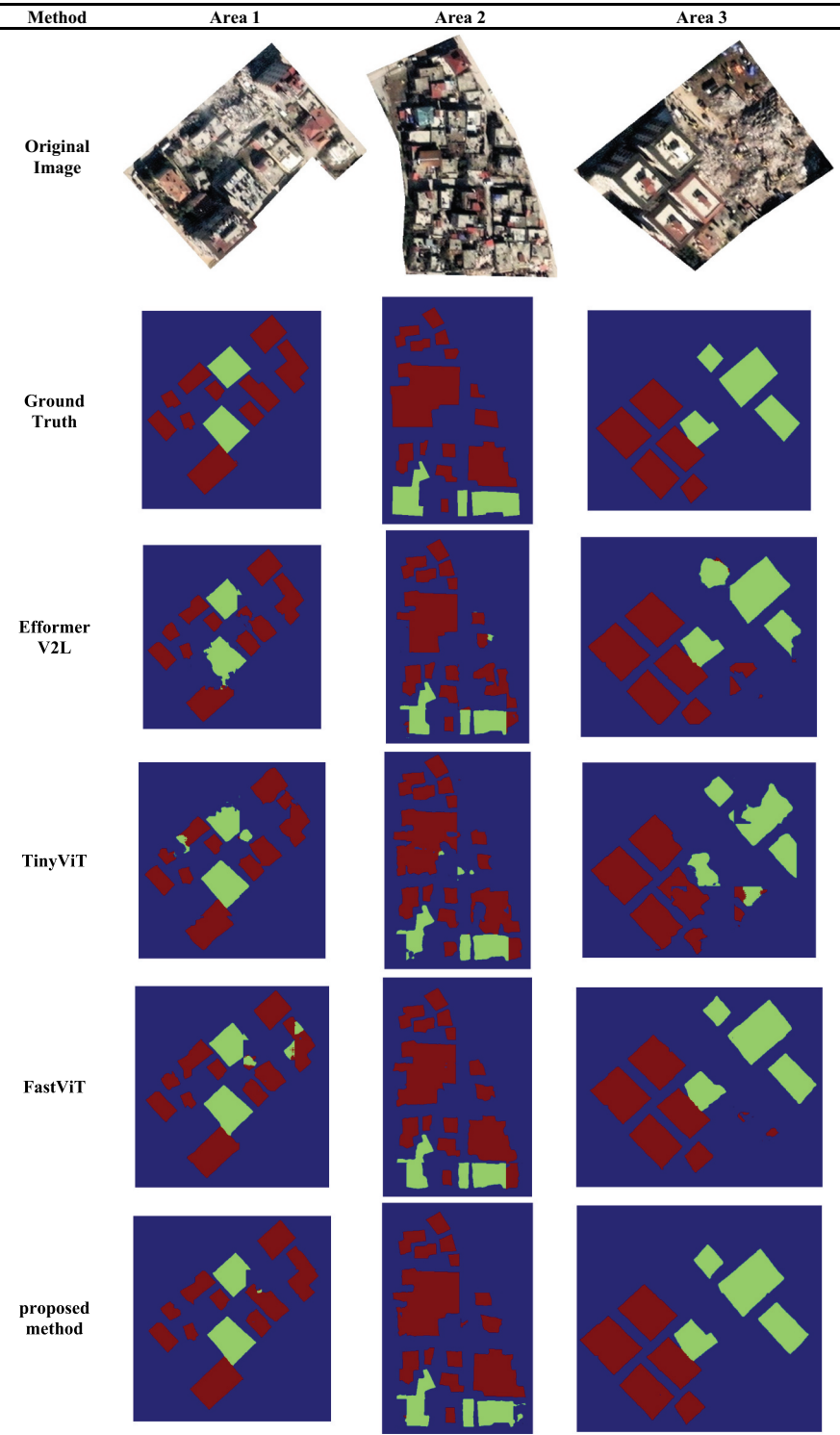


Figure 9. Comparison of the results of UNet-GCViT by different networks for Turkey earthquake.

