# Building Damage Mapping of the 2024 Noto Peninsula Earthquake, Japan, Using Semi-Supervised Learning and VHR Optical Imagery

Sesa Wiguna, Bruno Adriano, *Senior Member, IEEE*, Ruben Vescovo, Erick Mas, Ayumu Mizutani, and Shunichi Koshimura

***Abstract*— Deep learning (DL) models are generally less able to maintain their performance in out-of-domain (OOD) testing. Model transferability is crucial, especially when a model needs to be applied to a new dataset, such as in disaster emergency response, where the training samples are scarce. To solve the aforementioned issues, we propose a semi-supervised framework to improve model generalization using unlabeled samples from the target domain. The framework consists of two main steps: model initialization, which incorporates past events, and iterative fine-tuning. The latter step relies heavily on the pseudolabels inferred with high confidence from the former step. We tested our framework on the 2024 Noto Peninsula Earthquake. Our framework shows an improvement in model generalization indicated by higher scores in the tuned model compared with the initial model. The effect is even greater when the local context from the past event is included in the initial learning step. In this case, the score has increased by about 21% from 0.62 to 0.75. The proposed framework offers a promising solution for rapid disaster damage mapping.**

***Index Terms*— Building damage recognition, disaster resilience, emergency response, remote sensing (RS), semi-supervised.**

## I. Introduction

REMOTE sensing (RS) and deep learning (DL) have been widely studied in many applications, including building damage detection. Generally, model evaluation can be done in two streams: in-domain (IND) and out-of-domain (OOD). IND testing is undertaken by splitting samples into a training and test set. In this scheme, training and testing are drawn from the same distribution. In the OOD scheme, samples are unseen before and usually drawn from different distributions. Regarding these testing schemes, IND generally achieves a higher score as the models are already familiar with the distribution. For some applications, such as disaster emergency response, the OOD satisfies the real-world application.

In the aftermath of a disaster, training samples are unavailable, and data collection is time-consuming. Meanwhile, the damage information is required immediately to help the emergency response efforts. One possible approach would be to use a model trained on a big dataset and use it for inference in the new disaster-affected area [1]. In this setting, the testing is OOD, as the new disaster location will likely differ from the one the model was trained on. Previous studies (e.g., [2], [3], [4]), however, have reported a drop in performance when their models are applied to unseen disaster events.

In the lack of training samples, some approaches have been proposed to improve model generalizations, including domain adaptation, fine-tuning, and semi-/un-supervised learning. Kellenberger et al. [5] use various domain adaptation techniques to improve the model's generalization in land cover classification. Yang et al. [6] reported an improvement in model generalization by fine-tuning DL trained with satellite images with few samples from aerial photographs.

Semi-supervised learning frameworks use labeled and unlabeled data in the training loop. One of the approaches is to use a model trained on labeled data to predict label probability from the unlabeled target samples. The unlabeled data with a high probability (pseudolabels) are used in the training loop, e.g., via fine-tuning or retraining the model. Previous studies (e.g., [7], [8], [9]) have shown the usefulness of using these high-probability samples. In Saha et al. [9], a pseudolabel of a target data is used in incremental training schemes to align conditional distributions across domains. In Paul and Ganju [7], pseudolabels were generated using ensemble models and assimilated with original training data as input for the next round of training ensemble models. Similarly, Babakhin et al. [8] use multiple round training using assimilated pseudolabel samples and training data. The second-last models were used as an ensemble model to infer the final prediction.

Manuscript received 3 April 2024; revised 24 May 2024; accepted 27 May 2024. Date of publication 31 May 2024; date of current version 10 June 2024. This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI), under Grant 21H05001, Grant 22K21372, and Grant 22H01741; in part by the Ministry of Education, Culture, Sports, Science and Technology (MEXT); in part by the Co-Creation Center for Disaster Resilience, Tohoku University; and in part by the Cross Ministerial Strategic Innovation Promotion Program under Grant JPJ012289. *(Corresponding author: Shunichi Koshimura.)*

Sesa Wiguna and Ruben Vescovo are with the Department of Civil and Environmental Engineering, Tohoku University, Sendai, Miyagi 980-8579, Japan (e-mail: wiguna.sesa.p1@dc.tohoku.ac.jp; vescovo.ruben.t7@dc.tohoku.ac.jp).

