Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Constructing an Extensible Building Damage Dataset via Semi-supervised Fine-Tuning across 12 Natural Disasters

Zeyu Wang[1,2], Chuyi Wu[1,2], Feng Zhang[1,2*], and Junshi Xia[3]

[1]School of Earth Sciences, Zhejiang University, Hangzhou 310027, China. [2]Zhejiang Provincial Key Laboratory of Geographic Information Science, Hangzhou 310028, China. [3]Geoinformatics Team, RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan.

*Address correspondence to: zfcarnation@zju.edu.cn

Citation: Wang Z, Wu C, Zhang F, Xia J. Constructing an Extensible Building Damage Dataset via Semi-supervised Fine-Tuning across 12 Natural Disasters. *J. Remote Sens.* 2025;5:Article 0733. https://doi.org/10.34133/remotesensing.0733

Submitted 8 July 2024
Revised 17 June 2025
Accepted 18 June 2025
Published 7 August 2025

Post-disaster building damage assessment (BDA) is vital for emergency response. Deep learning (DL) models are increasingly being applied to achieve quick and automatic BDA on disaster remote sensing imagery, and their performance largely relies on the knowledge base offered by the dataset. However, constructing a BDA dataset requires intensive expert labeling work and a massive time, leading to a substantial lag in dataset enrichment and model development in the current research field. To address this, this paper introduces a new multidisaster BDA benchmark, the extensible building damage (EBD) dataset, which includes over 18,000 pre- and post-disaster image pairs from 12 recent disaster events, covering over 175,000 building annotations with 4-level damage labels. Unlike previous BDA datasets, EBD follows a semiautomatic labeling workflow and has reduced construction time by 80% compared to full manual labeling. In this process, the DL model served as the machine expert to perform automatic labeling. It was pretrained on the xView2 building damage dataset and then transferred to each new disaster scenario via semi-supervised fine-tuning (SS-FT). SS-FT not only leverages a few labeled samples for supervised fine-tuning but also incorporates both labeled and unlabeled samples into pixel-level contrastive learning. Results demonstrate that the DL model has considerably improved annotation performance under SS-FT. A series of analyses have proven EBD's building damage feature diversity, practical value in emergency mapping, and knowledge enrichment to the existing benchmark. EBD advances data renewal for natural disaster scenarios and supports the application of artificial intelligence in emergency response efforts.

Introduction

Every year, natural disasters have resulted in an increasing loss of life and infrastructure. Developing countries are more exposed to hazards due to the lack of technology reserves [1]. During disaster response, building collapse is one of the main causes of human casualties [2] and is a strong indicator of hazard information [3]. Compared to field research, remote sensing (RS) can provide in-time and large-range observation of building destruction when a natural disaster occurs [4,5].

Despite the abundant raw images, the damage information is sparse in the overall environment. Consequently, RS samples with fine-labeled damage information are valuable prior knowledge for disaster response. Table 1 shows existing open-source building damage datasets for disaster emergency response tasks. Notably, the xView2 building damage (xBD) dataset [6] is the largest disaster damage database, containing expert-labeled images from 19 historical disaster events. Apart from xBD, other datasets generally focus on a specific disaster type or event and provide damage information within particular domains.

Advanced artificial intelligence technologies, e.g., deep learning (DL), have been applied to disaster RS image interpretation tasks to achieve quicker disaster response [7]. Specifically, automatic building damage assessment (BDA) can be posed as the comprehensive results of building contour segmentation and damage level classification. Given sample coverage and labeling quality, the xBD dataset [6] is widely accepted as the benchmark to train and verify the performance of BDA models. In recent years, many efforts have been made to enhance the BDA model's performance by the attempts to investigate different model structures, including full convolutional neural networks [8–10] and transformer block [11,12]. It should be noted, however, that the disaster database extension is evolving at a relatively slow speed. This restrains the BDA model from absorbing more disaster knowledge.

The major obstacle to disaster sample labeling is expert knowledge and the massive time required for manual annotation. To overcome this, a promising solution is to apply intelligent models to aid sample annotation, transferring knowledge from existing datasets to new scenarios. Compared to full manual supervision, the construction efficiency of some recent RS datasets has been largely promoted by utilizing active learning [13,14], DL models [15,16], and pretrained (PT) large models [17,18]. Then, the main problem is ensuring the model's annotation