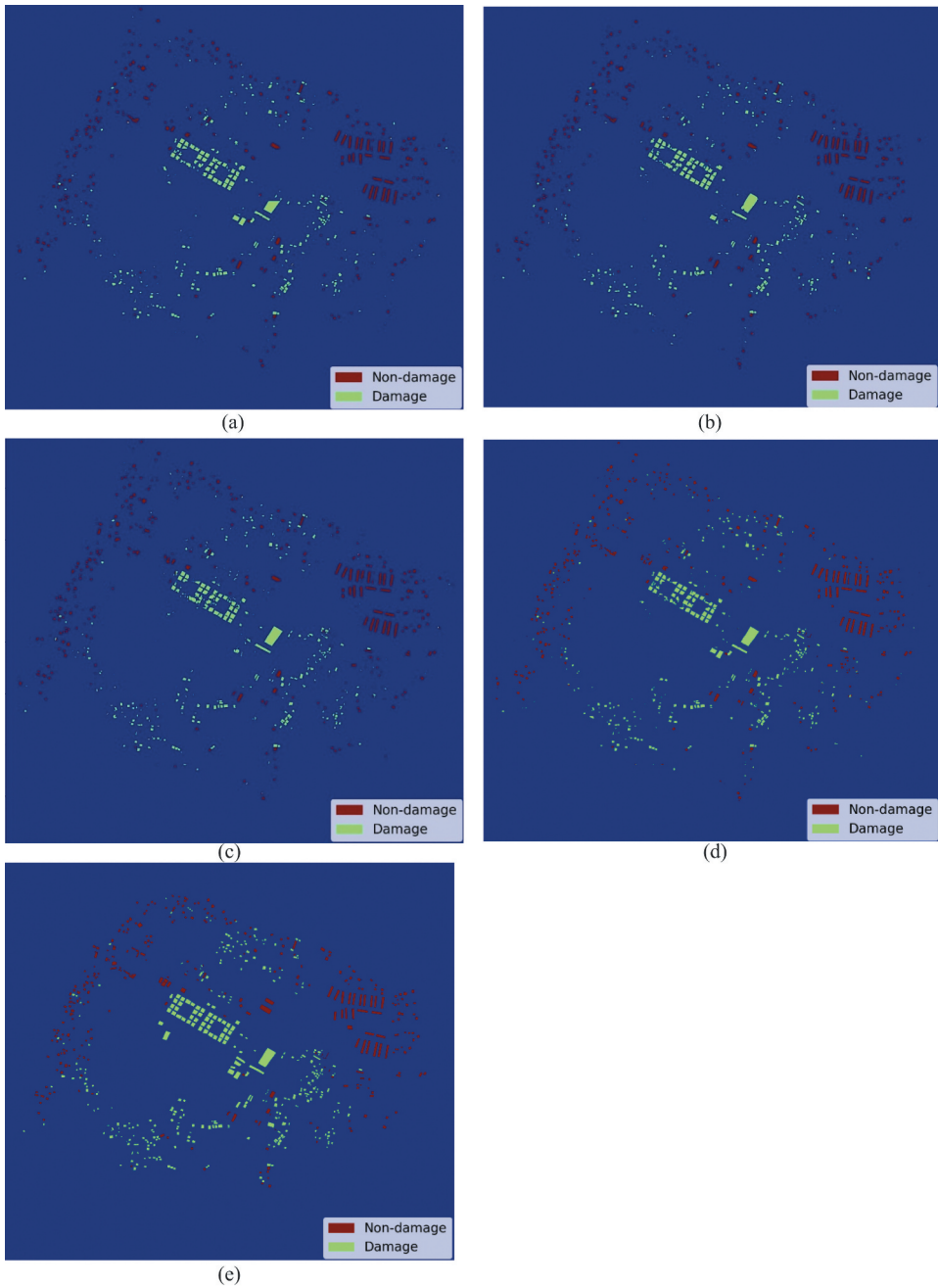


Figure 10. Comparison between the various networks for 2021 bata explosion: (a) EfficientformerV2L, (b) TinyViT, (c) FastViT, (d) proposed method, and (e) ground truth.

proposed method copes with it to a large extent and achieves favourable results for building damage detection. The EffformerV2 and FastViT (Figure 10a, c) miss some buildings and fail to detect the boundary of the building properly. In addition, the

TinyViT network (Figure 10b) fails to detect the buildings in some places and the results of this method are lower than those in the ground truth.

To better examine the performance of the proposed method, three areas from the 2021 Bata explosion dataset with different complexities are considered in Figure 11. The first area has a sparse urban region with industrial buildings that are completely damaged. The EfficformerV2L, TinyViT, and FastViT miss some damaged buildings, the FastViT misses the non-damaged building completely and these three networks also perform poorly in boundary reconstruction. On the other hand, the proposed network can cope with the complexities of this area and distinguish the damaged and non-damaged buildings accurately.

The second area has a sparse urban region with small buildings. The TinyViT and FastViT cannot reconstruct the boundary of small buildings properly. The EfficformerV2L has relatively good performance in small damaged and non-damaged building detection. The proposed method reconstructs the boundary of the small damaged and non-damaged buildings properly.

The third area has a very dense urban region which contains buildings with different shapes near each other. The proposed network can accurately distinguish the damaged buildings from the non-damaged ones. The TinyViT network failed to detect building boundaries accurately and also missed some damaged buildings. The FastViT network also misses some damaged buildings and cannot reconstruct the boundary of the building. Although the EfficformerV2L network provides relatively good results, it does not detect the boundary of one damaged building. The proposed network accurately distinguishes the damaged buildings from the non-damaged ones and reconstructs the boundary of buildings accurately.