Bruno Adriano, Erick Mas, Ayumu Mizutani, and Shunichi Koshimura are with the Disaster Geo-informatics Laboratory, International Research Institute of Disaster Science (IRIDeS), Tohoku University, Sendai 980-8572, Japan (e-mail: adriano@irides.tohoku.ac.jp; mas@irides.tohoku.ac.jp; ayumu.mizutani.c1@tohoku.ac.jp; koshimura@irides.tohoku.ac.jp).

Digital Object Identifier 10.1109/LGRS.2024.3407725

In this study, rather than using the pseudolabel information to train a model from scratch, we use it only for fine-tuning. This approach results in inexpensive computation. The contributions of this study are summarized as follows.

1) We propose a semi-supervised framework using unlabeled data to improve model generalizations in unseen data to contribute to the emergency response phase. We applied our approach to the 2024 Noto Peninsula Earthquake in Japan.
2) We evaluate the performance of a model trained only with the global dataset and a model trained with an additional local context from the same region.
3) We examine the classification results with higher resolution aerial photographs to understand the advantages and limitations of the model.

## II. Methodology

### A. Problem Settings

The study aims to use RS and DL technologies to detect building damage to support emergency response efforts. In disaster emergency settings, postdisaster imagery of the target area may be provided, but there is a lack of labeled data. However, we can access past event data, including imagery and corresponding damage labels. The task is then to predict the damage using this data availability. In other words, given the source domain dataset $D_s$ comprising satellite imagery $X_s$ and its corresponding damage label $Y_s$, we aim to predict a label $Y_t$ to the target domain $D_t$ imagery $X_t$.

For the initial rapid damage assessment purposes, we focus on two damage classes, namely, no-damage and damage. We perceive the problem as image classification, in which each building image will be assigned a class of either no-damage or damage.

As for the input images, we use multitemporal very-high-resolution (VHR) optical imagery acquired before and after the disaster. We operate under the assumption of perpetual access to predisaster images, sourced from platforms such as Google Satellite or Bing Satellite.

### B. Semi-Supervised Framework

In this study, we propose a semi-supervised framework using unlabeled data to help a model trained in the source domain to generalize in unseen data of the target domain (Fig. 1). The framework comprises two main steps: model initialization and iterative fine-tuning. Model initialization refers to training DL models with a large dataset. The second stage is the fine-tuning process using target domain unlabeled data. Since the tuning requires labels, we generate a pseudolabel $\hat{y}_t$ according to the damage probability inferred by the trained model. We discard all the inputs that score within a predefined threshold using Softmax (Con-1).

We iterate the tuning process multiple rounds where the predicted probability of each last tuned model is used for the next round of tuning. Note that the fine-tuned model is only used for inference. The tuning stops when the total pseudolabel is less than that of the previous round (Con-2).

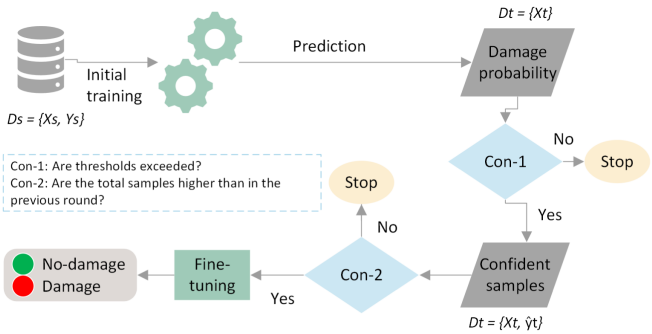


Fig. 1. Illustration of the semi-supervised iterative tuning framework. Past event dataset is fed to the initial training. The trained model is used to obtain damage probability of the 2024 Noto Peninsula Earthquake.