Table 1. Summary of publicly available RS datasets for BDA task

Dataset	Task	Disaster type	Events	Image source	Temporal	Resolution	Size
xBD [6]	Segmentation	(Multiple)	19	Satellite	Pre and post	1,024 × 1,024	22,068
ABCD [48]	Classification	Tsunami	1	Satellite	Pre and post	128 × 128	22,171
[49]	Classification	Earthquake	2	Aerial	Post	88 × 88	17,281
ISBDA [50]	Object detection	Hurricane	1	Aerial	Post	–	1,030
[51]	Object detection	Hurricane	1	Satellite and aerial	Post	–	–
Multi3net [52]	Segmentation	Hurricane	1	Satellite	Pre and post	–	–
SpaceNet8 [53]	Segmentation	Flood	2	Satellite	Pre and post	1,300 × 1,300	1,602
FloodNet [54]	Segmentation	Flood	1	Aerial	Post	3,000 × 4,000	2,343
RescueNet [55]	Segmentation	Hurricane	1	Aerial	Post	3,000 × 4,000	4,494
EBD (ours)	Segmentation	(Multiple)	12	Satellite	Pre and post	512 × 512	18,215

quality in a new, out-of-training-set scenario. In the BDA task, especially, the PT model will face great domain gaps when applied to new disasters due to the differences in geographic contexts and building damage characteristics [19,20]. Considering these issues, historical knowledge and new data in the target scenarios should be collaboratively leveraged to assist in disaster sample labeling.

Transfer learning (TL) and SSL (SSL) are 2 practical approaches when the model's application scenario lacks labeled data. TL focuses on how to transfer existing knowledge to the target scenario [21], which has been researched to improve the generalization ability of existing models to new disaster images. Specifically, fine-tuning (FT) is a model-based TL strategy implemented by pretraining the model on a previous dataset and then fine-tuning it on a few task-specific labeled samples. Studies have explored the effectiveness of FT in transferring models between different disaster types [22], man-made and natural disasters [23], and different geographical environments [24]. Accordingly, FT is adopted as the baseline setting in our disaster sample annotation task.

Beyond relying on a limited set of annotated data for model FT, SSL is a crucial method for extracting information from unlabeled data. The basic approach is to construct an additional loss function (L_{unsup}) to learn implicit feature representations from unlabeled data. Previous change detection (CD) and BDA studies most applied consistency-based SSL methods to construct L_{unsup}, such as posing data augmentation [25,26] and feature perturbation operations [27] on unlabeled data. These consistency-based methods focus on the model's robustness to nonsemantic noise [28–30], but not the distinction between the characteristics of multiple categories. Another mainstream method of SSL is contrastive learning. Its main idea is to attract positive samples and repulse the negative samples among multiple categories. Practically, van den Oord et al. [31] proposed the Information Noise Contrastive Estimation Loss (InfoNCE) loss to measure the similarity of positive. More recently, contrastive-based SSL methods including Semi-seg [32] and U2PL [33] have achieved state-of-the-art performance in semantic segmentation tasks. Following this line of thought, we design a pixel-level contrastive loss between different damage levels to enhance model performance in severe category-imbalance disaster scenarios.

In summary, 2 underlying problems need to be addressed to construct a disaster dataset efficiently. (a) How to utilize intelligent methods to reduce the manual workload of disaster data labeling, given the advancements of DL models in performing automatic BDA. (b) How to design effective mechanisms to utilize labeled and unlabeled samples to enhance sample annotation, especially maintaining and leveraging features from different damage levels.

On the basis of the deficiencies identified in current research, this paper adopts a semiautomatic annotation workflow to construct a new multidisaster BDA benchmark, the extensible building damage (EBD) dataset. Unlike all previously published BDA datasets that relied on expert labeling, this study adopts a machine-driven labeling process that can significantly reduce manual work brought by dense annotation, as Fig. 1 illustrates. The machine annotator should undergo a pretraining stage and a disaster-specific optimization stage before performing automatic labeling. In the second stage, specifically, this study introduces a semi-supervised fine-tuning (SS-FT) method to optimize the annotation model for the pixel-level building damage classification task.

Overall, our contributions can be summarized as follows:

1. Data contribution of the EBD dataset: It includes over 18,000 bitemporal, very-high-resolution (VHR) optical images from 12 recent natural disaster events, containing 4-level damage annotation on over 175,000 buildings. This dataset can serve as both an independent BDA benchmark and a crucial complementary resource to existing disaster RS materials. A series of validation work has been performed in this study to prove the high quality and application potential of EBD.
2. Method contribution for low-cost yet high-quality disaster sample annotation: The DL model optimized by the SS-FT method has hugely reduced annotation time and human effort in EBD's construction. This solution offers a response to key challenges in automatic annotation for new disasters. One challenge is the difficulty in distinguishing multilevel damage characteristics at a fine-grained scale, and the other is the limited availability of labeled samples.

