# Evaluation of Deep Learning Models for Building Damage Mapping in Emergency Response Settings

Sesa Wiguna, Bruno Adriano, *Senior Member, IEEE*, Erick Mas, and Shunichi Koshimura

***Abstract*—Integrated with remote sensing technology, deep learning has been increasingly used for rapid damage assessment. Despite reportedly having high accuracy, the approach requires numerous samples to maintain its performance. However, in the emergency response phase, training samples are often unavailable. Since no ground truth data is available, deep learning models cannot be trained for this specific situation and, thus, have to be applied to unseen data. Previous research has implemented transfer learning techniques to solve data unavailability. However, many studies do not accurately reflect the rapid damage mapping in real-world scenarios. This study illustrates the use of Earth observation and deep learning technologies in predicting damage in realistic emergency response settings. To this aim, we conducted extensive experiments using historical data to find the best model by examining multiple neural network models and loss functions. Then, we evaluated the performance of the best model for predicting building damage due to two different disasters, the 2011 Tohoku Tsunami and the 2023 Türkiye–Syria Earthquake, which were independent of the training samples. We found that a transformer-based model with a combined cross-entropy loss (CEL) and focal loss generates the highest scoring values. The testing on both unseen sites illustrates that the model can perform well in no-damage and destroyed classes. However, the scores dropped in the middle class. We also compared our transformer-based approach with other state-of-the-art models, specifically the xView-2 winning solution. The results show that the transformer-based models have stable generalization toward multiclass classification and multiresolution imagery.**

***Index Terms*—Building damage detection, deep learning (DL), disaster resilience, Earth observation, emergency response.**

## I. Introduction

In the aftermath of a disaster, damaged conditions, including affected buildings, are essential to detect for humanitarian activities. This information can help aid decision-makers in assessing the needs of affected populations, prioritizing response efforts, and allocating resources effectively. However, gaining the damage information in an emergency condition is challenging due to safety and accessibility reasons. Integrated with remotely sensed data, deep learning (DL) technology has been an emerging technique to serve in automatic building damage recognition.

Despite being superior in accuracy compared to other methods [1], [2], DL requires a large amount of data to maintain its performance. However, generating training samples is a laborious and time-consuming task that is not suitable for an emergency response setting [3] since the damage information is usually required for quick decision-making and action. The attention then turns into transfer learning techniques, that is, utilizing knowledge learned from a large dataset to test in a different but related task.

Previous researchers have studied transfer learning approaches for building damage detection. However, many studies do not accurately represent rapid damage mapping in real-world scenarios or realistic disaster emergency mapping. For example, the authors in [4] and [5] evaluated model generalization by utilizing the trained model to predict building damages in new instances unseen during the training. Gupta and Shah [4] split samples by the type of disaster, where some groups of a disaster event were used for training and other groups for testing. Benson and Ecker [5] used a similar split criteria as [4]. However, rather than using the whole testing set as in [4] and [5], split the testing set into three folds consisting of smaller disaster groups to reduce the computational cost. Although this reflects the transfer learning scenario and grouping can help to avoid an insufficient number of samples for testing, we argue that their split is less representative of realistic disaster emergency mapping. First, emergency mapping is conducted for an individual event with damage mechanisms that may differ from other hazards or events. Thus, grouping samples from various types of disasters in the testing stage might introduce biases to the model. Therefore, the testing should be done for an individual disaster type. Second, using multiple disaster events as the testing dataset is rather wasteful since the data can actually be used to enrich training processes. In a real scenario, one would maximize all available data for the training and then predict one new disaster [4].

Other studies applied transfer learning in an individual event rather than in a group of disasters. However, many focus on a single type of hazard. In fact, each disaster is unique. In other words, each disaster behaves differently so that a model may generalize well in one disaster but may not work equally in other events. It calls for more understanding of model generalization ability toward multiple disaster events. In [6], for

Manuscript received 28 September 2023; revised 27 December 2023; accepted 9 February 2024. Date of publication 20 February 2024; date of current version 6 March 2024. This work was supported in part by the JSPS KAKENHI Grants-in-Aid for Scientific Research under Grant 21H05001, Grant 22K21372, and Grant 22H01741, in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), and in part by the Co-creation Center for Disaster Resilience, Tohoku University. *(Corresponding author: Shunichi Koshimura.)*

Sesa Wiguna is with the Department of Civil and Environmental Engineering, Tohoku University, Sendai 980-8579, Japan (e-mail: wiguna.sesa.p1@dc.tohoku.ac.jp).

Bruno Adriano, Erick Mas, and Shunichi Koshimura are with the Disaster Geo-informatics Lab, International Research Institute of Disaster Science, Tohoku University, Sendai 980-8572, Japan (e-mail: adriano@irides.tohoku.ac.jp; mas@irides.tohoku.ac.jp; koshimura@irides.tohoku.ac.jp).

Digital Object Identifier 10.1109/JSTARS.2024.3367853

example, the model generalization was tested in both location and data variation; however, the focus was only on an earthquake. Zahs et al. [7] introduced a method involving machine learning and point-cloud data, then evaluated the model's transferability on earthquake-induced building damage. Similarly, Adriano et al. [8] assessed the cross-domain generalization by generating synthetic postevent SAR imagery focusing on tsunami damage.

The primary contribution of this study lies within the performance evaluation of DL models for building damage mapping in realistic emergency settings. We replicate rapid building damage assessment in a condition where samples are unavailable, which is the case in almost all disaster emergency situations. That said, we trained multiple DL models on a global disaster dataset and used the best-performing model to estimate damage in two unseen disaster events. Specifically, the main steps to our work's contribution are threefold.

1) We formulate the building damage recognition in a realistic disaster emergency scenario where the ground truth data is unavailable. Since no ground truth data is available, DL models cannot be trained for this specific situation and, thus, have to be applied to unseen data. We demonstrate the feasibility of DL models trained on a global dataset in predicting damage in new locations, namely the 2011 Tohoku Tsunami and the 2023 Türkiye–Syria Earthquake.
2) We conduct extensive experiments to find the best-performing models, including the state-of-the-art convolutional neural network (CNN) and modern transformer models. We also experiment with different input scenarios considering the optical image availability in the aftermath of disasters.
3) We compare the best-performing model with state-of-the-art DL approaches for building damage recognition. We evaluate the metrics of semantic segmentation-based models at the pixel level and building level to harmonize model performance of semantic segmentation and image classification tasks, two commonly used methods of building damage assessment.

### A. Related Works

Damage detection analyzes using remotely sensed data have been widely studied. A comprehensive review of this field has been made by several authors, for example, [9] for tsunami, [10], [11] for earthquake, and [12] for synthetic aperture radar-based building damage assessment. The use of remote sensing (RS) for damage detection can be grouped into regional and local levels. At the regional level, the estimation is usually made to estimate the affected area [13], [14]. For example, Yusuf et al. [15] used Landsat 7 images to detect affected areas due to the 2001 Gujarat Earthquake. While the regional-level approach provides an insight into the overview of the overall affected area, it is unable to give the details of the object being affected. The analysis is then targeted to inspect the damage at the local level, e.g., building. At the building level, the common approaches used include the following: 1) image classification, where a particular damage class is assigned to an image, e.g., [6], [16], [17] or where a damage category is assigned to each building patch and 2) semantic segmentation, where the categorization takes place at a pixel level, e.g., [18], [19]. Adriano et al. [20] proposed a semantic segmentation-based building damage mapping framework by considering data availability following a disaster. The proposed framework involves a multitemporal (utilizing pre- and postdisaster images) and multimodal using optical and radar imagery Earth observation dataset from large-scale earthquakes and tsunamis around the globe.

The analysis at the object level has been enhanced by the advancement of DL and data availability. Since winning the ImageNet competition in 2012, CNN-based models have become more popular in the computer vision field, including RS [21]. Nowadays, transformer models, which were initially developed in natural language processing, have been adopted for various computer vision tasks, e.g., [22]. Previous studies, such as the authors in [23] and [24] showed that the transformer-based model outweighs the CNN models. Chen et al. [23] designed a transformer-based damage assessment architecture. The model consists of a Siamese transformer encoder to extract features from multitemporal images. The extracted features are then fused by a multitemporal fusion module before feeding them into a lightweight dual-tasks decoder that aggregates multilevel features for final prediction. The experiments show that the designed model has surpassed CNN-based state-of-the-art models.

Moreover, a number of dataset benchmarks, such as OpenEarthMap [25], INRIA [26], WHU-OHS [27], and xBD [28] have been introduced to advance the research in the field. For example, Deng and Wang [29] used the xBD dataset to improve the U-Net model by designing a two-stage building damage assessment network in a semantic segmentation task. In the building segmentation stage, an extra skip connection and asymmetric convolution block were used to enhance the network's ability to segment buildings on different scales. Meanwhile, shuffle attention is utilized to improve the model by directing the network's attention to the correlation between buildings before and after the disaster. Similarly, Shen et al. [30] took advantage of xBD data to apply their proposed cross-directional attention module that explores the correlations between pre- and postdisaster images. Zheng et al. [31] proposed ChangeOS, a model that features an object generation module and an object classification module. The model was trained on the xBD dataset to develop an end-to-end building damage detection with high performance in both speed and accuracy. Yu et al. [3] utilized xBD to verify their SegDetector to detect small-scale targets and overlapping targets in target detection tasks. Xie et al. [32] developed two subnetworks named building subclass segmentation network for building subclass segmentation by combining binary building segmentation and multiclass building segmentation by utilizing the dataset.

On many occasions, DL models are designed for particular locations. This makes the ability of models to generalize to unseen data remains unknown [33] as the models may be dependent on the training dataset [34]. In fact, models would be more practically useful for real-world applications if the model can generalize to new unseen datasets [5], [35]. Furthermore, Benson and Ecker [5] insist that a model should be ranked through its ability to generalize to unseen (out of domain or OOD) data rather than in-domain (IND) datasets. In this case, IND refers to a proportion of samples that are intentionally split from all samples for testing purposes.

Previous studies, such as [5], [6], [8], [16], [35], [36], [37], have attempted to evaluate the generalization ability to serve emergency response efforts. Yang et al. [6] evaluated the generalization ability of several well-known CNN-based networks namely VGG-16, InceptionV3, ResNet50, and DenseNet121. In their study, the generalization of models was assessed in terms of geographic location (testing on satellite images of an unseen area) and data generalization (transferring from satellite to aerial photo), mainly in binary classification (damage or no-damage). Bouchard et al. [37] proposed a two-step model design comprised of BuildingNet for building localizer followed by DamageNet for damage classifier. Each model was trained before disasters to make them ready for inference, thus minimizing the postincident execution time.

The existing literature mostly focuses on binary class. For example, Nex et al. [36] evaluated the geographical transferability in three different datasets: satellite, airborne, and UAV in two building damage classes, namely damage and intact. Similarly, the authors in [33], [37], and [38] work in binary mapping. Bouchard et al. [37] combined no-damage and minor-damage into one class and major- and destroyed class into another. They argue that minor-damage levels do not require immediate emergency attention from humanitarian organizations. As for more general purposes [32], for example, for assessing monetary damages, damaged buildings should be classified into detailed classes [39].