The proposed framework was tested using a global dataset (xBD) [10] as a source domain ($X_s$, $Y_s$) and the recent 2024 Noto Peninsula Earthquake as a target domain $X_t$. Note that the 2024 Noto Peninsula Earthquake is only used for testing, so the model has not seen it during training. We also experiment with adding samples to the xBD, namely, the 2011 Great East Japan Earthquake and Tsunami (hereafter, 2011 Tohoku Tsunami). Specifically, we evaluate the model performance between training only with xBD and a combination of xBD and the 2011 Tohoku Tsunami. Since the xBD does not include samples from Japan, it may be beneficial to introduce the model with samples from the same region as the target area. This way, the model will have prior knowledge of the local context of the target domain, for example, its building characteristics. Details about the data used are described in Sections III-A and III-B.

## III. Experiments and Results

### A. Data Descriptions

This section explains the data used in this study. It includes xBD and the 2011 Tohoku Tsunami dataset, which was used for training, and the 2024 Noto Peninsula Earthquake as the target area of a site for testing.

*1) xBD Dataset:* The xBD dataset is initially released for xView-2 Challenge [10]. It is a collection of 19 disasters from around the globe from various disaster types, including bushfire, fire, wildfire, volcanic eruption, earthquake, flood, tsunami, hurricane, and tornadoes. The data comprise VHR satellite imagery acquired before and after the disasters. Each images have a ground sampling distance (GSD) of $0.5 \times 0.5$ m. Besides the imagery, the dataset is also provided with building footprints with the corresponding damage of four categories, namely, no-damage, minor-damage, major-damage, and destroyed. Since the current study focuses on binary classes, we grouped minor-, major-damage, and destroyed into the damage class. Meanwhile, we keep the no-damage class as it is.

*2) 2011 Tohoku Tsunami Dataset:* On March 11, 2011, a massive tsunami hit the eastern coast of Japan. The tsunami caused numerous deaths, missing, and massive damage to infrastructure. This study focuses on the Sendai City area.

The dataset for the study comprises a pair of WorldView3 imagery captured before (10 August 2010) and after