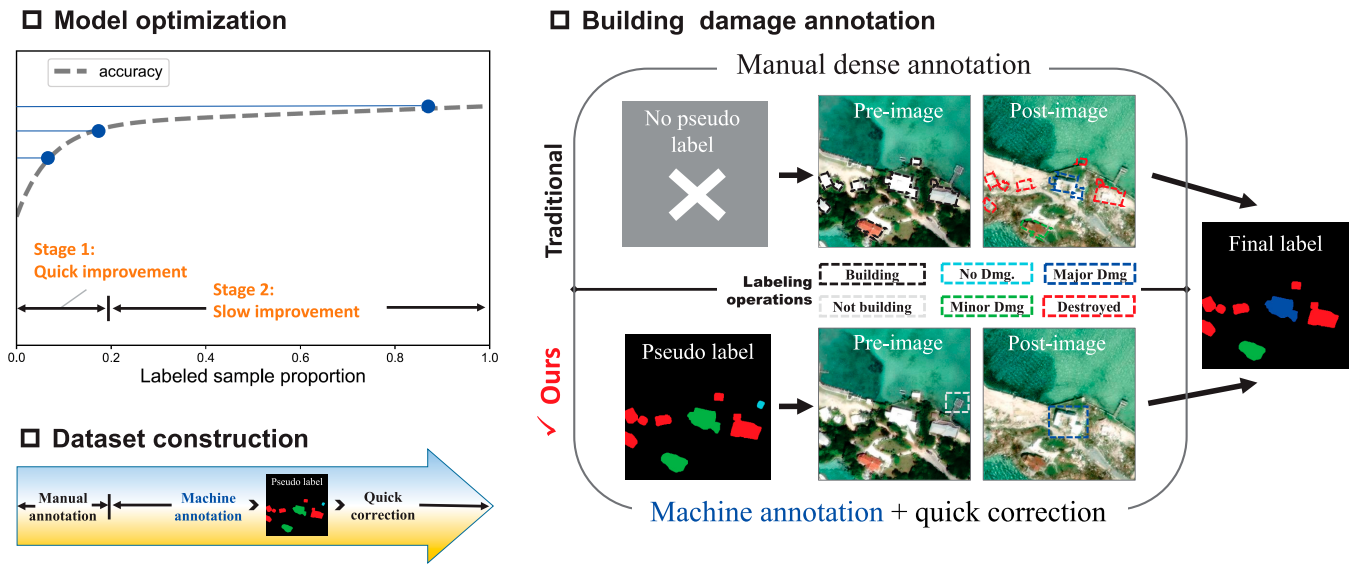

Fig. 1. Illustration of machine-driven dataset construction. With a small part of samples labeled manually, the PT model is optimized during the "quick improvement" stage and then performs automatic annotation for the remaining samples. The traditional way requires dense manual work in precise building contour annotation and damage level identification, while our machine-driven way can reduce it to quick corrections on pseudo labels. No Dmg, no damage; Major Dmg, major damage; Minor Dmg, minor damage.

Materials and Methods

Data sources

Materials for EBD were sourced from the Maxar Open-Data Program (https://www.maxar.com/open-data, accessed 2024 November 6), which provides near-real-time images of disaster-affected areas globally. The VHR optical images, with a spatial resolution of 0.3 to 0.5 m, were captured by the WorldView-3 satellite. As indicated in Table 2 and Fig. 2, EBD includes 12 disaster events that occurred around the globe and were not included in the xBD dataset. These events span a variety of disaster types, including hurricanes, tornadoes, volcanic eruptions, earthquakes, and flooding. Furthermore, all 12 events caused fatalities and economic losses in the affected regions, making them highly valuable for annotation as references for common local disasters.

For each event, we manually selected the affected region as the area of interest (AOI) for image filtering. To ensure the consistency of building position in pre- and post-disaster images, we performed georeferencing and resampling work on the bitemporal images. Then, the preprocessed RS images were cropped into 512 × 512 patch samples.

Damage scale criteria

In the EBD dataset, all post-disaster buildings are categorized into 4 damage levels (no damage, minor damage, major damage, and destroyed). Table 3 compares the criteria of each damage level in terms of different building damage characteristics, including structural, flooded, and volcanic-ash-covered. For one thing, these criteria generally follow the joint damage scale descriptions proposed in xBD [6] since xBD is regarded

Table 2. Disaster list of the EBD dataset

Disaster event	Center of AOI	Time	Building damage types
Hurricane Irma	93W, 31N	2017 Sep	Structural
Hurricane Dorian	79W, 26N	2019 Sep	Structural
Hurricane Laura	93W, 30N	2020 Aug	Structural
Hurricane Delta	65W, 18N	2020 Oct	Structural and flooded
Hurricane Ida	91W, 30N	2021 Oct	Structural and flooded
Hurricane Ian	82W, 27N	2022 Sep	Structural and flooded
Texas tornadoes	98W, 30N	2022 Mar	Structural
Pakistan flooding	68E, 26N	2022 Jul	Structural and flooded
Türkiye (Turkey) earthquake	37E, 37N	2023 Feb	Structural
Mount Semeru eruption	113E, 8S	2021 Dec	Ash-covered and structural
St. Vincent Volcano	61W, 13N	2021 Apr	Ash-covered
Tonga Volcano	175W, 20S	2022 Jan	Ash-covered