The third dataset exploited in this study is the 2010 haiti earthquake. The results of these dataset are examined by the proposed network and its performance compared to the other methods in Figure 12. The proposed network performs properly in this dataset and correctly distinguishes the damaged and non-damaged buildings. The FastViT (Figure 12c) misses some damaged and non-damaged buildings. Also, the TinyViT and EfficformerV2L (Figure 12a, b) fail to detect the buildings in some places and the results of this method are lower than those in the GT.

To analyse the performance of the proposed method, three areas from the 2011 haiti earthquake dataset with different complexities are considered in Figure 13. The first area has a very dense urban region and buildings with different boundaries that are difficult to distinguish. The EfficformerV2L network wrongly detects a building as the damaged one though it is not so. Additionally, it has poor performance in boundary reconstruction. The TinyViT has errors in the detection of damaged buildings and cannot reconstruct the boundaries. The FastViT has relatively better performance than the other networks but it considers two damaged buildings as one. The proposed method can handle the density of the buildings and the complexity of their boundaries in this area.

The second area has sparse urban regions. The FastViT misses some damaged and non-damaged buildings and has the worst performance. The EfficformerV2L has noise in some places. The TinyViT has better performance than the EfficformerV2L and FastViT but consider two damaged buildings that are near each other as one. The proposed method has the best performance and it can distinguish the damaged and non-damaged building accurately.

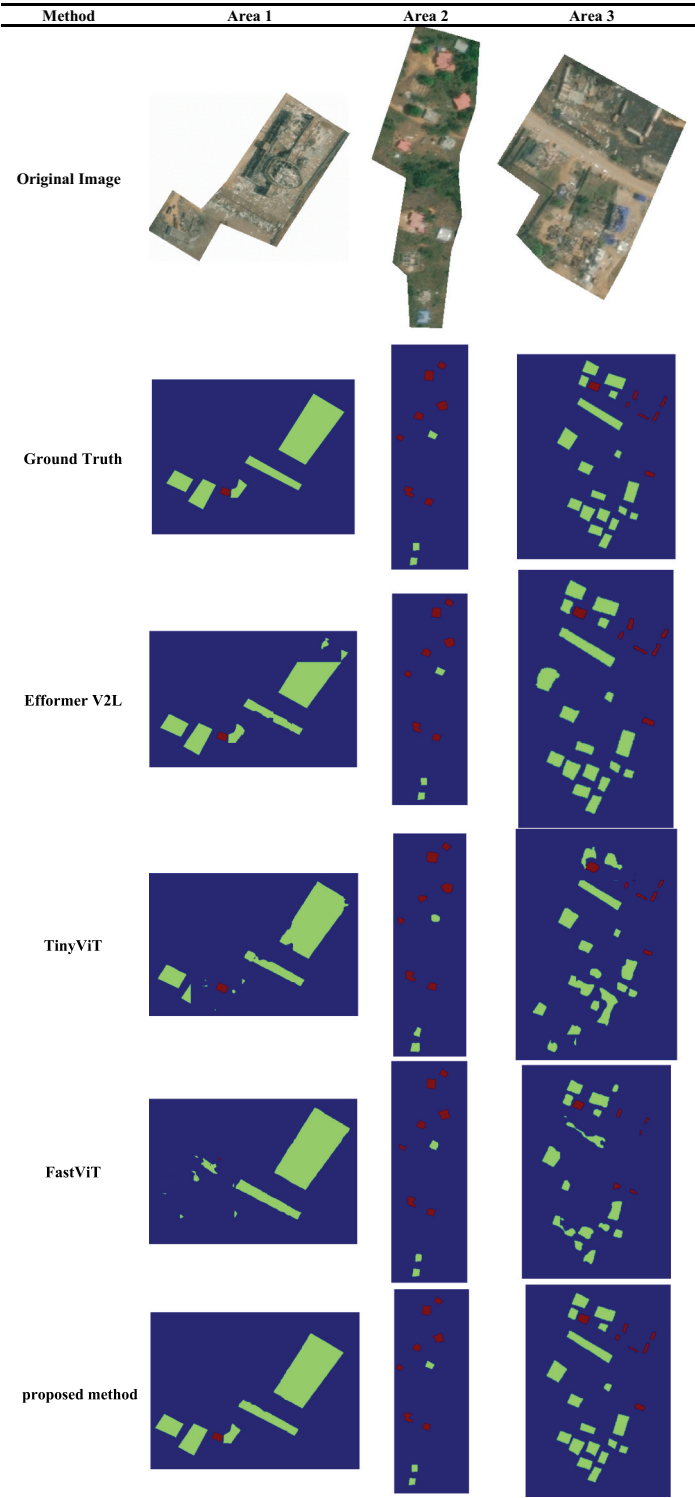


Figure 11. Comparison of the results of UNet-GCViT by different networks for bata explosion.