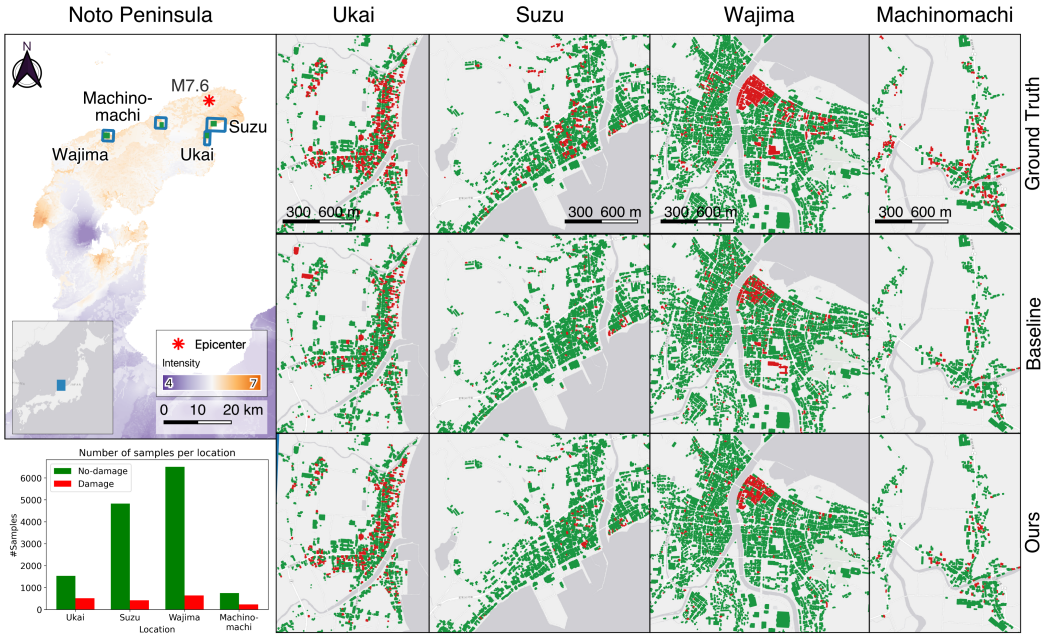


Fig. 2. (Top left) Epicenter and the intensity of the earthquake. The intensity is obtained from Quiet+ (https://app.quietplus.kke.co.jp/quakes?mode=quick). (Bottom left) Number of samples per location. (Right) Ground truth and damage distribution predicted by the baseline (direct testing using an xBD pretrained model) and our proposed framework. The predictions are generated using xBD + Tohoku training as the initial model. The graph in the bottom left shows the number of samples per class in the study area.

(11 March 2011) the tsunami and damage label. The images have a GSD of 0.5 × 0.5 m. As for the building damage label, we used a building polygon from the Ministry of Land, Infrastructure, Transport and Tourism of Japan (MLIT) with the corresponding damage labels based on a field survey [11]. The data are available at http://fukkou.csis.u-tokyo.ac.jp/dataset/list_all.

MLIT categorized the damage classes into seven classes, namely, no-damage, minor-damage, moderate-damage, major-damage, complete damage, collapsed, and washed-away. To meet the setting of the current work, we regrouped those classes based on the damaged appearance in the images. Specifically, we randomly selected samples from each sample of MLIT and checked the damage states of those samples. We then recategorized those samples into no-damage and damage classes.

### *B. 2024 Noto Peninsula Earthquake*

On 1 January 2024, a major earthquake of M7.6 hit the Noto Peninsula in Ishikawa Prefecture, Japan. As of March 12, the prefectural government confirmed 241 deaths in Ishikawa prefecture. The disaster also injured 1188 people, displaced 15 920 families, and damaged over 80 000 houses and buildings (https://www.pref.ishikawa.lg.jp/saigai/documents/higaihou_108_0312_1400.pdf).

The earthquake has triggered collateral hazards, such as tsunamis, landslides, and fires. To explore different types of damage, the study focuses on four areas: Ukai fishing port (hereafter, Ukai), Center of Suzu city (hereafter, Suzu), Center of Wajima city (hereafter, Wajima), and Machinomachi in Wajima city (herafter, Machinomachi). Each city has a unique damaging factor. Ukai and Suzu were affected by both the tsunami and strong motion. In Wajima, fires damaged many buildings. In Machinomachi, the earthquake was the main cause of the damage.

Postdisaster aerial photographs were taken by the Geospatial Information Authority of Japan (GSI) a few days after the disaster (https://www.gsi.go.jp/BOUSAI/20240101_noto_earthquake.html). For Wajima city, the image acquisition date is 11 January 2024, while for other areas, the images were captured on 2 January 2024. For some areas, oblique photographs are available from Kokusaki Kougyou (https://bois-free.bousai.genavis.jp/diarsweb). As for the predisaster image, we used Google Satellite imagery. Both the images have approximately 0.2 × 0.2 m. of GSD. To validate our model, we generated a damage label through visual interpretation by comparing the before- and after-disaster images. We classify the damage into two classes according to the structural conditions visible from the aerial photographs. No-damage refers to buildings with undisturbed structures. In contrast, buildings that are structurally disturbed are categorized as damage. The total samples of each studied city are illustrated in Fig. 2. Note that the dataset from Noto event is not included in the training.

### *C. Training Parameters*

In our experiments, we use the Shifted Windows (Swin) Transformer model [12] as the classification model. SwinT was demonstrated consistently superior to other architectures in a previous investigation [1]. Nonetheless, we performed additional tests using a ResNet backbone. Ultimately, SwinT again outperformed other architectures. Details of each architecture are outlined in our previous work.