Limited studies assess the OOD generalization in larger disaster types and damage classes. Benson and Ecker [5] developed four-class damage models and tested them in three test folds containing disasters driven by wind, fire, and water. Valentijn et al. [16] tested their models on performance on flood and tornado, also comparing the performance of binary and multidamage classes (no-damage, minor-damage, major-damage, destroyed). Both studies use the xBD dataset solely in their experiments. The data were then split in such a way as to meet the OOD evaluation scheme. For example, Benson and Ecker [5] hold several disaster events from xBD intentionally to be used only for OOD-generalization testing. Similarly, Valentijn et al. [16] excluded Nepal Flooding and Joplin Tornado of xBD datasets from the training and used those events only for OOD testing. Our study expands the generalization studies by utilizing data beyond xBD for testing. We believe that our setting reflects a more realistic disaster emergency condition where the dataset used for testing is generated independently from the training set. More specifically, both training and testing sets may vary in sensors to acquire the images and methods to determine the damage levels.

In summary, this study expands the literature by evaluating the usefulness of RS and DL technologies in realistic emergency settings. We maintain the model training and testing procedures as in real-world emergency mapping. We maximize the historical data and find the best-performing DL models among the state-of-the-art to predict damage in new areas.

The rest of this article is organized as follows. Section II presents the methods in which data research workflow, DL architectures, and loss function are described. Sections III and IV provide the experimental results and the discussion, respectively. Finally, Section V concludes this article.

## II. Methods

This section describes the methods used in this study. It starts by describing the approach of the study, then will describe the details of the methods.

### A. *Research Settings*

The study illustrates rapid building damage assessment in realistic emergency disaster response. In this scenario, obtaining ground truth data is limited, making it impractical to train a DL model from scratch. Therefore, the study relies on historical data to estimate the building damage in targeted locations.

In general, building damage detection can be perceived as semantic segmentation or image classification tasks. This study sets the current task as an image classification problem. Specifically, the model takes building patches as inputs and predicts the damage state of every building patch in the following four categories: no-damage, minor-damage, major-damage, or destroyed. Since a particular class is assigned to each image (patch) rather than to a pixel, the method is inexpensive in computation and, hence, suitable for disaster rapid assessment purposes. Moreover, building footprints that are usually required to create samples, e.g., to crop the satellite imagery to the extent of building polygon, are available publicly from a number of sources, including openstreetmap (OSM)[1] and Global ML Building footprints from Microsoft for large areas of the world.[2] In [40], damaged buildings were identified through a semantic segmentation approach; then, they summarized the model's output over the Microsoft Building footprint dataset. Our approach simplifies the process by directly categorizing the damage at the building level.

As depicted in Fig. 1, the study starts with preparing data inputs. Details about the data preprocessing are described in Section II-B. Second, the data inputs are used for model training. Here, we examine multiple DL architectures and loss functions to find the best-performing model. Details of which are described in Sections II-C and II-D. For all DL models, we also designed Siamese networks comprised of two encoders to utilize multitemporal images. This model selection process utilized the historical disaster events gathered in the xBD datasets. Finally, to illustrate the usefulness of the historical dataset in assisting the real emergency scenario, the best model is used to predict disasters that are not included in the training processes. To this aim, the 2011 Tohoku Tsunami and the 2023 Türkiye–Syria Earthquake datasets are used as real case testing scenarios.

Damage mapping studies generally involve images before- and after disasters. Although it is possible to use postdisaster images only, this approach may lose accuracy due to insufficient information [41], [42]. In contrast, multitemporal inputs make it possible to compare the degree of similarity between the two

[1] [Online]. Available: https://www.openstreetmap.org/

[2] [Online]. Available: https://github.com/microsoft/GlobalMLBuildingFootprints

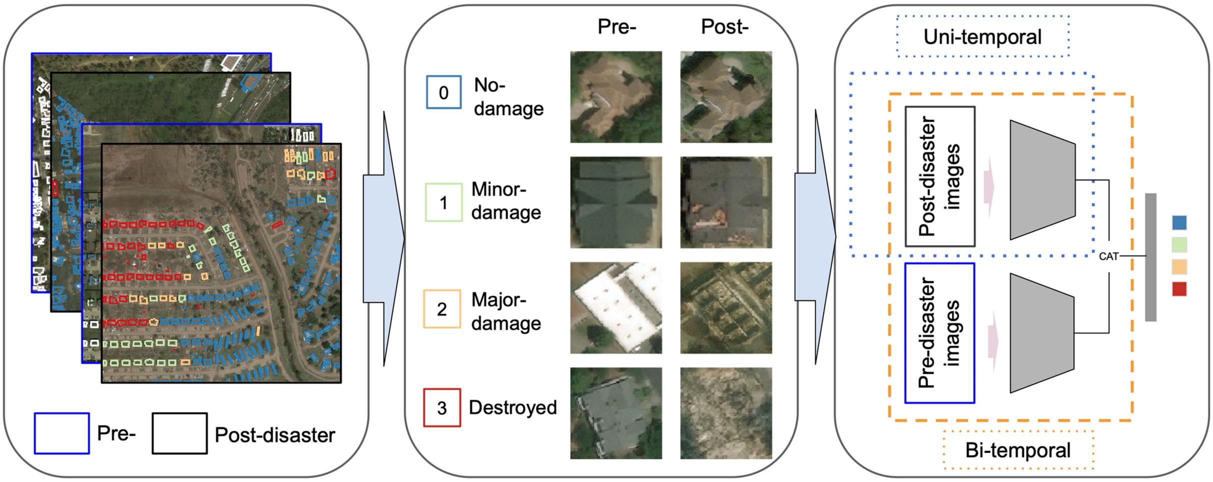


Fig. 1. Workflow of the research. Left: Data source containing Images and damage labels. Middle: Building patches. Right: Model training includes the best DL model selection. The best-performing model is then used for real-case testing.

samples [43]. As shown by [44] and [35], utilizing multitemporal images outperforms postonly inputs in building damage mapping. However, multitemporal images may not always be available. This study evaluates two input scenarios as follows.

1) *Unitemporal:* This scheme feeds only postdisaster images to the models. The setting fits conditions where only postdisaster images, short-range imagery, e.g., aerial photos, are available.
2) *Bitemporal:* The setting assumes that both pre- and postdisaster images are available, e.g., provided through disaster charters such as Maxar Open Data Program.[3]

### B. Data Preprocessing

Data preprocessing includes steps to prepare inputs for model training. The study utilizes the xBD Dataset [28] as the source of model training. The dataset was originally established for the xView-2 Challenge and is accessible at.[4] The database comprises building damage labels and satellite images. The damage information is stored in building footprint polygons containing four classes of damage, namely no-damage, minor-damage, major-damage, and destroyed. The images are very high-resolution satellite images with a ground sampling distance (GSD) of 0.5 × 0.5 m acquired before and after 19 major disasters from nine disaster types, including earthquake, fire, bushfire, wildfire, floods, hurricane, tornado, tsunami, and volcanic eruption. Details of the making of the xBD dataset and the damage description for each class can be referred to [45]. The samples are collected from around the globe. However, the samples are biased toward the United States, where the country holds the majority of the sites. The location distribution of samples used in the study is illustrated in Fig. 2.

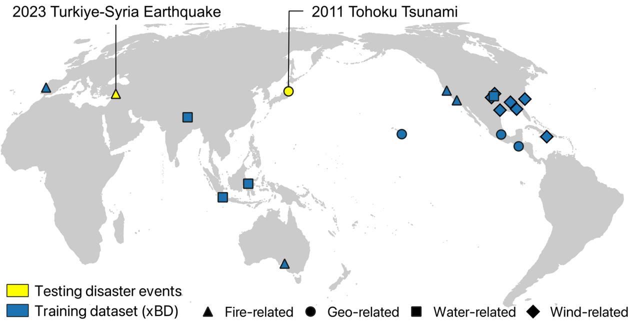


Fig. 2. Distribution of xBD and test locations. xBD samples are used for training processes while testing disaster events are independent of xBD. The disaster types are grouped based on triggering factors. Fire-related includes bushfires, fire, and wildfire. Geo-related covers volcanic eruptions and earthquakes. Water-related comprises of flooding and tsunamis. Wind-related consists of hurricanes and tornadoes.

Training samples were obtained by cropping the pre- and postdisaster images with the bounding box of each building footprint. A buffer of 2 pixels (1 m) was added to each bounding box to capture the situation surrounding the building, i.e., existence of debris [6], [16]. Having cropped the images, we select buildings with sizes of a minimum of 14 m at their width and length by considering a 28 × 28 to discard small buildings. This step results in 181 254 pair patches from pre- and postdisaster images.

In the xView-2 Challange, the original xBD dataset was split into the following four folders: Train, Tier3, Test, and Hold-out. This study also takes samples from Train, Tier3, and Test folders as train sets and uses the hold-out folder as the test set. The train set was split at a ratio of 0.8 and 0.2 for training and validation, respectively. The number of samples for training, validation, and test set per damage class is illustrated in Fig. 3.

### C. DL Models Architecture

The study compares the performance of multiple DL models, including ResNet and ResNeXt of CNN and swin transformer of the transformer models. They are among the state-of-the-art from earlier periods to the latest DL models in numerous applications. They vary in feature and model complexity.

[3] [Online]. Available: https://www.maxar.com/open-data/
[4] [Online]. Available: https://xview2.org/dataset

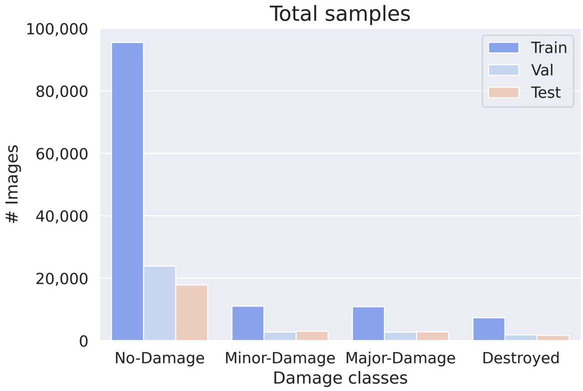


Fig. 3. Number of samples generated from xBD dataset. Train, val, and test refer to training, validation, and test set, respectively.

*1) ResNet:* Generally, the model's robustness could be enhanced by increasing the depth of networks. However, it is found that adding more layers may saturate the accuracy and instead reduce the performance due to the vanishing gradient problem. Residual Network [46] then solved the issues by introducing residual blocks with skip connections that add intermediate input to the output of a series of convolution blocks. Through this mechanism, ResNet ensures that the gradient signal does not become too small, which helps prevent the vanishing gradient problem and, thus, allows the addition of more layers. In the authors' experiments, ResNet outperforms previous state-of-the-art architectures on the ImageNet dataset, such as VGG and GoogLeNet. Since the establishment of ResNet, it soon became the state-of-the-art CNN model in multiple DL tasks, including image classification, object detection, and semantic segmentation. Many variants of ResNet architectures utilize the same concepts yet vary in the number of layers. In this study, ResNet34 was used. The network has $21.5 \times 10^6$ paremeters.