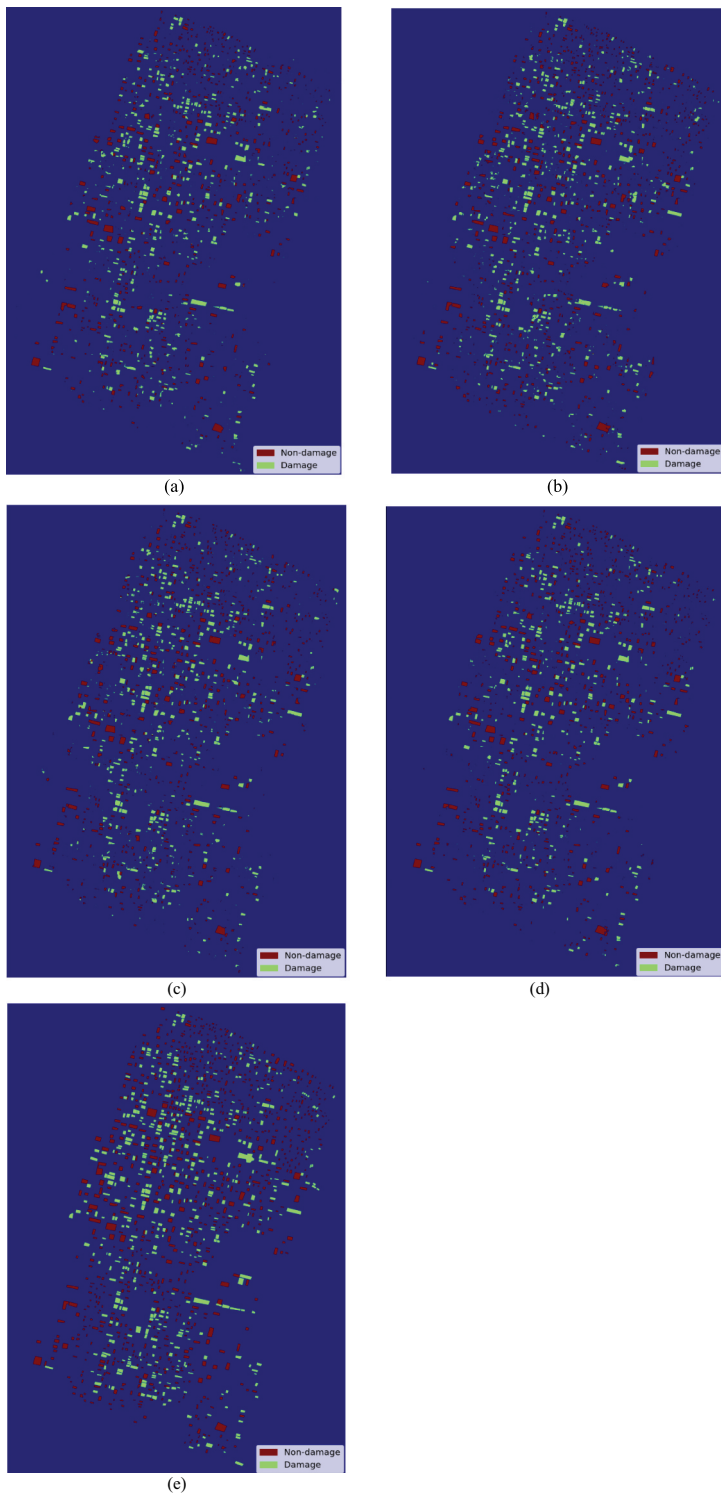


Figure 12. Comparison between the various networks for the 2010 haiti earthquake: (a) EfficientformerV2L, (b) TinyViT, (c) FastViT, (d) proposed method, and (e) ground truth.