We designed a Siamese structure comprising two identical encoders to meet multitemporal input images. One encoder is responsible for extracting features from predisaster images, and the other for postevent images. The features extracted

from each encoder are concatenated before being fed to the additional linear layers for classification.

We used a pretrained model to achieve better convergence and accuracy. The model was pretrained on ImageNet. The model was then retrained on the xBD dataset with a batch size of 64 for 30 epochs. We used BinaryCrossEntropy (BCE) as a loss function. An initial learning rate (LR) of $1 \times 10^{-4}$ was set. Cosine annealing LR with a minimum LR of $1 \times 10^{-6}$ was used as LR Scheduler. AdamW was used as an optimizer algorithm. To tackle randomness in the experiments, we used a random seed number of 100. Data augmentation techniques, including rotation, resizing, color transformation, and noise transformation, were used to deal with overfitting and increase the variability of models. All the experiments were implemented using the PyTorch framework on an NVIDIA Quadro RTX 6000 GPU machine.

For fine-tuning, we freeze all the blocks except the linear layers and the last Swin Transformer block. All other training settings were kept the same as in the initial training.

To evaluate model performance, we use $F_1$. It considers both precision and recall, which makes it suitable for imbalanced datasets.

### D. Results

This section evaluates the proposed method for predicting damage in a new disaster event. To assess its effectiveness, we compare the results with a baseline model (xBD pretrained model without fine-tuning).

For the fine-tuning process, we experiment with six different thresholds to obtain the pseudolabel. Each threshold determines the pseudolabel's confidence level, which ranges from lowest to highest. Each threshold value is randomly set based on its strictness in determining the confidence level, affecting the number of pseudolabels, hence the model performance. These thresholds range from loose to strict. Details on the values per each threshold, as well as the average $F_1$, are summarized in Fig. 3. The figure shows that the highest score is obtained by threshold#5 (0.01 & 0.90 for no-damage and damage, respectively). Thus, the results presented in this section are calculated based on threshold#5. The performance evaluation is reported in Table I, and the prediction maps are depicted in Fig. 2.

The table shows that overall, our approach improves the generalization, indicated by a higher score for iterative tuning compared with the baseline. In the xBD input model, our approach increases the overall score by 7% from 0.57 to 0.61. All the areas also show improvements. For instance, for Machinomachi, the tuning approach has improved the score by 20% from 0.50 to 0.60.

The impact is even greater in cases where the 2011 Tohoku Earthquake data are included in the initial training. Overall, our method has increased the overall score by 21% from 0.62 to 0.75, with a significant improvement in the damage class with an increase of 67% from 0.33 to 0.55. In this case, all the areas experience significant improvements. It appears that combining the past events from the local area and our approach gives significant generalization ability of the model. Introducing a local context may have helped models learn the pattern better since the model is familiar with local characteristics such as building styles.

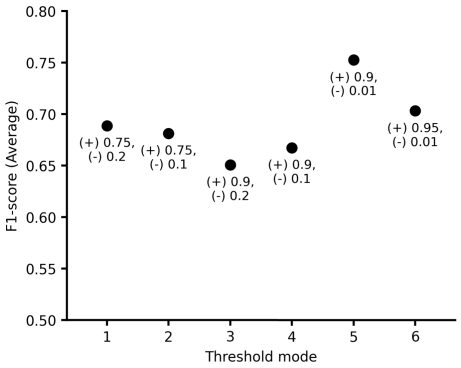


Fig. 3. Ablation study with different threshold values. (−) and (+) indicates the threshold for no-damage and damage class, respectively.