*2) ResNeXt:* ResNeXt model is proposed by [47] on top of ResNet. The authors introduced a new dimension called cardinality, which refers to the number of independent paths within a residual block. Traditional residual blocks in ResNet have a single path that connects the input to the output through a series of convolutional layers, while ResNeXt introduces multiple paths, each of which performs a subset of the total convolutional operations. Their experiments show that accuracy can be gained more efficiently by increasing the cardinality than going deeper or wider. The model architecture follows the split-transform-merge paradigm, where input is split into a number of paths, transformed independently, and merged by aggregating the outputs of different paths. Each path shares the same topology, requiring fewer parameters while adding more layers to this architecture. Through cardinality and aggregated transformations, the model can learn richer and more diverse feature representations. Wu et al. [19] compare ResNet, squeeze-and-excitation networks, ResNeXt, and dual path net as the backbone for the U-net model. Their experiments show that the ResNeXt model with attention suppressed other studied models. This study utilized ResNeXt50, which has $25 \times 10^6$ parameters.

TABLE I
NEW ADDITIONAL LAYERS ADDED TO EACH NETWORK

| Layer names | Details (in, out) + Layers |
|---|---|
| FC1 | (Encoder output, 512) + Activation function, Dropout |
| FC2 | (512, 512) + Activation function, Dropout |
| FC3 | (512, no. classes) |

*3) Shifted Windows Transformer:* Shifted Windows (Swin) Transformer was introduced in 2021 by [48]. The model architecture builds hierarchical feature maps by starting from small-sized patches (local windows) and gradually merging neighboring patches in deeper transformer layers. Self-attention layers compute within each local window rather than within the entire image patches. This mechanism results in lower computation and allows the model to learn finer features, which is useful for other tasks such as semantic segmentation. Swin transformer reduced the number of patching through a path-merging approach, which concatenates features of neighboring patches and applies a linear layer. A shifted window approach is introduced to ensure the connections between local windows. It utilizes two partitioning configurations, where the first layer uses regular partitioning, and the second layer uses a windowing configuration shifted by half of the window size. This shifted window introduces connections between neighboring windows while keeping the local computation within each nonoverlapping window. Swin transformer has multiple backbones in regards to model's size and computational complexity. In ascending order of those parameters, the Swin variants are T ($29 \times 10^6$), S ($50 \times 10^6$), B ($88 \times 10^6$), and L. SwinT, for example, has 96 channels of hidden layers in the first stage. In comparison, Swin L has 192 total channels at the same stage. Xia et al. [25] compared multiple DL models, including CNN and transformer-based models in mapping land use, and they found that swin transformer-based models are among the best among the studied models. Ogawa et al. [49] designed a model to estimate the built year and structure of a building using omnidirectional street view images captured using an onboard camera. The classification model was trained using CNN-based and transformer-based networks. The results show that swin transformer model effectively improves prediction accuracy.

Our study evaluates two input scenarios: unitemporal and bitemporal. To meet this research setting, we modified the studied networks. Three additional linear layers were added, and the last layer of each network was modified to accommodate four classes of output. Each linear layer is followed by an activation function. In addition, dropout layers are added to the first two linear layers. The detail of each layer is summarized in Table I. In the unitemporal input, the models are fed only with postdisaster images.

For the bitemporal input scenario, we designed a Siamese structure. The Siamese networks have two identical encoders. Each encoder extracts features from pre- and postdisaster images individually. Features maps extracted from each image were concatenated and then fed into the additional linear layers for classification.

### D. Loss Functions

Loss functions are an important part of the training process in DL. They provide a measure of how well the model is performing and guide the optimization algorithm to improve its performance. Studies such as [50] show that loss function selection greatly influences the modeling results. This study investigates the two most used loss functions to gain the best performance. The studied loss functions include CE, focal, and CE-Focal.

*1) Cross-Entropy Loss:* CEL [51] measures the difference between two probability distributions (ground truth data and predicted distribution). CEL has become popular for many DL tasks, including image classification. This study uses categorical CE for multiclass classification tasks. The CEL is expressed as follows:

$$\mathcal{L}_{\mathrm{CE}(y,\hat{y})} = -\sum_{i=1}^{M} y_i \log \hat{y}_i \tag{1}$$

where $M$ is the number of classes; $y_i$ represents the true class label for class $i$, and $\hat{y}_i$ is the predicted probability of class $i$.

*2) Focal Loss:* Focal Loss is proposed by [50] to address imbalanced samples. In focal loss, the authors refined the standard CEL by adding a modulating factor to reduce the relative loss for well-classified samples and putting more focus on hard, misclassified samples. The focal loss is expressed as follows:

$$\mathcal{L}_{\mathrm{Focal}(\hat{y})} = -(1-\hat{y})^{\gamma}\log(\hat{y}) \tag{2}$$

where $\hat{y}$ is the predicted probability of the true class and $\gamma$ is the focusing parameter that modulates the contribution of well-classified examples to the loss. Their experiments on object detection show that introducing the modulating factor can solve the data imbalanced issues. The loss function is used in our study to combat the class imbalance of the xBD samples. This study uses $\gamma$ of 2, which shows the best performance in the authors' study.

*3) CE-Focal Loss:* Besides the aforementioned losses, the study uses CE-Focal loss, an aggregated loss of CEL and focal loss. The CE-Focal loss takes the sum of the two losses. Formally, the loss is expressed as follows:

$$\mathcal{L}_{\text{CE-Focal}} = \mathcal{L}_{\mathrm{CE}} + \mathcal{L}_{\mathrm{Focal}}. \tag{3}$$

### E. Evaluation Metrics

Model evaluation is an important part of DL studies. A number of metrics are widely used in the model evaluation, including accuracy, precision, recall, and $F_1$. Accuracy is the proportion of correctly classified samples out of the total samples. Precision measures the proportion of correctly classified samples over total samples predicted as positive by the model, while recall measures the proportion of correctly classified samples out of all actual positive samples in the dataset. In imbalanced datasets where the number of samples of each class is not the same, the accuracy metric can be misleading as the model may achieve high accuracy by predicting the majority class even though the minority class performs poorly. Meanwhile, $F_1$ considers both precision and recall, which makes it suitable for imbalanced datasets. As described in Fig. 3, the number of samples in each class is unequal. Therefore, we selected $F_1$ to evaluate the models' performance. The formula to obtain $F_1$ is mathematically expressed as follows:

TABLE II
COMPUTATION RESOURCES AND HYPERPARAMETER SETTINGS FOR THE EXPERIMENTS

| Parameters | Details |
|---|---|
| PC resource | NVIDIA Quadro RTX 6000 |
| Operating system | Ubuntu 22 |
| Program | Python 3.8 |
| DL framework | Pytorch Version 1.12 |
| Optimizer | AdamW |
| Initial LR | $1 \times 10^{-4}$ |
| Number of epochs | 30 |
| Batch size | 64 |

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where TP is true positive (rate of positive instances correctly classified), FP is false positive (rate of negative instances misclassified), and FN is false negative, indicating the rate of positive instances misclassified.

## III. Experiments

This section contains the training settings and the experiments' results. In the experiment results, we first describe the DL model selection processes involving multiple DL models and loss functions. Second, we report the performance of the selected model on real case scenarios. Finally, we show the model performance when the task is simplified from four-class to three-category classification.

### A. Training Settings

Details of the resources and hyperparameters used in the study are listed in Table II. We utilized PyTorch as the DL framework. We initialized the first layer weights using networks pretrained on ImageNet to improve the training speed and achieve better convergence and accuracy [49]. The models were trained on 30 epochs with a batch size of 64 and were optimized using AdamW optimizer [52]. An initial learning rate (LR) of $1 \times 10^{-4}$ was set for the training. Cosine Annealing LR with a minimum LR of $1 \times 10^{-6}$ was incorporated as LR Scheduler. All experiments were done on Ubuntu 22 operating system on NVIDIA Quadro RTX 6000 machine.

To deal with overfitting problems, we used data augmentation techniques. Data augmentation increases the variability and robustness of the model since the model becomes new due to the modified versions of the input data [19], [53]. The data augmentation techniques were implemented only in training samples and included image rotation, resizing, color transformation, and noise transformation. All samples were resized into

TABLE III
ACCURACY ASSESSMENT OF THE STUDIED DL NETWORKS IN UNITEMPORAL AND BITEMPORAL INPUTS

| Networks | Training settings | | $F_1$ (%) | | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|---|
| | Input images | Loss func. | No-dmg | Minor | Major | Destr | mean | mean |
| ResNet34 | Unitemporal | CEL | 90.97 | 55.62 | 58.95 | 86.38 | 72.98 | 69.55 |
| ResNeXt50 | Unitemporal | CEL | 91.17 | 54.58 | 61.18 | 86.23 | 73.29 | 69.89 |
| SwinT | Unitemporal | CEL | **92.12** | **60.12** | **64.41** | **88.17** | **76.20** | **73.59** |
| ResNet34 | Bitemporal | CEL | 91.52 | 57.69 | 60.61 | 87.45 | 74.32 | 71.18 |
| ResNeXt50 | Bitemporal | CEL | 91.18 | 58.59 | 60.01 | 87.66 | 74.36 | 71.29 |
| SwinT | Bitemporal | CEL | **92.49** | **60.64** | **66.22** | **88.94** | **77.07** | **74.56** |

Bolds indicate the highest score for each damage class in each input images.

TABLE IV
ACCURACY ASSESSMENT OF THE STUDIED LOSS FUNCTIONS CALCULATED ON THE xBD TEST SET INCLUDING CEL, FOCAL LOSS, AND CE-FOCAL (COMBINED OF CEL AND FOCAL)

| Networks | Training settings | | $F_1$ (%) | | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|---|
| | Inputimages | Loss func. | No-dmg | Minor | Major | Destr | mean | mean |
| SwinT | Unitemporal | CEL | 92.12 | 60.12 | 64.41 | 88.17 | 76.20 | 73.59 |
| SwinT | Unitemporal | Focal | 91.99 | 59.48 | 64.56 | 88.06 | 76.02 | 73.36 |
| SwinT | Unitemporal | CE-Focal | 92.16 | 59.98 | 65.27 | 87.83 | 76.31 | 73.76 |
| SwinT | Bitemporal | CEL | 92.49 | 60.64 | 66.22 | 88.94 | 77.07 | 74.56 |
| SwinT | Bitemporal | Focal | 92.49 | 60.65 | 65.60 | 88.81 | 76.89 | 74.34 |
| SwinT | Bitemporal | CE-Focal | 92.52 | **61.63** | 65.64 | 89.09 | 77.22 | 74.78 |
| EnsSwin | Unitemporal | CE-Focal | **92.51** | **61.39** | **66.12** | **88.55** | **77.14** | **74.74** |
| EnsSwin | Bitemporal | CE-Focal | **92.77** | 61.44 | **67.46** | **89.49** | **77.79** | **75.40** |

Bolds indicate the highest score for each damage class.

128 × 128 pixels and normalized by the mean and standard deviation of training samples. The image augmentation was done using Albumentations library.[5]

### B. DL Models Architecture Evaluation

This part compares the performance of ResNet34, ResNeXt50, and swin transformer calculated on the xBD testing set. In this evaluation, all models were trained using the same settings as in Table II and used CEL as the loss criterion. The training process and score calculation were done using the xBD dataset. While the training and validation sets were used for training, the test set was utilized to evaluate the model performance.