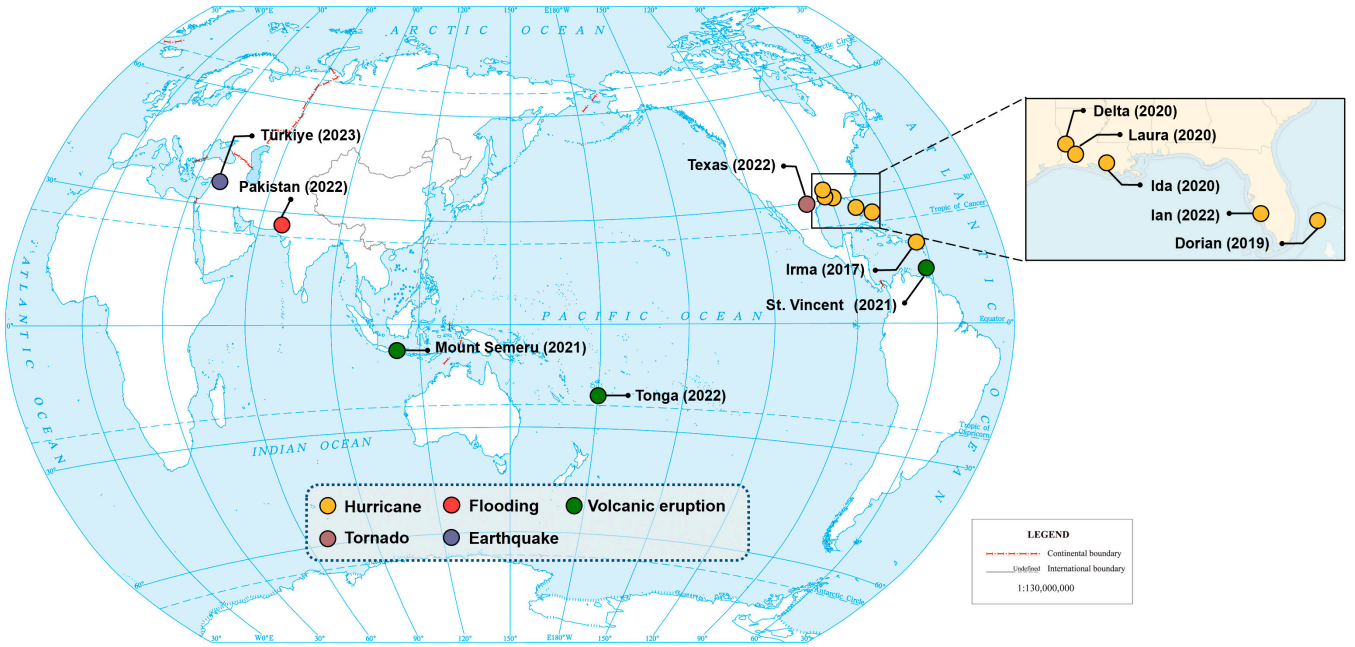

Fig. 2. Location of all disasters in the EBD dataset.

Table 3. Descriptions of the building appearance under different damage levels

Damage level		No damage	Minor damage	Major damage	Destroyed
Description	Structural	No sign of structural damage	Roof elements missing or visible cracks	Partial roof or wall collapsed	Completely collapsed or no longer present
	Flooded	Undisturbed by flood	Partly surrounded by flood	Completely surrounded by flood/mud	Covered with flood/mud or washed away
	Ash-covered	No sign of volcanic ash on the roof	Roof color changed with volcanic ash covered	Partial roof collapsed because of volcanic ash fall	Completely collapsed because of heavy volcanic ash fall
Damage scale in EMSR		No visible damage, possibly damaged	Damaged (minor)	Damaged (major)	Destroyed

as the prior knowledge of our EBD. For another, our EBD's criteria focus more on the hazards of inclusive disaster types and are also compatible with the Copernicus Emergency Management Service Rapid (EMSR) mapping's damage scale (https://emergency.copernicus.eu/mapping/ems/damage-assessment, accessed 2024 November 6).

Sample labeling workflow

Figure 3 shows the overall semiautomatic labeling workflow for samples of a new disaster. To achieve the balance between efficiency and accuracy, the annotation model is iteratively optimized with limited human supervision. The major steps of the workflow are as follows:

- (Step 1) Base model preparation and automatic labeling: Train a base annotation model on the xBD dataset by following the supervised setting. Then, bitemporal images are generated into patch samples to be labeled. For the first run, each sample will get the pseudo labels.
- (Step 2) Accuracy inspection: Randomly select a proportion of samples as the inspection set. The labels annotated by experts serve as the ground truth for calculating the accuracy of pseudo labels. If the accuracy exceeds the threshold, proceed to step 4 for sample postprocessing; otherwise, proceed to step 3 for model optimization.
- (Step 3) Model optimization by SS-FT: Manually select (or add) a proportion of samples as the FT set, which should provide abundant building damage information for the target disaster. These manual-labeled samples along with the remaining unlabeled samples are altogether applied to the model's semisupervised optimization process. Then, another round of automatic labeling is performed by the optimized model.