TABLE I
ACCURACY ASSESSMENT COMPARING THE BASELINE AND ITERATIVE TUNING

| Area | Mode | Input | $F_1$ | | Mean |
|---|---|---|---|---|---|
| | | | No-dmg | Damage | |
| Ukai | Baseline | 1 | **0.81** | 0.30 | 0.56 |
| | Ours | 1 | 0.78 | **0.43** | **0.60** |
| | Baseline | 2 | 0.84 | 0.41 | 0.62 |
| | Ours | 2 | **0.91** | **0.67** | **0.79** |
| Suzu | Baseline | 1 | **0.92** | 0.13 | 0.52 |
| | Ours | 1 | 0.89 | **0.16** | 0.52 |
| | Baseline | 2 | 0.94 | 0.18 | 0.56 |
| | Ours | 2 | **0.96** | **0.34** | **0.65** |
| Wajima | Baseline | 1 | **0.94** | 0.31 | 0.62 |
| | Ours | 1 | 0.93 | **0.38** | **0.66** |
| | Baseline | 2 | 0.94 | 0.38 | 0.66 |
| | Ours | 2 | **0.97** | **0.58** | **0.77** |
| Machino-machi | Baseline | 1 | 0.86 | 0.14 | 0.50 |
| | Ours | 1 | **0.88** | **0.31** | **0.60** |
| | Baseline | 2 | 0.87 | 0.34 | 0.60 |
| | Ours | 2 | **0.90** | **0.50** | **0.70** |
| All area | Baseline | 1 | **0.91** | 0.24 | 0.57 |
| | Ours | 1 | 0.90 | **0.32** | **0.61** |
| | Baseline | 2 | 0.92 | 0.33 | 0.62 |
| | Ours | 2 | **0.95** | **0.55** | **0.75** |

Model input: 1 = xBD; 2 = xBD+Tohoku

The classification performance is illustrated in Figs. 4 and 5. In general, our model can improve damage detection in collapsed buildings. For example, in Case #1, the postdisaster image shows that the building has collapsed, as indicated by the flattened roof in the postdisaster image. The oblique image also verifies this. The baseline model, however, predicted it as no-damage. After the fine-tuning process, the model correctly classified it as damage.

In Case #2, the building is damaged, indicated by a structural change where the building has rotated. Meanwhile, no significant changes appear in the roof. For this instance, the baseline and our model misclassified it as no-damage. The two cases indicated that our model can enhance damage detection when there is an alteration in the roof. However, buildings with damaged structures with no significant changes in the roof tend to be classified as no-damage. This may be related to the use of satellite imagery in the xBD. The nature of the nadir view of the images tends to learn the damage states from the top, in this case, building roofs.

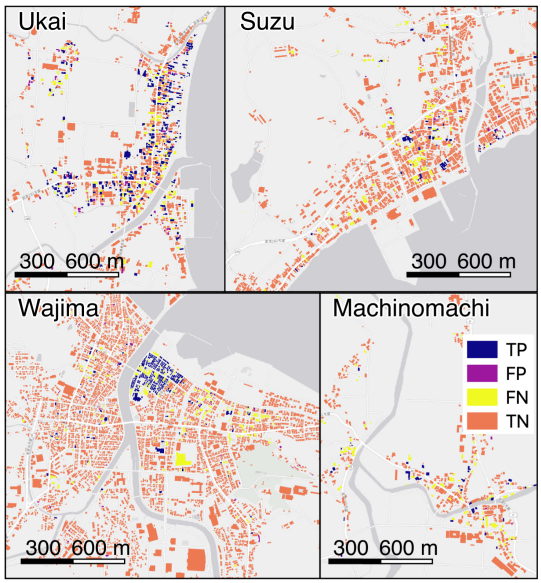


Fig. 4. Classification results calculated for the damage class. TP: True positive, FP: False positive, FN: False negative, TN: True negative.