The results of each studied model are listed in Table III. Generally, all models perform best in no-damage and destroyed classes and score lower in middle classes (minor- and major-damage). Swin transformer, however, scores higher in all classes than ResNet and ResNeXt. Moreover, the transformer model scores much higher in the middle classes. For example, swin transformer scores 64.41 in the major-damage class in the unitemporal model. Meanwhile, the scores of the class for ResNet and ResNeXt are 61.18 and 58.95, respectively.

The aforementioned pattern is also found in bitemporal scheme. The scores for both ResNet and ResNeXt are much lower than that of swin transformer. For example, the average $F_1$ for ResNet is 74.32, and ResNeXt is 74.36 compared to 77.07 for the swin transformer. In comparison with the unitemporal input, swin transformer with the Siamese model has a slightly higher score, whereas the unitemporal model achieved 76.20.

These findings show that besides scores higher than ResNet and ResNeXt, swin transformer is more consistent in both input scenarios. Therefore, we used the swin transformer as the baseline model for the loss function evaluation.

### C. Loss Functions Evaluation

This section reports the comparison of loss functions performance. As in the previous section, the scores were calculated on the xBD test set. The results are summarized in Table IV. Among the studied losses, the score of each loss does not vary significantly. All loss functions have a similar pattern where the middle classes remain the minority score compared to no-damage and destroyed classes. For the macroaverage, the scores of all losses and image inputs range from 76.02 to 77.22. As per the harmonic mean, the score is slightly lower considering the score differences among the damage classes.

CE-Focal loss achieved the highest score in both unitemporal and bitemporal models. In unitemporal inputs, CE-Focal achieved an overall score of 76.31 as the highest among others. CEL, with a slightly lower score (76.20), stood in second place. Similar patterns are also observed in bitemporal input images where the average score in descending order is CE-Focal (77.22), CE (77.07), and focal loss (76.89). By looking at the scores in both inputs, CE-Focal shows to be superior to other losses. Therefore, we selected CE-Focal as the standard loss function for the next experiments.

Having found the best scoring model and its loss function, we implemented an ensemble technique involving multiple backbones of swin transformer, namely S, B, and T. In addition, we apply test-time augmentation techniques, including image rotations to each testing sample. We took the average value of each vector resulting from each backbone before feeding it into the Softmax activation function for classification. The ensemble approaches have been implemented in the xBD winning approach

[5] [Online]. Available: https://albumentations.ai/

and [31] to help improve the final results. In addition, ensembling multiple models is reported to outweigh the performance of single models [54] and have a better generalization [55].

The results of the Ensemble models EnsSwin are illustrated in Table IV. Overall, ensemble models yield better performance in all damage levels, even when compared to other loss functions. For unitemporal models, for instance, the highest score was initially achieved by CE-Focal loss with a score of 76.31. By assembling multiple models, the score outweighs this single model. The same pattern is also found in bitemporal input. The ensemble model has increased the overall $F_1$ from 77.22 acquired by a single model of swin transformer to 77.79. In conclusion, assembling multiple models can achieve a higher score than a single model. We then used the swin transformer ensemble model to predict building damage in real-case testing. In addition, the similarity in score between the two input scenarios indicates that postevent-only images can yield satisfactory results in data scarcity.

### D. Application to 2011 Tohoku Tsunami

This section evaluates the generalization performance of the trained model on the 2011 Tohoku Tsunami. It starts by describing the dataset and then reports the prediction results.

*1) Tohoku Dataset:* As in training, the 2011 Tohoku Tsunami dataset also consists of image pairs and building damage annotations. The images used are WorldView3 imagery sensed before August 10, 2010 and after March 11, 2021, the disaster and were resampled into 0.5 × 0.5 m GSD. The images were acquired by the International Research Institute for Desaster Science (IRIDeS), Tohoku University. As for the damage labels, the data were obtained from [56]. The data are available at.[6]

The xBD and MLIT labels differ in several ways, including the number of classes and the method to determine the damage level. For the number of classes, xBD has four damage levels. In contrast, MLIT segregated the building status into seven classes, including no-damage, minor-damage, moderate-damage, major damage, complete damage, collapsed, and washed away. In addition, unlike the xBD dataset, where the damage status is primarily determined through visual interpretation of overhead imagery, MLIT considered the damage level through field inspections. Besides, although some class names are common, they may vary in the class definition. Due to the differences in the number of classes and damage class definition, the two datasets need to be harmonized. In this regard, the Tohoku dataset was reclassified as many xBD classes in the following two different ways.

1) *Semantic Meaning:* In this approach, the class mapping was done according to the semantic meaning of each damage class. For example, the no-damage of Tohoku remains a no-damage class since it has a similar class name in xBD. This method illustrates a scenario where attention is paid to comparing damage descriptions used in different sources.
2) *RS-Based Meaning:* In this technique, the class was mapped based on an RS point-of-view. Satellites are positioned in space and observe the Earth's surface from above, which makes Earth objects, including houses, observed from a top-down perspective, i.e., from the roof. That said, RS only evaluates the building damage from the top of the building rather than from other angles or perspectives. As mentioned in [56], however, the damaged houses were also evaluated from wall inspection and height of inundation. Those factors are not visible from the sky. Therefore, this scenario ensures that the model determines the damage of two datasets from the same perspective. Thus, the samples from each MLIT class are inspected visually and compared with those from xBD. Technically, a number of images per damage class from MLIT were randomly selected and compared to the xBD samples. For obvious classes, such as no-damage and washed away, 80 samples were selected. For the middle classes, one percent of the samples were inspected. Those samples are matched based on visual similarity in depicting building damage. Since the xBD is treated as a reference, each MLIT class was mapped to the xBD classes. Based on the visual inspection, each damage class from MLIT has a general agreement toward a particular class of xBD. For example, major-damage images of the Tohoku dataset do not show damage from the RS images. In this case, the major-damage class of the Tohoku is assigned as the no-damage class.

The class mapping methods between xBD and Tohoku dataset are summarized in Table V and Fig. 4 for the image sample comparison.

*2) Prediction Results:* This part reports the prediction of our model (swin transformer-based) trained on xBD data for the 2011 Tohoku Tsunami event. First, we present the results of different labeling schemes, and finally, we compare the result with the first place of the xView-2 Challenge.

The performance comparison is illustrated in Table VI and Fig. 5. Comparing labeling techniques illustrates that semantic meaning yields a lower score than RS-based meaning. For instance, the average score of semantic-based labeling for bitemporal is 27.30. Meanwhile, the value for RS-based labeling is 47.29. In other words, scores for all classes increased when RS-based labeling was used. Specifically, the no-damage class score rose significantly from 12.57 to 75.18 in the unitemporal model.

In this study area, the bitemporal network yields a higher score than single temporal inputs, especially in no-damage and destroyed classes. The changes between pre- and postdisaster images may cause this. Since the bitemporal model takes input from both images, any significant change, such as in destroyed classes or unchanged images as in no-damage classes, is more recognized by the model than changes in the middle classes (minor- and major-damage).

*3) Comparison With the xView-2 Challenge Solution:* The winner adopted a Siamese network structure with bitemporal inputs, the details of which are available at.[7] Since the xView-2 Challenge was designed for segmentation tasks, it results in a semantic map. We then calculated the score at both pixel and

[6] [Online]. Available: http://fukkou.csis.u-tokyo.ac.jp/dataset/list_all

[7] [Online]. Available: https://github.com/DIUx-xView/xView2_first_place

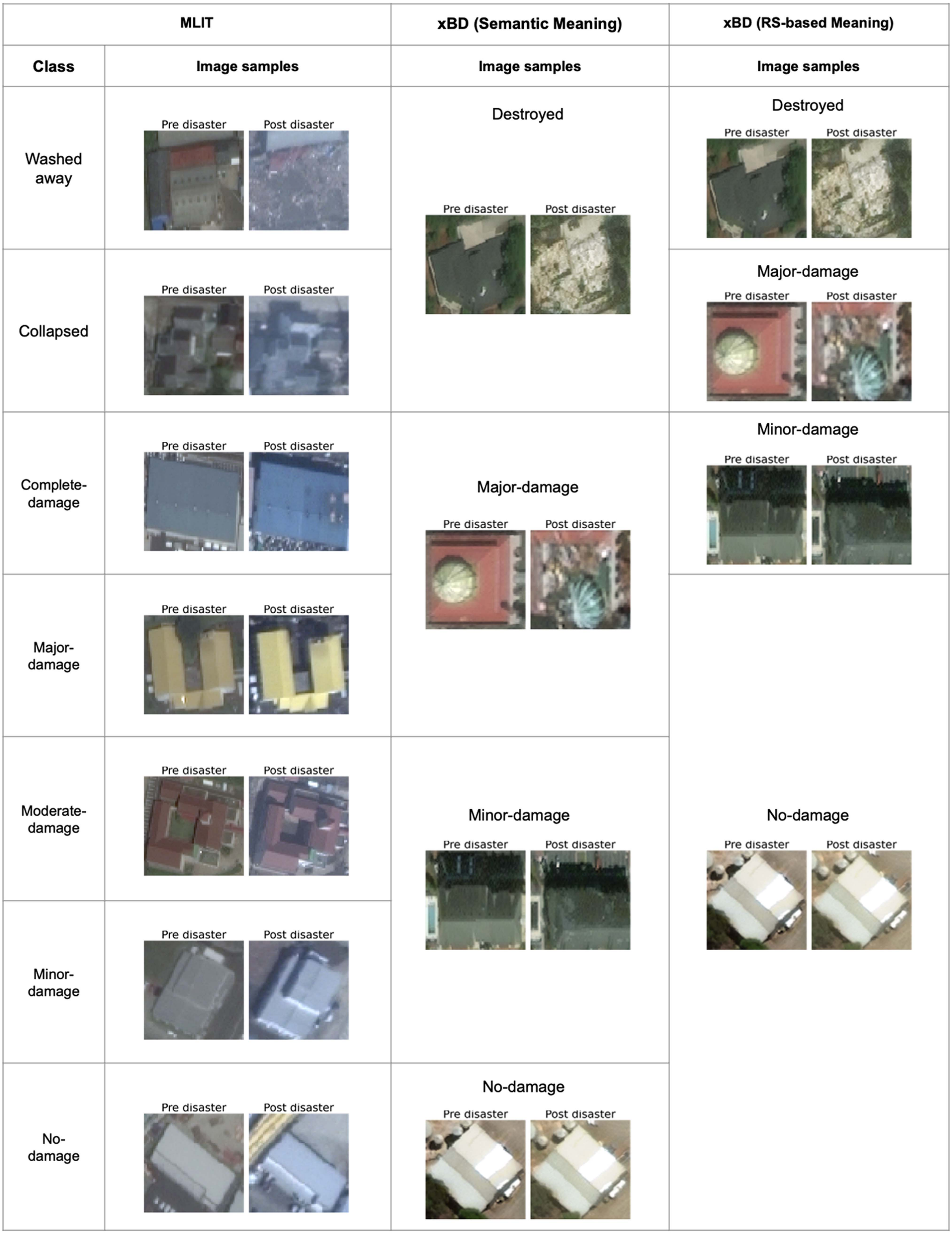


Fig. 4. Image samples of each damage class. Semantic Meaning matches the xBD and MLIT based on damage descriptor similarity. RS-based Meaning maps the two datasets based on the visual similarity of each sample.