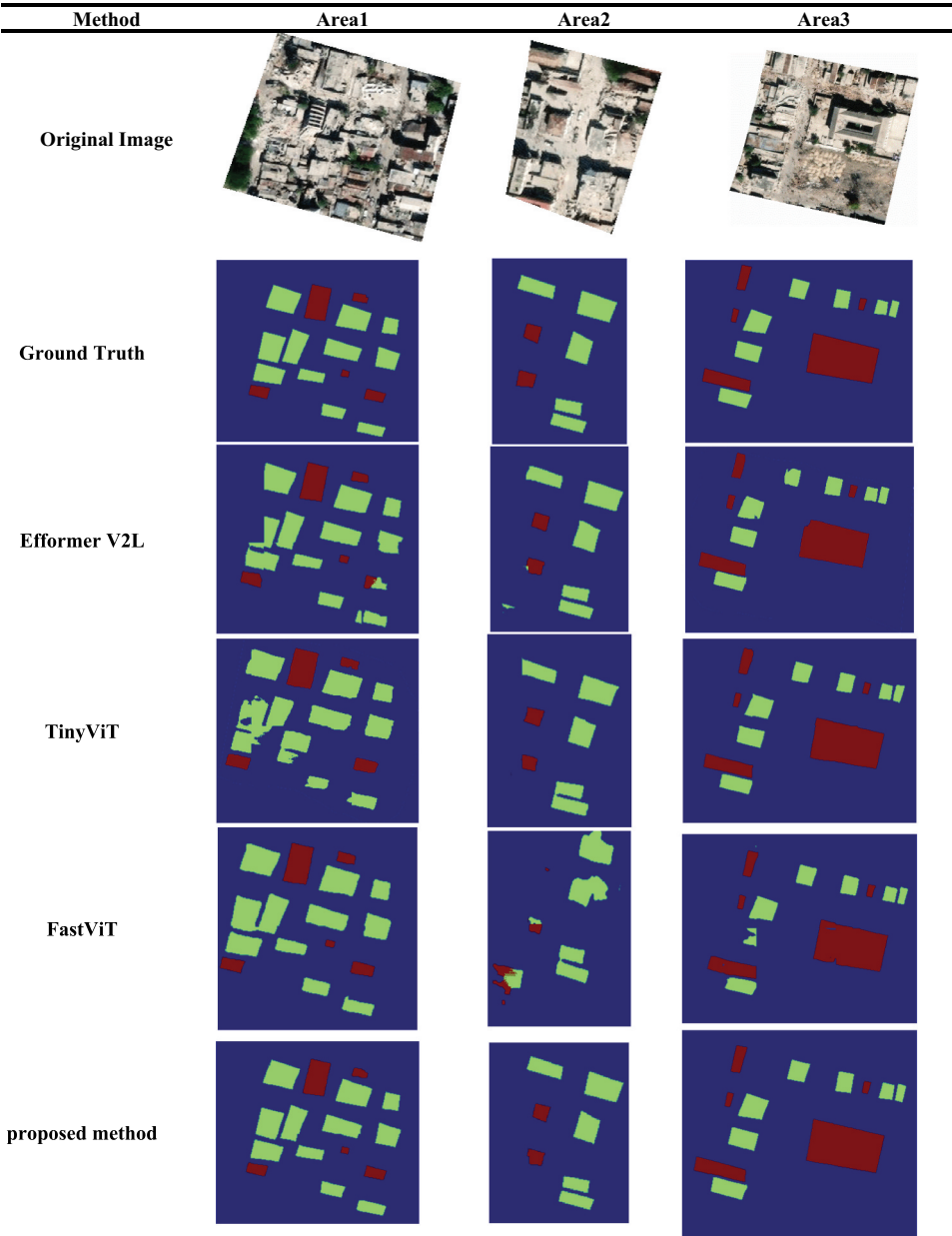


Figure 13. Comparison of the results of UNet-GCViT by different networks for the Haiti earthquake.

The third area has a relatively dense urban region and buildings with different shapes and boundaries. The EfformerV2L cannot reconstruct the boundaries properly. The FastViT misses a damaged building and has some noise. The TinyViT has a better performance than EfformerV2L and FastViT, but the proposed method is the best one, it accurately distinguishes the damaged buildings from non-damaged ones and reconstructs the boundaries properly.

Generally, the proposed method can detect damage and non-damage buildings of different sizes and boundaries. The proposed method can cope with the complexity of different urban areas with different environments; for instance, in some urban areas within three datasets, there are dense buildings with different shapes and the proposed method can effectively detect the damaged and non-damaged buildings in such regions. In other areas, where buildings are spars and have complex boundaries, the proposed method performs well in detecting damaged and non-damaged buildings and accurately reconstructing the boundaries.

6. Conclusion

A novel method is presented in this study based on the UNet architecture using the GCViT blocks to generate a map of the damaged buildings after natural and man-made disasters. The performance of this network is then compared with other new DL networks based on transformers. In this study, the GCViT architecture has been pre-trained by the ImageNet dataset; in other words, the semi-supervised transfer learning technique is employed. Three datasets are used in this study, the first one is related to the WorldView-II satellite images from the city of Nurdagi located in Turkey, which suffered much damage due to an earthquake (2023), made publicly available by Maxar Technologies Inc. The second one is the WorldView-II satellite images of the Bata explosion and the third one is the WorldView-II satellite images of the Haiti earthquake. In this study, only post-event images are employed to generate a map of the damaged buildings, as one of the advantages compared to methods that required both pre- and post-event data. The proposed method also achieved OA and KC equal to 96.71%, 98.71%, 98.75, and 0.85, 0.81, and 0.72, respectively, for the first, second, and third datasets, indicating its proper performance. Moreover, the performance of the given method in terms of training speed is compared with other networks such as FastViT, revealing that the proposed method has an acceptable speed for post-disaster building damage map generation. Which is appropriate for post-event crisis management planning at the earliest possible time.