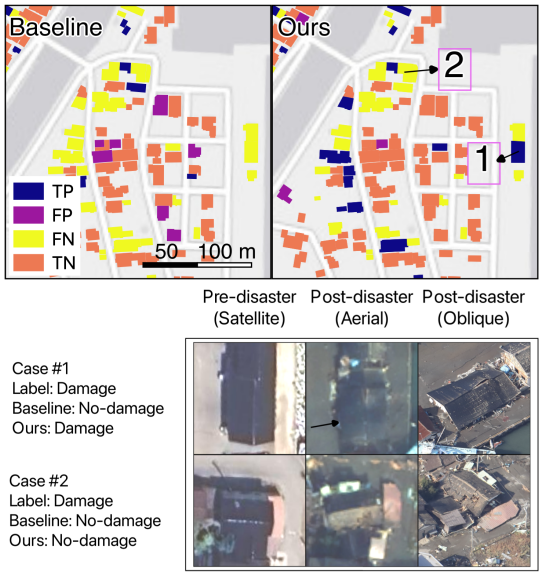


Fig. 5. Classification results calculated for the damage class in Ukai City. Case #1 illustrates that our model improves damage detection in collapsed buildings by looking at the significant change in the roof (e.g., flattened in the postdisaster image). Case #2 shows that damaged buildings with no significant changes in the roof tend to be classified as no-damage.

In our experiments, training took approximately 7 h. Once the initial model and the target data were available, each round of tuning and inference took approximately 20 min. In addition, since past data are s available in most cases, it is possible to train a model to prepare for a possible upcoming disaster. During the emergency response, the trained model can be directly used for inference and tuning processes. Due to its generalization ability and speed in the process, the proposed approach can be promising in helping disaster emergency operations.

## IV. Conclusion

In this study, we propose a semi-supervised iterative tuning process using unlabeled data of the target area to help map the building damage in emergency settings. The benefit of our proposed method is that it operates with unlabeled samples, which is often the case in disaster emergencies. Meanwhile, since the model uses fine-tuning methods, this approach is inexpensive in computation. We initiate a big model by training a damage classification model with a global dataset (xBD). As a part of the experiment, we also add damage dataset to the global data. We used the trained models to predict unseen data. This inference process results in a damage probability. We filter the samples of the testing location with high confidence and assign a pseudolabel to each sample based on the probability. We use these samples to fine-tune the initial model.

We tested the framework on the recent 2024 Noto Peninsula Earthquake. Our framework shows an improvement in the model's generalization. The effect is even greater when the similar building context from the past event is included in the initial learning step. In this case, the score increased by about 21% from 0.62 to 0.75. We believe that the framework offers a promising solution for rapid disaster damage mapping. As a final note, the proposed approach should not be seen as a competitor to supervised and unsupervised methods, rather as a complementary to them.

## References

[1] S. Wiguna, B. Adriano, E. Mas, and S. Koshimura, "Evaluation of deep learning models for building damage mapping in emergency response settings," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 5651–5667, 2024.

[2] 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, p. 2839, Sep. 2020.

[3] 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. Photogramm. Remote Sens.*, vol. 140, pp. 45–59, Jun. 2018.

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

[5] B. Kellenberger, O. Tasar, B. B. Damodaran, N. Courty, and D. Tuia, "Deep domain adaptation in Earth observation," in *Deep Learning for the Earth Sciences*. Wiley, 2021, ch. 7, pp. 90–104. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119646181.ch7

[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, p. 504, Jan. 2021.

[7] S. Paul and S. Ganju, "Flood segmentation on Sentinel-1 SAR imagery with semi-supervised learning," 2021, *arXiv:2107.08369*.

[8] Y. Babakhin, A. Sanakoyeu, and H. Kitamura, "Semi-supervised segmentation of salt bodies in seismic images using an ensemble of convolutional neural networks," in *Pattern Recognition*, G. A. Fink, S. Frintrop, and X. Jiang, Eds. Cham, Switzerland: Springer, 2019, pp. 218–231. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-33676-9_15#citeas

[9] S. Saha, S. Zhao, and X. X. Zhu, "Multitarget domain adaptation for remote sensing classification using graph neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

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

[11] Ministry of Land, Infrastructure Transport and Tourism (MLIT). (2011). *Summary of the Investigation on Reconstruction Methods for Tsunami-Damaged Areas After the Great East Japan Earthquake and Tsunami (in Japanese)*. Accessed: May 19, 2023. [Online]. Available: https://www.mlit.go.jp/toshi/toshi-hukkou-arkaibu.html

[12] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.