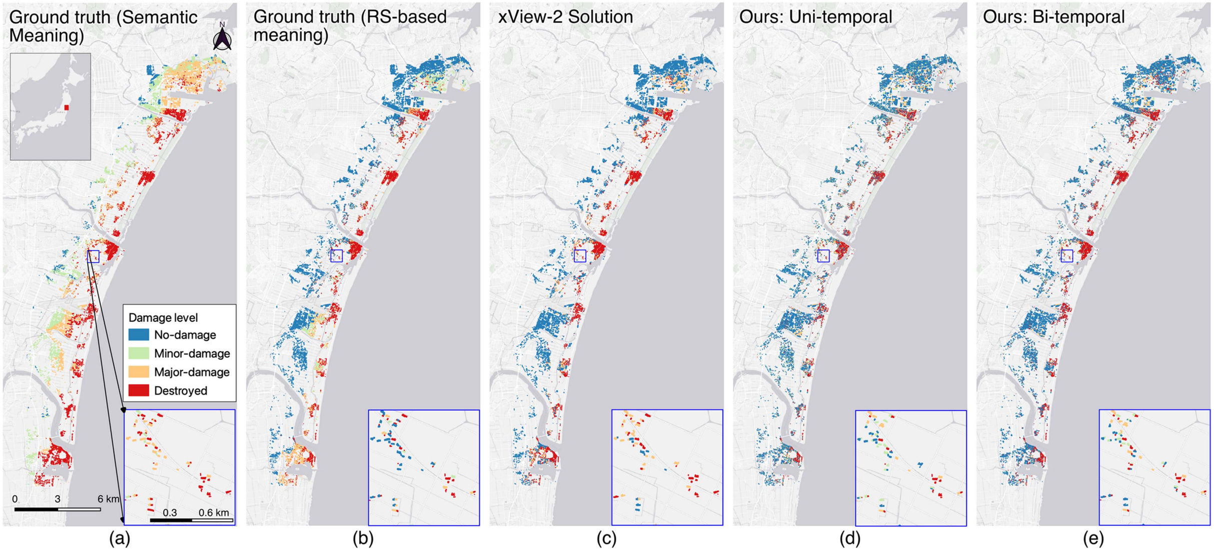
TABLE V
SUMMARY OF DAMAGE CLASS MATCHING BETWEEN XBD AND MLIT

| Code | xBD | MLIT (Semantic Meaning) | | MLIT (RS-based Meaning) | |
|---|---|---|---|---|---|
| | | Classes | # Samples | Classes | # Samples |
| 0 | No-damage | No-damage | 1565 | No-, minor-, moderate-, major-damage | 16 725 |
| 1 | Minor-damage | Minor-, moderate-damage | 9163 | Complete damage | 1697 |
| 2 | Major-damage | Major-, complete damage | 7694 | Collapsed | 3581 |
| 3 | Destroyed | Collapsed, washed away | 12 352 | Washed away | 8771 |

TABLE VI
ACCURACY ASSESSMENT OF XVIEW-2 SOLUTION AND OUR MODEL CALCULATED ON 2011 TOHOKU TSUNAMI

| Networks | Training settings | | $F_1$ (%) | | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|---|
| | Input images | Label | No-dmg | Minor | Major | Destr | mean | mean |
| xView-2 | Pixel | Semantic | 1.68 | 0.00 | **22.08** | 33.75 | 14.38 | NaN |
| xView-2 | Building | Semantic | **17.37** | 0.00 | 20.06 | 74.96 | **28.10** | NaN |
| Ours | Unitemporal | Semantic | 12.57 | **16.50** | 11.93 | 63.80 | 26.20 | **16.69** |
| Ours | Bitemporal | Semantic | 13.29 | 9.99 | 10.77 | **75.17** | 27.30 | 14.21 |
| xView-2 | Pixel | RS-based | 14.86 | 0.00 | 15.19 | 33.74 | 15.95 | NaN |
| xView-2 | Building | RS-based | 79.95 | 0.00 | **27.38** | **81.01** | 47.08 | NaN |
| Ours | Unitemporal | RS-based | 75.18 | **10.84** | 15.64 | 74.50 | 44.04 | **21.87** |
| Ours | Bitemporal | RS-based | **80.60** | 9.24 | 14.77 | 84.57 | **47.29** | 19.98 |

Bolds indicate the highest score for each damage class in each label setting.



(a) (b) (c) (d) (e)

Fig. 5. Building damage distribution of Tohoku Tsunami estimated by xView-2 solution and our model. (a) Ground truth reclassified from MLIT data using the semantic meaning method. (b) Ground truth obtained from the RS-based meaning approach. (c) Building damage predicted by the xView-2 solution. (d) and (e) Predicted results of our model using uni, and bitemporal inputs, respectively.

building-levels. For the pixel level, $F_1$ is directly estimated by comparing the semantic maps with ground truth maps. As for the building level, the damage status of a building is determined by the majority voting of pixels falling into each building [57].

Regarding the xView-2 approach, pixel-level scores are the lowest among all testing. Moreover, it could not detect minor-damage classes. The lower score in the pixel-level evaluation may be caused by pixel-wise methods being sensitive to different values between predicted and ground truth data. However, the sensitivity is reduced at the building level since the label decision depends on the majority pixel class falling into a building. For building-level testing, the patterns are similar to that of the swin transformer-based model. Although the xView-2's average score is higher than our unitemporal model, it is slightly lower than the bitemporal model. In addition, as in the pixel-level score, the result from semantic mapping could not detect minor-damage classes.

### E. Application to 2023 Türkiye–Syria Earthquake

Besides the 2011 Tohoku Tsunami, we utilize our model to predict the damage caused by the 2023 Türkiye–Syria Earthquake. On February 6, 2023, two earthquakes with a magnitude of 7.8 and 7.5 hit Kahramanmaraş, Türkiye [58]. According to [59], the disaster affected 11 provinces and have killed 45 089 lives, and displaced 1 971 589 people.

The study focuses on the Islahiye City of Türkiye as one of the areas that were heavily affected by the disaster. The event and study area locations are depicted in Fig. 6. In this evaluation, the images for both pre- and postdisaster are obtained from the

TABLE VII
ACCURACY ASSESSMENT OF XVIEW-2 SOLUTION AND OUR MODEL CALCULATED ON SATELLITE IMAGERY OF 2023 TÜRKIYE–SYRIA EARTHQUAKE

| Networks | Training settings | | $F_1$ (%) | | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|---|
| | Input images | Images | No-dmg | Minor | Major | Destr | mean | mean |
| xView-2 | Pixel | Satellite | 63.72 | 0.00 | 6.01 | 37.93 | 26.92 | NaN |
| xView-2 | Building | Satellite | 77.12 | 0.00 | **7.55** | 43.56 | 32.06 | NaN |
| Ours | Unitemporal | Satellite | 91.50 | 0.00 | 5.41 | 32.89 | 32.45 | NaN |
| Ours | Bitemporal | Satellite | **92.95** | **1.64** | **9.74** | 40.87 | **36.30** | **5.35** |

Bolds indicate the highest score for each damage class.

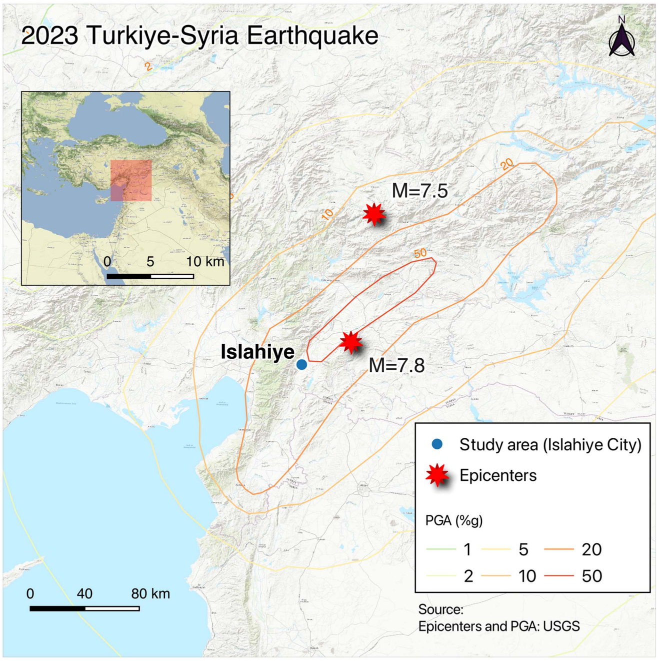
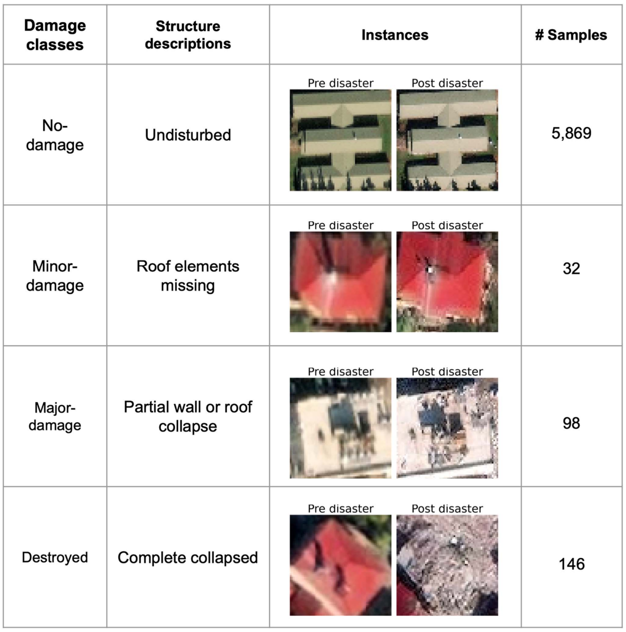


Fig. 6. Overview of the 2023 Türkiye–Syria Earthquake. A blue dot indicates Islahiye City as the study area.



| Damage classes | Structure descriptions | Instances | # Samples |
|---|---|---|---|
| No-damage | Undisturbed | Pre disaster Post disaster | 5,869 |
| Minor-damage | Roof elements missing | Pre disaster Post disaster | 32 |
| Major-damage | Partial wall or roof collapse | Pre disaster Post disaster | 98 |
| Destroyed | Complete collapsed | Pre disaster Post disaster | 146 |

Fig. 7. Damage descriptions for building damage inspection in 2023 Türkiye–Syria Earthquake.

Maxar Open Data Program.[8] The images were acquired from the WoldView3 satellite on December 27, 2022 and February 7, 2023 for pre- and postevent, respectively. Both temporal images have a GSD of approximately 0.3 m. As in the 2011 Tohoku Tsunami, the Türkiye Dataset was also not included in the training.

Besides the satellite imagery, the testing also involved an aerial photo. The photo after the disaster was acquired on February 15, 2023 and is available publicly from OpenAerialMap.[9] In addition, the OSM building footprint was used to generate the building patches. The polygons were updated and aligned following the satellite images. The building footprints were removed or added according to the building's existence in the predisaster image.