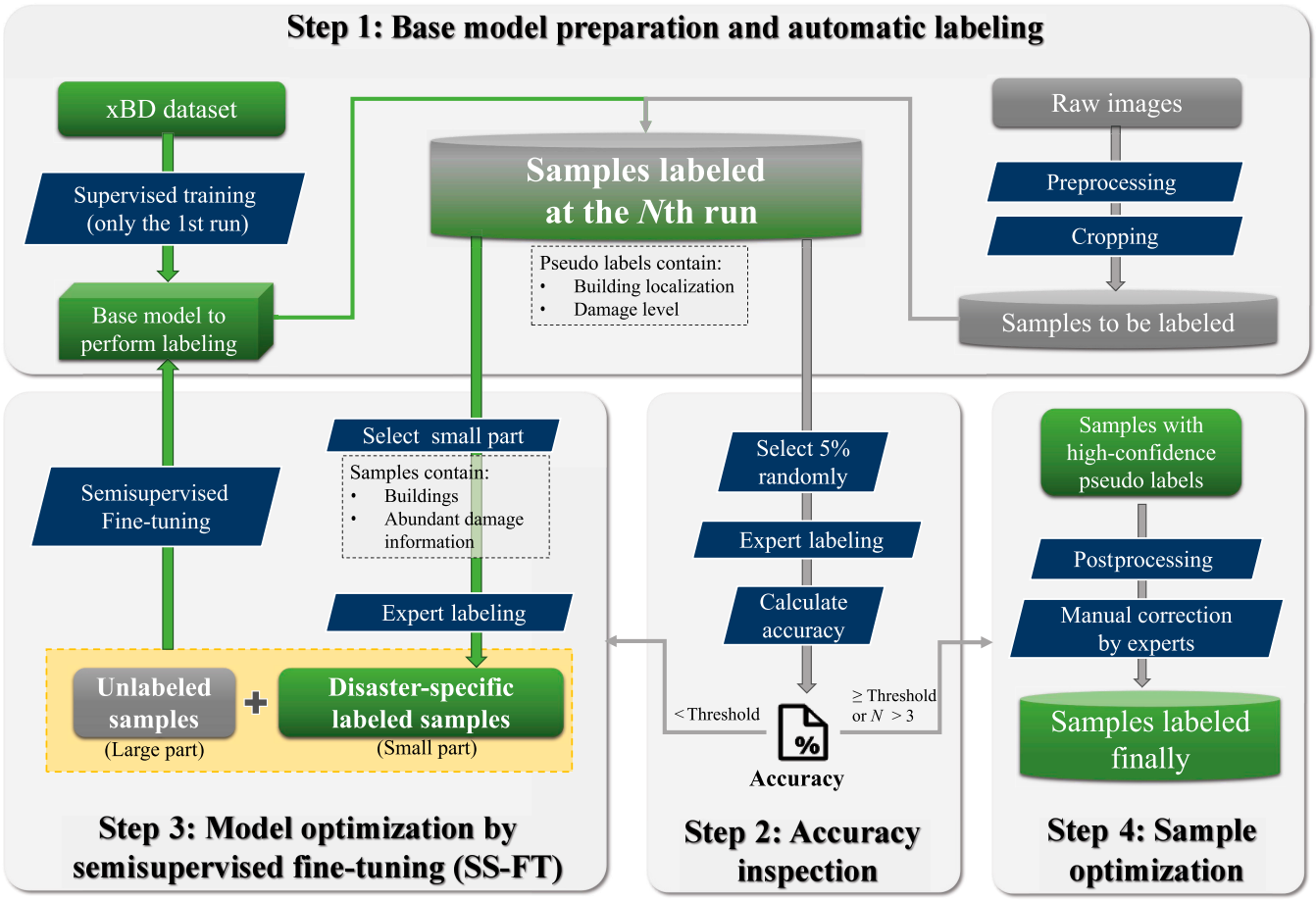

Fig. 3. The overall workflow of sample labeling.

- (Step 4) Sample postprocessing and optimization: Implement an object voting postprocessing work on pseudo labels. Last, experts perform a quick check for all samples and correct those mislabeled bad cases.

Considering the cost of manual labeling, steps 2 and 3 are iterated for a maximum of 3 rounds. After the generation of pseudo labels, the manual optimization work can compensate for the model's lack of knowledge. Human supervision is mainly introduced to the semiautomatic labeling workflow in 3 aspects: (a) manual labeling of the inspection set for calculating the accuracy of pseudo labels; (b) manual labeling of a small number of disaster-specific samples for model FT; and (c) postcheck for all samples and correction for bad cases. To avoid human errors, the manual annotation or correction for each single sample should be done by experts in pairs. Here, experts refer to people equipped with domain knowledge for disaster analysis.

Automatic annotation models

With bitemporal images, the basic idea of BDA is to compare the difference between pre- and post-disaster images of the affected area [34,35]. Following this idea, 2 annotation models have been proposed for sample automatic labeling. Notably, this paper develops the SS-FT method on the first model.