Regarding the proper performance of this method for building damage map generation, it is suitable to apply for mapping buildings damaged by other natural disasters, such as floods, storms, and fires. Furthermore, it is suggested to use this network to generate a 3D map of damaged buildings after a natural and man-made disaster using UAV or LiDAR data.

Acknowledgment

Many thanks to Maxar Technologies Inc. for providing satellite images of Turkey's earthquake.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Masoomeh Gomroki  <http://orcid.org/0009-0009-7858-1830>

Mahdi Hasanlou  <http://orcid.org/0000-0002-7254-4475>

Jocelyn Chanussot  <http://orcid.org/0000-0003-4817-2875>

Danfeng Hong  <http://orcid.org/0000-0002-3212-9584>

References

- Adriano, B., J. Xia, G. Baier, N. Yokoya, and S. Koshimura. 2019. "Multi-Source Data Fusion Based on Ensemble Learning for Rapid Building Damage Mapping During the 2018 Sulawesi Earthquake and Tsunami in Palu, Indonesia." *Remote Sensing* 11 (7): 886.
- Bolloorani, A. D., M. Darvishi, Q. Weng, and X. Liu. 2021. "Post-War Urban Damage Mapping Using InSAR: The Case of Mosul City in Iraq." *ISPRS International Journal of Geo-Information* 10 (3): 140.
- Chen, H., E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley. 2022. "Dual-Tasks Siamese Transformer Framework for Building Damage Assessment." In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium Kuala Lumpur, Malaysia*, 1600–1603. IEEE.
- Chen, J., S. Kao, H. He, W. Zhuo, S. Wen, C.-H. Lee, and S.-H. G. Chan. 2023. "Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Vancouver, Canada*, 12021–12031.
- Chen, J., H. Liu, J. Hou, M. Yang, and M. Deng. 2018. "Improving Building Change Detection in VHR Remote Sensing Imagery by Combining Coarse Location and Co-Segmentation." *ISPRS International Journal of Geo-Information* 7 (6): 213.
- Cotrufo, S., C. Sandu, F. Giulio Tonolo, and P. Boccoardo. 2018. "Building Damage Assessment Scale Tailored to Remote Sensing Vertical Imagery." *European Journal of Remote Sensing* 51 (1): 991–1005. <https://doi.org/10.1080/22797254.2018.1527662>.
- D'Addabbo, A., G. Pasquariello, and A. Amodio. 2021. "Urban Change Detection from Vhr Images via Deep-Features Exploitation." In *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021*, 487–500, Volume 2, London. Springer.
- Dell'acqua, Fabio, and Paolo Gamba. 2012. "Remote Sensing and Earthquake Damage Assessment: Experiences, Limits, and Perspectives." In *Proceedings of the IEEE*, 2876–2890.
- Eguchi, Ronald T., Charles K. Huyck, Shubharoop Ghosh, Beverley J. Adams, and Anneley McMillan Showalter, Pamela S., Lu, Yongmei. 2010. "Utilizing New Technologies in Managing Hazards and Disasters." *Geospatial Techniques in Urban Hazard and Disaster Analysis*: 295–323 doi:10.1007/978-90-481-2238-7_15 . Dordrecht: Springer Netherlands.
- ElGharbawi, T., and F. Zarzoura. 2021. "Damage Detection Using SAR Coherence Statistical Analysis, Application to Beirut, Lebanon." *Isprs Journal of Photogrammetry & Remote Sensing* 173:1–9. <https://doi.org/10.1016/j.isprsjprs.2021.01.001>.
- Gomroki, M., M. Hasanlou, and J. Chanussot. 2023. "Automatic 3D Multiple Building Change Detection Model Based on Encoder-Decoder Network Using Highly Unbalanced Remote Sensing Datasets. 16: 10311–10325. doi:10.1109/JSTARS.2023.3328561." In.
- Gomroki, M., M. Hasanlou, and P. Reinartz. 2023. "STCD-EffV2T Unet: Semi Transfer Learning EfficientNetv2 T-Unet Network for Urban/Land Cover Change Detection Using Sentinel-2 Satellite Images." *Remote Sensing* 15 (5): 1232.
- Gupta, R., and M. Shah. 2021. "Rescuenet: Joint Building Segmentation and Damage Assessment from Satellite Imagery." 2020 25th International Conference on Pattern Recognition (ICPR) Milan, Italy, 4405–4411. IEEE.
- Haciefendioğlu, K., H. B. Başağa, and G. Demir. 2021. "Automatic Detection of Earthquake-Induced Ground Failure Effects Through Faster R-CNN Deep Learning-Based Object Detection Using Satellite Images." *Natural Hazards* 105 (1): 383–403. <https://doi.org/10.1007/s11069-020-04315-y>.
- Hatamizadeh, A., H. Yin, G. Heinrich, J. Kautz, and P. Molchanov. 2023. "Global Context Vision Transformers." *International Conference on Machine Learning. PMLR Proceedings of the 40 th International Conference on Machine Learning, Honolulu, Hawaii, USA., 12633–12646*.