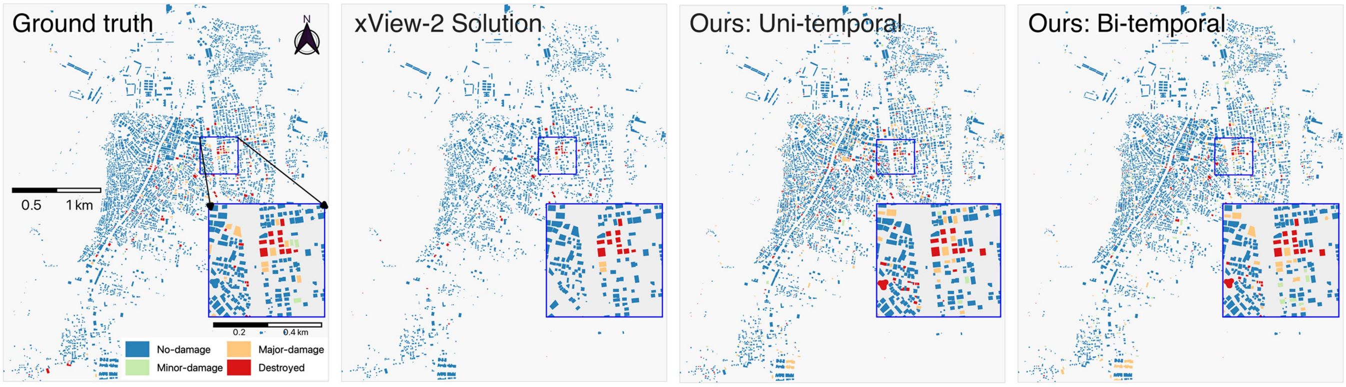
To evaluate the performance of our model, we constructed reference labels of 6145 buildings through visual interpretation of satellite imagery. The adjusted building footprint was classified into four classes to comply with the xBD dataset. Moreover, the damage description used in xBD was also adopted for the Islahiye damage inspection. The details of the damage description are illustrated in Fig. 7.

*1) Testing on Satellite Imagery:* The results of model testing for the Türkiye–Syria Earthquake are reported in Table VII. Like the Tohoku Tsunami results, the $F_1$ for Türkiye Earthquake varies in every damage class. Generally, the model can predict well in the no-damage class with a score of 91.50 for unitemporal and 92.95 for bitemporal. The second highest scoring class is the destroyed class, with a score of 32.89 for unitemporal and 40.87 for bitemporal. The minor- and major-damage classes score much lower, which causes the average score to reach only 32.45 and 36.30 for unitemporal and bitemporal, respectively.

For segmentation-based results, the score for pixel-based evaluation is much lower than building-based. A similar scoring pattern of the transformer-based model is also found in segmentation-based results. The no-damage and destroyed are two chiefly scoring classes. In segmentation-based scores, however, the model could not detect minor-damage classes. On average, our model acquired higher accuracy compared to the semantic-segmentation-based model. This is true for both input scenarios. The building damage distribution predicted by the models is depicted in Fig. 8.

[8] [Online]. Available: https://www.maxar.com/open-dataturkey-earthquake-2023

[9] [Online]. Available: https://openaerialmap.org/



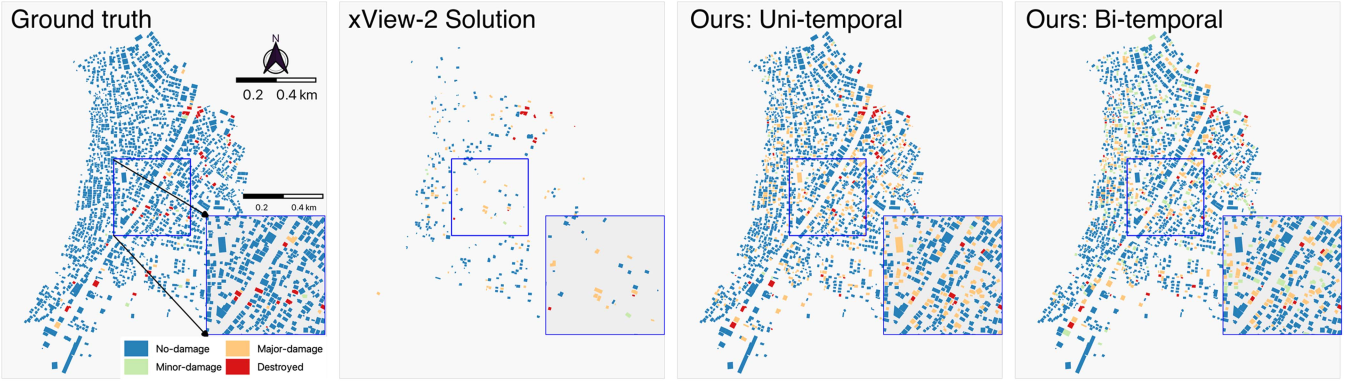
(a) (b) (c) (d)

Fig. 8. Building damage distribution in the Islahiye City of Türkiye Earthquake estimated by xView-2 solution and our model. The maps were calculated on satellite imagery. (a) Ground truth obtained through the visual interpretation of satellite imagery. (b) Building damage predicted by the xView-2 solution at the building level. (c) and (d) Predicted results of our model using uni, and bitemporal inputs, respectively.

TABLE VIII
ACCURACY ASSESSMENT OF XVIEW-2 SOLUTION AND OUR MODEL CALCULATED ON THE AERIAL PHOTO OF 2023 TÜRKIYE–SYRIA EARTHQUAKE

| Networks | Training settings | | $F_1$ (%) | | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|---|
| | Input images | Images | No-dmg | Minor | Major | Destroyed | mean | mean |
| xView-2 | Pixel | Aerial photo | 15.78 | 0.00 | 2.68 | 11.32 | 7.44 | NaN |
| xView-2 | Building | Aerial photo | 18.99 | 0.00 | **7.90** | 15.09 | 10.50 | NaN |
| Ours | Unitemporal | Aerial photo | **87.45** | 0.00 | 4.05 | **50.53** | 35.51 | NaN |
| Ours | Bitemporal | Aerial photo | 84.79 | 0.00 | 4.76 | 42.86 | **33.10** | NaN |

Bolds indicate the highest score for each damage class.



(a) (b) (c) (d)

Fig. 9. Building damage distribution of Türkiye Earthquake estimated by xView-2 solution and our model. The maps were calculated on the aerial photo. (a) Ground truth obtained through the visual interpretation of satellite imagery. (b) Building damage predicted by the xView-2 solution at the building level. (c) and (d) Predicted results of our model using uni, and bitemporal inputs, respectively.

*2) Testing on Aerial Photo:* The aerial photo has a GSD of around 8 centimeters. However, it only has postdisaster (acquired on February 14, 2023) and covers a smaller area than satellite imagery. The pre-event image of the satellite imagery was then used as an image pair by resampling and cropping it according to GSD and the extent of the aerial image. The results of the transferability test on aerial imagery are depicted in Table VIII for the summary and Fig. 9 for the damage distribution.

The scores of aerial photo testing for our model have a similar pattern as in satellite imagery. Although the model could not detect the minor-damage class, it scored higher in the no-damage and destroyed class, resulting in a higher average score. In contrast, the xView-2 solution drops in scores for both building- and pixel levels. Besides not detecting minor-damage class, the scores for no-damage and destroyed dropped significantly.

As depicted in Fig. 9(b), it is noticeable that xView-2 solution's map has fewer buildings than other maps. This is mainly because the building damage is determined by the majority class of each pixel falling into each building polygon. In a semantic segmentation task, there is any possibility that pixels are classified as background instead of any

TABLE IX
ACCURACY ASSESSMENT OF xVIEW-2 SOLUTION AND OUR MODEL IN THREE-CLASS OF DAMAGE

| Networks | Training settings | | $F_1$ (%) | | | Arithmetic | Harmonic |
|---|---|---|---|---|---|---|---|
| | Input images | Test set | No-damage | Damage | Destroyed | mean | mean |
| xView-2 | Pixel | xBD | 84.65 | 67.17 | 71.38 | 74.40 | 73.69 |
| xView-2 | Building | xBD | 90.77 | **80.70** | 78.78 | 83.42 | 83.10 |
| Ours | Unitemporal | xBD | 92.51 | 73.92 | 88.55 | 84.99 | 84.20 |
| Ours | Bitemporal | xBD | **92.77** | 74.41 | **89.49** | **85.56** | **84.76** |
| xView-2 | Pixel | Tohoku (Semantic) | 1.68 | 17.03 | 33.75 | 17.49 | 4.39 |
| xView-2 | Building | Tohoku (Semantic) | **17.40** | 12.20 | 75.00 | 34.87 | 19.64 |
| Ours | Unitemporal | Tohoku (Semantic) | 12.57 | **29.52** | 63.80 | 35.30 | **23.24** |
| Ours | Bitemporal | Tohoku (Semantic) | 13.29 | 18.01 | **75.17** | **35.49** | 20.82 |
| xView-2 | Pixel | Tohoku (RS-based) | 14.86 | 25.26 | 33.74 | 24.62 | 21.98 |
| xView-2 | Building | Tohoku (RS-based) | 79.90 | **32.90** | 81.00 | **64.60** | **54.29** |
| Ours | Unitemporal | Tohoku (RS-based) | 75.18 | 28.98 | 74.50 | 59.55 | 48.99 |
| Ours | Bitemporal | Tohoku (RS-based) | **80.60** | 23.57 | **84.57** | 62.91 | 45.01 |
| xView-2 | Pixel | Islahiye (Satellite) | 63.70 | 5.10 | 37.90 | 35.57 | 12.60 |
| xView-2 | Building | Islahiye (Satellite) | 77.10 | 5.80 | **43.60** | 42.17 | 14.40 |
| Ours | Unitemporal | Islahiye (Satellite) | 91.50 | 6.47 | 32.89 | 43.62 | 15.31 |
| Ours | Bitemporal | Islahiye (Satellite) | **92.95** | **9.78** | 40.87 | **47.87** | **21.82** |
| xView-2 | Pixel | Islahiye (Aerial) | 15.80 | 4.60 | 11.30 | 10.57 | 8.13 |
| xView-2 | Building | Islahiye (Aerial) | 19.00 | **9.00** | 15.10 | 14.37 | 13.04 |
| Ours | Unitemporal | Islahiye (Aerial) | **87.45** | 7.42 | **50.53** | **48.47** | **18.07** |
| Ours | Bitemporal | Islahiye (Aerial) | 84.79 | 5.49 | 42.86 | 44.38 | 13.81 |

Bolds indicate the highest score for each damage class in each test set.

damage class. In this case, when most pixels inside a building polygon are classified as "background," the class for the building is background. This indicates that many buildings are unidentifiable by the models and, thus, become nonbuilding, hence, no polygons. These experiments show that an image classification scheme yields more stable results when applied to images from different sources, even though they vary in resolution.

### *F. Reduced Classes*

The previous section shows that the models perform well in no-damage and destroyed classes. However, the scores drop in the middle classes. The study then experiments with combining the middle classes (minor- and major-damage) into a single class called damage. The classification results for the xBD testing set, Tohoku, and Islahiye are outlined in Table IX.