General annotation model

This model is pretrained on xBD and can be applied to different disaster scenarios with structural or flood-affected building damage. The overall model consists of 2 parts, localization network (F_{loc}) and classification network (F_{cls}). The optimization process is to establish the mapping functions from bitemporal images (x_{pre} and x_{post}) to the building footprints and damage levels (P_{building} and P_{damage}):

$$P_{\text{building}} = F_{\text{loc}}\left(x_{\text{pre}}; \theta_{\text{loc}}\right) \tag{1a}$$

$$P_{\text{damage}} = F_{\text{cls}}\left(x_{\text{pre}}, x_{\text{post}}; \theta_{\text{cls}}\right) \tag{1b}$$

where θ_{loc} and θ_{cls} represents the combined parameters of 2 networks. Here, F_{loc} is parameterized by a U-Net with the backbone of SE-ResNeXt50; F_{cls} is parameterized by 2 weight-sharing U-Net (named Siamese U-Net), each branch of which has the same structure as F_{loc} and initializes its parameters with θ_{loc}. Detailed information about model structure and training settings follows the work proposed by Wu et al. [8].

Object-level few-shot classification model

This model is specific to volcanic eruption scenarios. Since xBD only contains very limited volcanic eruption building samples, the labeling work of volcanic-ash-covered damage was implemented by a 2-stage model. In the first stage, all building contours are extracted; in the second stage, an object-level classification model is fine-tuned by few-shot samples and then implements prediction for remaining unclassified buildings. Model structure

details were introduced in our previous work [36], specifying the sample labeling of Mount Semeru Eruption, St. Vincent Volcano, and Tonga Volcano.

Evaluation metrics

$F1_{\text{loc}}$ and $F1_{\text{dam}}$ are adopted as the evaluation metrics to measure the model performance on building localization and damage classification tasks. In addition, on these 2 metrics, the accuracy of model-annotated pseudo labels is calculated by being compared with manually annotated ground truth. Formally, the $F1$ score is defined by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2a}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2b}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2c}$$

where TP, FP, and FN represent true positive, false positive, and false negative, respectively. As Eqs. (3a) and (3b) shows, $F1_{\text{loc}}$ is the normal $F1$ score for the binary classification task of building extraction; for the multilevel classification task, $F1_{C_i}$ is the $F1$ score for the building classification of damage level C, and $F1_{\text{dam}}$ is the harmonic mean of $F1_{C_i}$ from all damage levels.

$$F1_{\text{loc}} = F1 \tag{3a}$$

$$F1_{\text{dam}} = \frac{C}{\sum_{i=1}^{C} \frac{1}{F1_{C_i}}} \tag{3b}$$

SS-FT method

High-accuracy building damage identification is a challenging task with only a few manual-labeled samples. Accordingly, we proposed the SS-FT method to fully utilize both labeled and unlabeled samples during F_{cls} optimization. The target is to minimize the overall loss, which can be formulated as:

$$\mathcal{L} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{contra}}\mathcal{L}_{\text{contra}} \tag{4}$$

where $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{contra}}$ represent the supervised loss and contrastive loss respectively. Figure 4 illustrates the overall SS-FT method, with the following components:

- Teacher–student framework: The parameters of student model (θ^{stu}) is optimized by minimizing the total loss $\mathcal{L}$, while the parameters of teacher model (θ^{tea}) is iteratively updated with θ^{stu} by exponential moving average (EMA). For the training step n:

$$\theta_n^{\text{tea}} = \alpha\theta_{n-1}^{\text{tea}} + (1-\alpha)\theta_n^{\text{stu}} \tag{5}$$

where $\alpha \in [0,1]$ is a smoothing coefficient, usually set to 0.99.

- Supervised FT on labeled samples ($\mathcal{L}_{\text{sup}}$): On the few labeled samples, θ^{stu} is fine-tuned to predict the building damage in a specific disaster domain. We use cross-entropy (CE) loss as the supervised FT loss $\mathcal{L}_{\text{sup}}$, which is computed on the ground-truth y_i and the model prediction $\hat{y}_i$.