- Hendrycks, D., and K. Gimpel. 2016. "Gaussian error linear units (gelus)." *arXiv Preprint arXiv: 160608415*.
- Hong, Z., H. Zhang, J. Gao, X. Tong, R. Zhou, H. Pan, Y. Zhang, Y. Han, J. Wang, and S. Yang. 2023. "Building Damage Assessment Using Improved Combined Adjustment and Automatic Building Corners Extraction Method." *International Journal of Remote Sensing* 44 (23): 7464–7491. <https://doi.org/10.1080/01431161.2023.2285740>.
- Hu, J., L. Shen, and G. Sun. 2018. "Squeeze-And-Excitation Networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Salt Lake City, UT, USA, 7132–7141.
- Janalipour, M., and A. Mohammadzadeh. 2019. "A Novel and Automatic Framework for Producing Building Damage Map Using Post-Event LiDAR Data." *International Journal of Disaster Risk Reduction* 39:101238. <https://doi.org/10.1016/j.ijdrr.2019.101238>.
- Ji, M., L. Liu, R. Du, and M. F. Buchroithner. 2019. "A Comparative Study of Texture and Convolutional Neural Network Features for Detecting Collapsed Buildings After Earthquakes Using Pre-And Post-Event Satellite Imagery." *Remote Sensing* 11 (10): 1202.
- Ji, M., L. Liu, R. F. Zhang, and M. Buchroithner. 2020. "Discrimination of Earthquake-Induced Building Destruction from Space Using a Pretrained CNN Model." *Applied Sciences* 10 (2): 602.
- Khankeshizadeh, E., A. Mohammadzadeh, H. Arefi, A. Mohsenifar, S. Pirasteh, E. Fan, H. Li, and J. Li. 2024. "A Novel Weighted Ensemble Transferred U-Net Based Model (WETUM) for Post-Earthquake Building Damage Assessment from UAV Data: A Comparison of Deep Learning-And Machine Learning-Based Approaches. 16: 1–17." In.
- Li, M., L. Cheng, J. Gong, Y. Liu, Z. Chen, F. Li, G. Chen, D. Chen, and X. Song. 2008. "Post-Earthquake Assessment of Building Damage Degree Using LiDAR Data and Imagery." *Science in China Series E: Technological Sciences* 51 (S2): 133–143. <https://doi.org/10.1007/s11431-008-6014-1>.
- Li, Y., W. Hu, H. Dong, and X. Zhang. 2019. "Building Damage Detection from Post-Event Aerial Imagery Using Single Shot Multibox Detector." *Applied Sciences* 9 (6): 1128.
- Li, Y., G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren. 2022. "Efficientformer: Vision Transformers at Mobilenet Speed." *Advances in Neural Information Processing Systems* 35:12934–12949.
- Lian, X., W. Yuan, Z. Guo, Z. Cai, X. Song, and R. Shibasaki. 2020. "END-TO-END BUILDING CHANGE DETECTION MODEL IN AERIAL IMAGERY AND DIGITAL SURFACE MODEL BASED ON NEURAL NETWORKS." In The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B2-2020:1239–1246. Nice, France:." *The International Archives of Photogrammetry, Remote Sensing & Spatial Information Sciences* 43:1239–1246 doi:10.5194/isprs-archives-XLIII-B2-2020-1239-2020 Copernicus GmbH.
- Liu, W., L. Chen, and Y. Chen. 2018. "Age Classification Using Convolutional Neural Networks with the Multi-Class Focal Loss." In IOP Conference Series: Materials Science and Engineering, 012043. IOP Publishing.
- Merlin, G. S., and G. W. Jiji. 2019. "Building Damage Detection of the 2004 Nagapattinam, India, Tsunami Using the Texture and Spectral Features from IKONOS Images." *The Journal of the Indian Society of Remote Sensing* 47 (1): 13–24. <https://doi.org/10.1007/s12524-018-0858-z>.
- Mohammadian, A., and F. Ghaderi. 2023. "SiamixFormer: A Fully-Transformer Siamese Network with Temporal Fusion for Accurate Building Detection and Change Detection in Bi-Temporal Remote Sensing Images." *International Journal of Remote Sensing* 44 (12): 3660–3678. <https://doi.org/10.1080/01431161.2023.2225228>.
- Naito, S., H. Tomozawa, Y. Mori, T. Nagata, N. Monma, H. Nakamura, H. Fujiwara, and G. Shoji. 2020. "Building-Damage Detection Method Based on Machine Learning Utilizing Aerial Photographs of the Kumamoto Earthquake." *Earthquake Spectra* 36 (3): 1166–1187.
- Nex, F., D. Duarte, F. G. Tonolo, and N. Kerle. 2019. "Structural Building Damage Detection with Deep Learning: Assessment of a State-Of-The-Art CNN in Operational Conditions." *Remote Sensing* 11 (23): 2765.
- Seydi, S. T., M. Hasanlou, J. Chanussot, and P. Ghamisi. 2023a. "Leveraging Involution and Convolution in an Explainable Building Damage Detection Framework." *European Journal of Remote Sensing* 56 (1). <https://doi.org/10.1080/22797254.2023.2252166>.