As listed in Table IX, the overall performance in all testing datasets has increased by merging the middle classes. For the unitemporal of the xBD test set, the score has risen from 77.14 in the four-class to 84.99 in the three-class. No-damage and destroyed classes remain the two chiefly classes with the values of 92.51 and 88.55, respectively. The damage class stood as the smallest score. However, the merged class has a higher score than the original classes. After merging, the score is 73.92, rising from only 61.39 and 66.12 in the original class for minor- and major-damage classes, respectively. This pattern is similar in bitemporal with slightly higher scores.

For the emergency scenario testing, the generalization scores have a similar pattern to the four-class scheme. However, the overall scores have increased in the reduced-class scheme. Moreover, unlike in the four-class, where no-damage is generally not recognized by the xView-2 model, all classes in the three-class are detectable. xView-2 solution generalized better than the transformer-based model in RS-based labeling of Tohoku Tsunami yet scored lower in Semantic-based labeling. Meanwhile, the transformer-based performed better in Islahiye for both satellite and aerial images.

## IV. DISCUSSION

In this study, we did extensive experiments in selecting a best-performing model to predict damage from new disaster events, illustrating the rapid damage assessment in emergency settings. First, we computed the score on the xBD test set to find the best model. Then, we used the model to recognize damage in the unseen dataset. Overall, the model performs well in the xBD test set. However, when the performance per damage class is compared, the score is unequal. Specifically, the descending order of the damage class in terms of $F_1$ is no-damage, destroyed, major-, and minor-damage. This may be affected by several causes, including intraclass discrepancy and sample imbalance.

The fact that the no-damage and destroyed classes are the two chiefly classes indicates that the model performs well according to the intraclass discrepancy. As illustrated in Fig. 4, the figures for the destroyed and no-damage classes are robust. For the destroyed class, the image has completely changed from a building into ruins. For the no-damage class, images in both pre- and postdisaster show no change. In contrast, the change is not so distinguishable for the middle class. For minor-damage classes, for instance, some changes are only slightly visible. This may be the source of confusion for the model in recognizing the damage in the middle class. Besides, this may also reflect one of the challenges of optical RS. As the system is limited to nadir-looking, the image may also have a limited perspective of the objects. Any geometrical changes will be difficult to detect in optical RS. In fact, the degree of changes in buildings' geometry may indicate the level of damage experienced by the buildings. Moreover, as depicted in Fig. 3, the no-damage class occupies most of the total samples. It could be another reason why the no-damage class is the top score in all models and all input scenarios. As the samples for this class are abundant, the model has more chances to recognize the pattern for the no-damage class.

Although achieved high scores on the xBD test split, the scores generally dropped when the trained model was used in new unseen datasets. A similar finding is also reported by

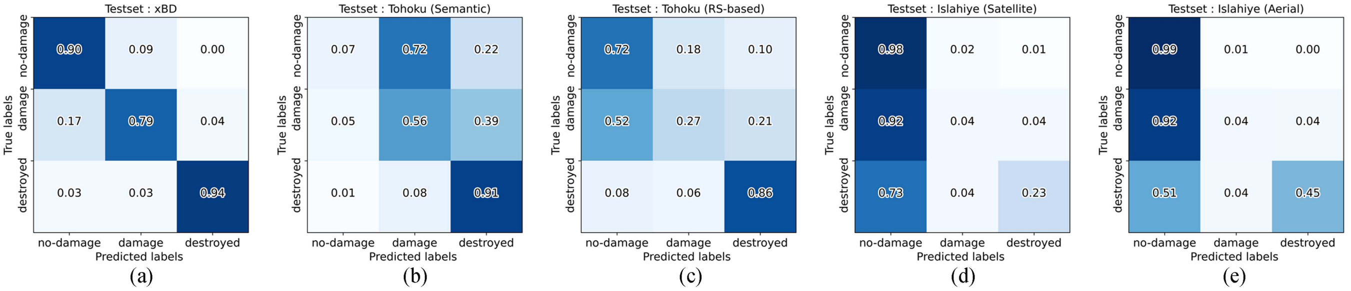


Fig. 10. Confusion matrices of normalized $F_1$ of the reduced classes. The middle (damage) class is a merged classes of minor- and major-damage. The matrices are generated from unitemporal model. (a)–(e) refers to the different datasets used for testing where (a) xBD, (b) Tohoku using semantic labeling, (c) Tohoku under RS-based labeling, (d) satellite imagery of Islahiye, and (e) aerial photo of Islahiye city.

some studies, including [5], [6], [35], [36], and [34]. In [60], due to the fall in OOD score, few samples from the target area were collected for fine-tuning the model. Yang et al. [6] pointed out that one of the reasons for the drop is a difference in geographical locations. The building variation in size, structure, and surrounding environment makes it difficult to generalize in new locations. In addition, the difference in the images used for training (source) and testing (target) may have been the cause of the fall in performance in the unseen datasets. As the images were acquired in a different sensing condition, e.g., atmospheric conditions [36], both source and target images will have different data distributions and may suffer from domain shifts. Additional techniques, such as domain adaption, which allows the source and target domain to have invariant features, can be a direction for future studies.

To evaluate the results further, confusion matrices are provided in Fig. 10. The figure provides the normalized $F_1$ of our model testing on different datasets computed using the unitemporal model. The matrix for the xBD test set shows that most samples are classified correctly, indicating that the model learned relevant features from the data. This is understandable since the model learned from the same data subset. For the OOD testing, however, the matrices show undesirable results. For Tohoku Tsunami, samples from a particular class tend to be misclassified as their adjacent classes. The pattern is more visible in semantic-based labeling. While for the RS-based labeling, the model tends to be classified as a no-damage class, especially from the middle class.

For the Türkiye earthquake, the model has a higher tendency to predict no-damage classes. This pattern is true for both satellite and aerial photo images. The fact that the model has a tendency to predict no-damage in OOD is probably caused by the sample imbalance. As shown in Fig. 3, the no-damage class occupies the majority of samples. This makes the model likely to predict a no-damage class, as the model learned more about no-damage features.

Merging middle classes (minor- and major-damage classes) aims to improve the middle classes' performance. Although insignificant, this approach can increase the overall scores, especially the merged classes. The increase is also found in both the Tohoku and Islahiye datasets. In [16], the multiclass is compared with the binary class. Their experiments show that reducing class means reducing the complexity of the task and, hence, can improve the results when the task is simpler. Similarly, Yang et al. [6] found that multiclass acquires lower performance than binary models. In [17], utilizing only unitemporal images acquires high accuracy in binary mapping.

A similar approach as ours has been done by [40]. xBD dataset was used as weight initialization. Although the xBD dataset has four damage categories, they only use binary classes (damaged/undamaged) for their final output. This indicates that utilizing optical RS images and current DL models still works better in smaller damage classes, especially for OOD generalization purposes. To improve the generalization in middle classes, a multimodal approach will be required, for example, by taking advantage of SAR images. With their side-looking ability, SAR systems can add more perspective and stronger characteristics to each category, which eventually can help models identify different classes.

As for the generalization test, we use the 2011 Tohoku Dataset as one of the studies. As the Tohoku Dataset and xBD vary in terms of total damage class and damage determination, we experiment with two label mapping techniques: based on semantic meaning and RS-based meaning. The increase in the score due to the change in labeling approaches shows that the model is sensitive to the damage definitions. The score gets higher when the samples are mapped based on visual similarity rather than semantic definition. Specifically, MLIT determined the damage level by looking at the wall condition, which is not visible from the roof. Meanwhile, the model only decides the damage states from the top. In this case, the model may determine such buildings as no damage since no ruins are visible from the top. Owing to this factor, a higher score was achieved when the class definition was mapped according to the visual similarity between each class. In other words, visual-based matching gains higher performance since the ground truth data is defined in a similar method as the model was trained on.

This experiment shows that label matching is an important factor to determine, especially when the training and testing are of different sources. Specifically, the class matching should be conducted as close as possible to the scheme used for training processes. Since the xBD labels (and most RS-based detection) are obtained from visual interpretation, class matching would preferably use the same approach. This also emphasizes the limitation of RS methods, where its ability is mostly limited to roof appearance.

## V. Conclusion

This study replicates building damage identification in an emergency response context. In this setting, ground truth data is unavailable. Therefore, we rely on historical data to predict building damage from an event unseen during the training. We utilized a large disaster damage dataset compiled from multiple major-disaster events around the globe. Extensive experiments have been done involving multiple loss functions and CNN-based models, including transformer-based models. Besides, we also compared two inputs based on data availability, namely unitemporal and bitemporal where we designed a Siamese model. The best-performing model was then used to predict building damage caused by the 2011 Tohoku Tsunami and the 2023 Türkiye–Syria earthquake. We also compared the best model to the state-of-the-art building damage recognition (xView-2 Challenge solution).

Calculated on the xBD test set, the experiments show that swin transformer with CE-Focal Loss is the best combination. The model performs best in no-damage and destroyed classes. However, the model scores lower in the middle classes (minor- and major-damage). When the trained model is used to predict the damage of two different disasters, the model can maintain its high performance in no-damage and destroyed classes. However, the scores dropped in both minor- and major-damage classes, making the average scores fall. This pattern is found in both input scenarios, indicating that unitemporal input can perform satisfyingly in case of data scarcity. However, bitemporal input shows a higher score in almost all cases. Therefore, when data are available, utilizing predisaster images is preferred.

In comparison with xView-2 solution, since there is still a decrease in performance, there is comparably stable generalization. While the transformer-based model maintains its performance in satellite and aerial images, the xView-2 model's score drops in predicting aerial imagery damage. Moreover, the xView-2-based model generally can only detect no-damage, major-damage, and destroyed classes. In contrast, our model, especially bitemporal inputs, can detect all damage classes, including minor-damage ones.

To summarize, our study shows that current DL models perform inferior in predicting damage to buildings on a new event. This may be accounted to the geographical difference and domain shift between the source (where the model is trained on) and target domains (unseen disaster events). In future research, we will extend our study by implementing domain adaptation techniques to reduce the domain gap between the source and target domains and involving SAR imagery for a multimodal approach.

## References

[1] M. A. Günen, "Performance comparison of deep learning and machine learning methods in determining wetland water areas using Eurosat dataset," *Environ. Sci. Pollut. Res.*, vol. 29, no. 14, pp. 21092–21106, 2022.

[2] A. Vetrivel, M. Gerke, N. Kerle, F. Nex, and G. Vosselman, "Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 140, pp. 45–59, 2018.

[3] Z. Yu et al., "Segdetector: A deep learning model for detecting small and overlapping damaged buildings in satellite images," *Remote Sens.*, vol. 14, no. 23, 2022, Art. no. 6136.

[4] R. Gupta and M. Shah, "RescueNet: Joint building segmentation and damage assessment from satellite imagery," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4405–4411.

[5] V. Benson and A. Ecker, "Assessing out-of-domain generalization for robust building damage detection," 2020, *arXiv:2011.10328*.

[6] W. Yang, X. Zhang, and P. Luo, "Transferability of convolutional neural network models for identifying damaged buildings due to earthquake," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 504.