- Seydi, S. T., M. Hasanlou, J. Chanussot, and P. Ghamisi. 2023b. "BDD-Net+: A Building Damage Detection Framework Based on Modified Coat-Net." In.
- Shen, Y., S. Zhu, T. Yang, C. Chen, D. Pan, J. Chen, L. Xiao, and Q. Du. 2021. "Bdanet: Multiscale Convolutional Neural Network with Cross-Directional Attention for Building Damage Assessment from Satellite Images." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–14.
- Talreja, P., S. S. Durbha, R. C. Shinde, and A. V. Potnis. 2021. "Real-Time Embedded HPC Based Earthquake Damage Mapping Using 3D LiDAR Point Clouds." In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS Brussels, Belgium, 8241–8244. IEEE.
- Tan, M., and Q. Le. 2021. "Efficientnetv2: Smaller Models and Faster Training." In International Conference on Machine Learning. PMLR, 10096–10106.
- Ünlü, R., and R. Kiriş. 2022. "Detection of Damaged Buildings After an Earthquake with Convolutional Neural Networks in Conjunction with Image Segmentation." *The Visual Computer* 38 (2): 685–694. <https://doi.org/10.1007/s00371-020-02043-9>.
- Vasu, P. K. A., J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan. 2023. "FastViT: A Fast Hybrid Vision Transformer Using Structural Reparameterization. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 5762–5772. Paris, France. doi:10.1109/ICCV51070.2023.00532." *arXiv Preprint arXiv: 230314189*.
- Wang, C., Y. Zhang, T. Xie, L. Guo, S. Chen, J. Li, and F. Shi. 2022. "A Detection Method for Collapsed Buildings Combining Post-Earthquake High-Resolution Optical and Synthetic Aperture Radar Images." *Remote Sensing* 14 (5): 1100.
- Woo, S., S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. 2023. "Convnext V2: Co-Designing and Scaling Convnets with Masked Autoencoders." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Vancouver, BC, Canada, 16133–16142.
- Wu, C., F. Zhang, J. Xia, Y. Xu, G. Li, J. Xie, Z. Du, and R. Liu. 2021. "Building Damage Detection Using U-Net with Attention Mechanism from Pre-And Post-Disaster Remote Sensing Datasets." *Remote Sensing* 13 (5): 905.
- Wu, K., J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan. 2022. "Tinyvit: Fast Pretraining Distillation for Small Vision Transformers." In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI, 68–85. Springer.
- Yadav, R., A. Nascetti, and Y. Ban. 2022. "BUILDING CHANGE DETECTION USING MULTI-TEMPORAL AIRBORNE LIDAR DATA." In The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2022:1377–1383. Nice, France:." *arXiv Preprint arXiv: 220412535*.
- Yeom, J., M. Jung, and Y. Kim. 2017. "Detecting Damaged Building Parts in Earthquake-Damaged Areas Using Differential Seeded Region Growing." *International Journal of Remote Sensing* 38 (4): 985–1005. <https://doi.org/10.1080/01431161.2016.1274445>.
- Zheng, Z., Y. Zhong, J. Wang, A. Ma, and L. Zhang. 2021. "Building Damage Assessment for Rapid Disaster Response with a Deep Object-Based Semantic Change Detection Framework: From Natural Disasters to Man-Made Disasters." *Remote Sensing of Environment* 265:112636. <https://doi.org/10.1016/j.rse.2021.112636>.