[7] V. Zahs, K. Anders, J. Kohns, A. Stark, and B. Höfle, "Classification of structural building damage grades from multi-temporal photogrammetric point clouds using a machine learning model trained on virtual laser scanning data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103406.

[8] B. Adriano, N. Yokoya, J. Xia, G. Baier, and S. Koshimura, "Cross-domain-classification of tsunami damage via data simulation and residual-network-derived features from multi-source images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 4947–4950.

[9] S. Koshimura, L. Moya, E. Mas, and Y. Bai, "Tsunami damage detection with remote sensing: A review," *Geosciences*, vol. 10, no. 5, 2020, Art. no. 177.

[10] L. Dong and J. Shan, "A comprehensive review of earthquake-induced building damage detection with remote sensing techniques," *ISPRS J. Photogrammetry Remote Sens.*, vol. 84, pp. 85–99, 2013.

[11] G. Taskin, E. Erten, and E. O. Alatas, "A review on multi-temporal earthquake assessment damage using satellite images," in *Change Detection and Image Time Series Analysis 2: Supervised Methods*. Hoboken, NJ, USA: Wiley, 2021, Art. no. 155.

[12] P. Ge, H. Gokon, and K. Meguro, "A review on synthetic aperture radar-based building damage assessment in disasters," *Remote Sens. Environ.*, vol. 240, 2020, Art. no. 111693.

[13] K. Kouchi and F. Yamazaki, "Characteristics of tsunami-affected areas in moderate-resolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1650–1657, Jun. 2007.

[14] S.-W. Chen, X.-S. Wang, and S.-P. Xiao, "Urban damage level mapping based on co-polarization coherence pattern using multitemporal polarimetric SAR data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2657–2667, Aug. 2018.

[15] Y. Yusuf, M. Matsuoka, and F. Yamazaki, "Damage assessment after 2001 Gujarat earthquake using Landsat-7 satellite images," *J. Indian Soc. Remote Sens.*, vol. 29, pp. 17–22, 2001.

[16] T. Valentijn, J. Margutti, M. van den Homberg, and J. Laaksonen, "Multi-hazard and spatial transferability of a CNN for automated building damage assessment," *Remote Sens.*, vol. 12, no. 17, 2020, Art. no. 2839.

[17] S. T. Seydi, M. Hasanlou, J. Chanussot, and P. Ghamisi, "BDD-Net+: A building damage detection framework based on modified coat-net," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4232–4247, Apr. 2023.

[18] Y. Bai et al., "Pyramid pooling module-based semi-siamese network: A benchmark model for assessing building damage from XBD satellite imagery datasets," *Remote Sens.*, vol. 12, no. 24, 2020, Art. no. 4055.

[19] 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.

[20] B. Adriano et al., "Learning from multimodal and multitemporal earth observation data for building damage mapping," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 132–143, 2021.

[21] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogrammetry Remote Sens.*, vol. 152, pp. 166–177, 2019.

[22] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[23] H. Chen, E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley, "Dual-tasks siamese transformer framework for building damage assessment," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 1600–1603.

[24] L. Wang, S. Fang, X. Meng, and R. Li, "Building extraction with vision transformer," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625711.

[25] J. Xia, N. Yokoya, B. Adriano, and C. Broni-Bediako, "Openearthmap: A benchmark dataset for global high-resolution land cover mapping," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6254–6264.

[26] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, “Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[27] J. Li, X. Huang, and L. Tu, “WHU-OHS: A benchmark dataset for large-scale hersepctral image classification,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 113, 2022, Art. no. 103022.

[28] R. Gupta et al., “XBD: A dataset for assessing building damage from satellite imagery,” 2019, *arXiv:1911.09296*.

[29] L. Deng and Y. Wang, “Post-disaster building damage assessment based on improved U-Net,” *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 15862.

[30] 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, May 2021, Art. no. 5402114.

[31] 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.

[32] H. Xie, X. Hu, H. Jiang, and J. Zhang, “BSSNet: Building subclass segmentation from satellite images using boundary guidance and contrastive learning,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7700–7711, Aug. 2022.

[33] Y. Hu and H. Tang, “On the generalization ability of a global model for rapid building mapping from heterogeneous satellite images of multiple natural disaster scenarios,” *Remote Sens.*, vol. 13, no. 5, 2021, Art. no. 984.

[34] C. M. Gevaert and M. Belgiu, “Assessing the generalization capability of deep learning networks for aerial image classification using landscape metrics,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 114, 2022, Art. no. 103054.

[35] J. Z. Xu, W. Lu, Z. Li, P. Khaitan, and V. Zaytseva, “Building damage detection in satellite imagery using convolutional neural networks,” 2019, *arXiv:1910.06444*.

[36] F. Nex, D. Duarte, F. G. Tonolo, and N. Kerle, “Structural building damage detection with deep learning: Assessment of a state-of-the-art CNN in operational conditions,” *Remote Sens.*, vol. 11, no. 23, 2019, Art. no. 2765.

[37] I. Bouchard, M.-È. Rancourt, D. Aloise, and F. Kalaitzis, “On transfer learning for building damage assessment from satellite imagery in emergency contexts,” *Remote Sens.*, vol. 14, no. 11, 2022, Art. no. 2532.

[38] G. Abdi, M. Esfandiari, and S. Jabari, “Building damage detection in post-event high-resolution imagery using deep transfer learning,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 531–534.

[39] Y. Endo, B. Adriano, E. Mas, and S. Koshimura, “New insights into multiclass damage classification of tsunami-induced building damage from SAR images,” *Remote Sens.*, vol. 10, no. 12, 2018, Art. no. 2059.

[40] C. Robinson et al., “Turkey building damage assessment,” *Microsoft*, 2023.

[41] J. Su et al., “Technical solution discussion for key challenges of operational convolutional neural network-based building-damage assessment from satellite imagery: Perspective from benchmark XBD dataset,” *Remote Sens.*, vol. 12, no. 22, 2020, Art. no. 3808.

[42] 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.

[43] T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, and A. Plaza, “A siamese network based U-Net for change detection in high resolution remote sensing images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2357–2369, Mar. 2022.

[44] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, “Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15193–15202.

[45] R. Gupta et al., “Creating XBD: A dataset for assessing building damage from satellite imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 10–17.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[47] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[48] Z. Liu et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[49] Y. Ogawa, C. Zhao, T. Oki, S. Chen, and Y. Sekimoto, “Deep learning approach for classifying the built year and structure of individual buildings by automatically linking street view images and GIS building data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1740–1755, Jan. 2023.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[51] D. R. Cox, “The regression analysis of binary sequences,” *J. Roy. Statist. Soc. Ser. B: Statist. Methodol.*, vol. 20, no. 2, pp. 215–232, 1958.

[52] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, *arXiv:1711.05101*.

[53] S. Kaur et al., “Transfer learning-based automatic hurricane damage detection using satellite images,” *Electronics*, vol. 11, no. 9, 2022, Art. no. 1448.

[54] D. Cao, H. Xing, M. S. Wong, M.-P. Kwan, H. Xing, and Y. Meng, “A stacking ensemble deep learning model for building extraction from remote sensing images,” *Remote Sens.*, vol. 13, no. 19, 2021, Art. no. 3898.

[55] X. Zhao, J. Hu, L. Mou, Z. Xiong, and X. X. Zhu, “Cross-city landuse classification of remote sensing images via deep transfer learning,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103358.

[56] Ministry of Land, Infrastructure Transport and Tourism (MLIT), “Summary of the investigation on reconstruction methods for tsunami-damaged areas after the Great East Japan earthquake and tsunami (in Japanese),” 2011. Accessed: May 19, 2011. [Online]. Available: https://www.mlit.go.jp/toshi/toshi-hukkou-arkaibu.html

[57] 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.

[58] United States Geological Survey (USGS), “The M7.8 and M7.5 Kahramanmaraş earthquake sequence struck near Nurdağı, Turkey (Türkiye) on Feb. 6, 2023,” 2023. Accessed: Apr. 30, 2023. [Online]. Available: https://earthquake.usgs.gov/storymap/index-turkey2023.html

[59] Disaster and Emergency Management Authority (AFAD), “The 2023 Kahramanmaraş, Turkey, earthquake sequence,” 2023. Accessed: Apr. 30, 2023. [Online]. Available: https://en.afad.gov.tr/press-bulletin-36-about-the-earthquake-in-kahramanmaras

[60] C. Robinson et al., “Rapid building damage assessment workflow: An implementation for the 2023 rolling fork, mississippi tornado event,” 2023, *arXiv:2306.12589*.

**Sesa Wiguna** received the B.Sc. degree in geography from the University of Indonesia, Jawa Barat, Indonesia, in 2012 and the M.Sc. degree in geography from the University of Auckland, Auckland, New Zealand, in 2019. He is currently working toward the Doctoral degree in civil and environmental engineering with the Graduate School of Engineering, Tohoku University, Sendai, Japan.

Since 2014, he also works for the National Disaster Management Authority of Indonesia or Badan Nasional Penanggulangan Bencana (BNPB) as a Disaster Risk Analyst. His research interests include the use of deep learning and Earth observation technology for disaster management, particularly supporting rapid building damage mapping and GIS for disaster risk assessment.

**Bruno Adriano** (Senior Member, IEEE) received the M.Eng. degree in disaster management from the National Graduate Institute for Policy Studies, Tokyo, Japan, in 2010 and the Ph.D. degree in civil and environmental engineering from the Department of Civil and Environmental Engineering, Graduate School of Engineering, Tohoku University, Sendai, Japan, in 2016.

From 2016 to 2018, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS), International Research Institute of Disaster Science (IRIDeS), Tohoku University. From 2018 to 2023, he was a Research Scientist with the Geoinformatics Unit, RIKEN Center for Advanced Intelligence Project. Since 2023, he has been an Associate Professor with IRIDeS, Tohoku University. His research interests include Earth observation, machine learning, and high-performance computer simulation technologies with applications to disaster management and environmental monitoring.

**Erick Mas** was born in Lima, Peru. He received the Graduate degree in civil engineering from the Faculty of Civil Engineering, National University of Engineering (UNI), Lima, Peru, in 2004, and the master's degree in disaster risk management from UNI, and the Ph.D. degree in civil and environmental engineering from the Graduate School of Engineering, Tohoku University, Sendai, Japan, in 2009 and 2012, respectively.

He has professional experience in disaster risk management with regional and local governments in Peru. He is currently an Associate Professor with the International Research Institute of Disaster Science (IRIDeS) and has been appointed to the Tough Cyberphysical AI Research Center, the Co-Creation Center for Disaster Resilience, and the Core Research Cluster of Disaster Science, Tohoku University. He is a Technical Advisor to the RTi-cast firm. His research interests include agent-based modeling, tsunami risk reduction, evacuation simulation, and geoinformatics for disaster science.

**Shunichi Koshimura** received the Ph.D. degree in civil engineering from the Graduate School of Engineering, Tohoku University, Sendai, Japan, in 2000.

He is currently a Professor with the International Research Institute of Disaster Science, Tohoku University. He is also a co-founder of RTi-cast, a technology firm to offer real-time tsunami inundation damage forecast services to government organizations and commercial clients. His research interests include developing a real-time natural hazard forecast system and on estimating social impacts by integrating numerical modeling, Earth observation, and geo-informatics.