$$\mathcal{L}_{\text{sup}} = \text{CE}(\hat{y}_i, y_i) \tag{6}$$

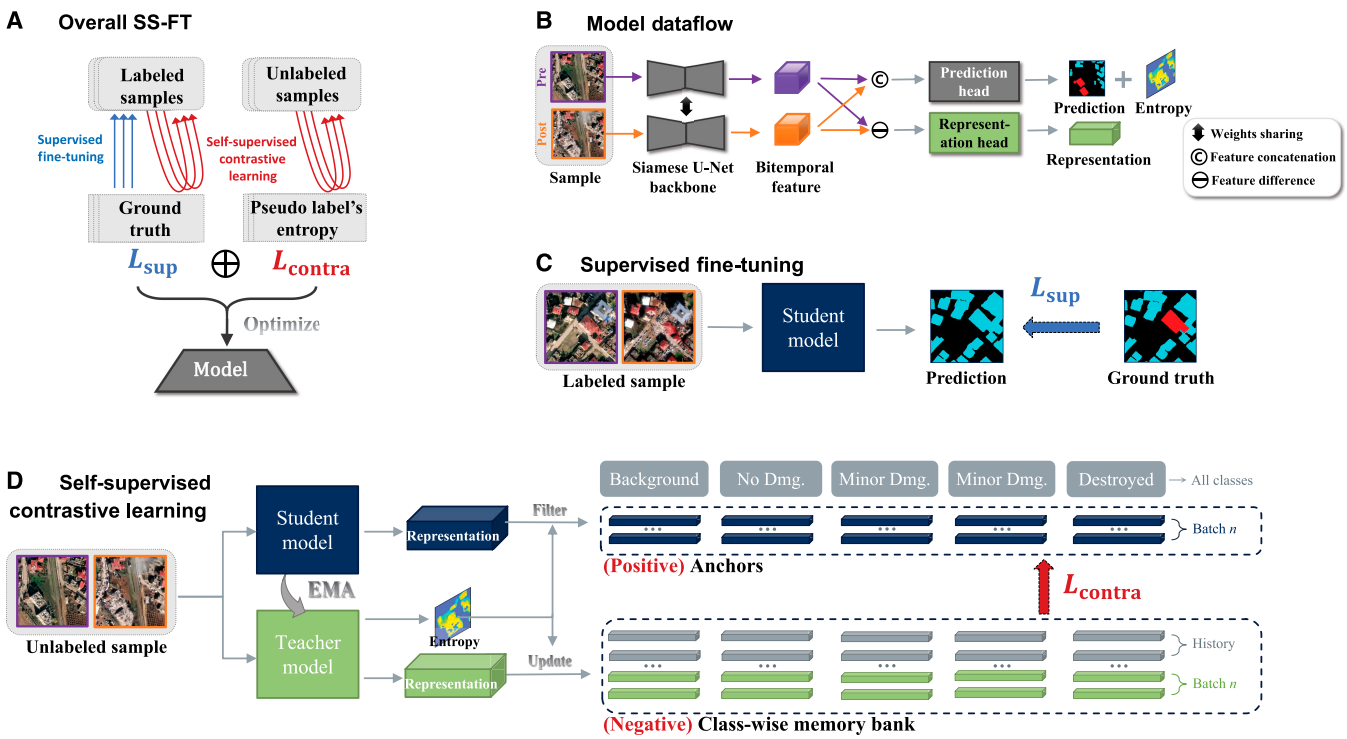

Fig. 4. Illustration of the proposed SS-FT framework. (A) and (B) show the overall SS-FT and the model's dataflow. (C) and (D) elaborate on the supervised FT and the self-supervised contrastive learning processes. For each minibatch, L_{contra} is calculated on positive "queries" and negative representations stored in the category-wise memory bank on a pixel level.

- Pixel-level contrastive learning ($\mathcal{L}_{\text{contra}}$): The contrastive loss further helps to discriminate the characteristics of each damage level. Our building damage classification network has 3 parts: (a) $f(\cdot)$, the Siamese U-Net backbone to extract features from bitemporal images; (b) $h(\cdot)$, the prediction head to give probabilities of each class; and (c) $g(\cdot)$, the representation head to give extra representation. As Eq. (7) shows, for the pixel j of ith sample, its softmax probabilities p_{ij} and representation z_{ij} are sourced from the prediction head and representation head, respectively. Then, the class with the highest probability is assigned as the pseudo label $\hat{y}_{ij}$, and the entropy of probabilities among all classes $\mathcal{H}\left(p_{ij}\right)$ is seen as the reliability of pseudo label.

$$p_{ij} = h \circ f\left(x_{ij}\right), z_{ij} = g \circ f\left(x_{ij}\right) \tag{7a}$$

$$\hat{y}_{ij} = \arg\max_c p_{ij}(c) \tag{7b}$$

$$\mathcal{H}\left(p_{ij}\right) = -\sum_{c=0}^{C-1} p_{ij}(c) \log p_{ij}(c) \tag{7c}$$

Following the recent popular contrastive learning paradigm in distinguishing representation, $\mathcal{L}_{\text{contra}}$ is set to the sum of pixel-level infoNCE losses [31] under all classes:

$$\mathcal{L}_{\text{contra}} = \sum_{c=0}^{C-1} \mathcal{L}^c_{\text{pixelNCE}} \tag{8}$$

where C is the number of classes. For each minibatch of images, a set of queries $\mathcal{Q}$ are filtered as typical representations under different classes, and a category-wise memory bank $\mathcal{N}$ is updated to store the corresponding negative representations. The whole process is shown in Algorithm 1. Then, the pixel-level infoNCE loss under class c is calculated among queries, positive samples, and negative samples:

$$\mathcal{L}_{\text{pixelNCE}} = -\sum_{i=1}^{M} \log \left[\frac{\exp\left(\left\langle \mathbf{q}_i, \mathbf{p}_i \right\rangle / \tau\right)}{\exp\left(\left\langle \mathbf{q}_i, \mathbf{p}_i \right\rangle / \tau\right) + \Sigma_{j=1}^{N} \exp\left(\left\langle \mathbf{q}_i, \mathbf{n}_{ij} \right\rangle / \tau\right)} \right] \tag{9}$$

where M is the total number of queries and $\langle \cdot, \cdot \rangle$ is the cosine similarity between 2 representations with the temperature τ; each query $\mathbf{q}_i$ is followed with a positive sample $\mathbf{p}_i$ calculated as the center of $\mathcal{Q}_c$, and N negative samples $\mathbf{n}_i$ randomly selected from $\mathcal{N}_c$. This loss setting is expected to maximize the distance between queries $\mathcal{Q}_c$ and negative samples $\mathcal{N}_c$ and also to minimize the intraclass distance of $\mathcal{Q}_c$. To ensure that the number of negative samples for each class is balanced, we use a queue for each category as the memory bank to store negative samples. It is worth noting that queries $\mathcal{Q}$ are sourced from the student model, which requires gradient backpropagation, and the memory bank $\mathcal{N}$ is updated by the teacher model.

Here is a further explanation of how to filter queries and negative samples for contrastive learning. If a pixel has the true label or reliable (low-entropy) pseudo label of class c, it will be filtered into queries as the typical representation. Likewise, a negative sample should not belong to class c according to its true label or pseudo label. In our settings, with the category probability order $\text{argsort}\left(p_{ij}\right)$ of an unlabeled pixel, class c should not be included in the first 2 positions. In addition, the prediction entropy of unlabeled negative samples should be relatively high, which implies the difficulty for the model to discriminate its category.

Object-wise voting postprocessing

Pixel-level pseudo labeling of building damage (D_{pxl}) can be obtained by model prediction. However, D_{pxl} may face semantic inconsistency within each building. To overcome the problem, an object-wise voting (OV) approach is applied during postprocessing work. First, each building proposal is marked as B_{obj} through a connected component labeling algorithm. Then, 2 thresholds $\alpha_{\min}$ and $\alpha_{\max}$ are set to filter the B_{obj} with different area sizes. As Eq. (10) shows, there are 3 situations: (a) the B_{obj} less than $\alpha_{\min}$ will be regarded as background; (b) the B_{obj} larger than $\alpha_{\max}$ will not be postprocessed since it might contain several adhesive buildings; and (b) otherwise, the B_{obj} will get a object-wise damage level through the weighted voting of all pixels within the area. Parameters are set as follows: $\alpha_{\max} = 8{,}000$, $\alpha_{\min} = 100$, $\left\{w_c | 1 \le c \le 4\right\} = \{1, 2, 1.5, 1.2\}$.

$$D_{\text{obj}} = \begin{cases} 0, & \text{if } B_{\text{obj}} < \alpha_{\min}; \\ D_{\text{pxl}}, & \text{else if } B_{\text{obj}} > \alpha_{\max}; \\ \arg\max_c \left(w_c \cdot \dfrac{\text{count}\left(D^c_{\text{pxl}}\right)}{\text{count}\left(B_{\text{obj}}\right)} \right), & \text{otherwise.} \end{cases} \tag{10}$$

Results

Implementation details of the annotation model

The annotation model is implemented with PyTorch 1.7.1 and CUDA 11.0, and all the experiments are conducted on 2 NVIDIA RTX 3090ti graphics processing units with 24-GB memory. During the SS-FT process of the classification model, each batch contains 8 labeled samples and 8 unlabeled samples. The optimization algorithm is AdamW, with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay $= 1 \times 10^{-6}$. The learning rate is initially set to 5×10^{-5} with a poly learning rate schedule. For the teacher and student framework, the EMA parameter update coefficient α is first set to 0.99 and follows the cosine schedule. The loss weights λ_{sup} and λ_{contr} are 1.0 and 0.01, respectively.

Initially, the performance of the PT model on xBD is as follows: $F1_{\text{loc}} = 0.868$, $F1_{\text{dam}} = 0.752$, and $\left\{F1_{C_i} | 1 \le c \le 4\right\} = \{0.920, 0.583, 0.750, 0.847\}$. During model optimization, the total epoch is set to 100. The entropy threshold δ is first set as the quantile 20% to categorize pixels into the top 20% reliable part and the other high-entropy part and gradually decrease according to the epoch t [see Eq. (11)].

$$\delta_t = \delta_0 \cdot \left(1 - \frac{t}{100}\right) \tag{11}$$

Algorithm 1: Contrastive learning architecture

Require: Labeled images $\mathcal{B}^l$ (with label y^l), unlabeled images $\mathcal{B}^u$; student model θ^{stu}, teacher model θ^{tea}; entropy threshold δ

for $x_i \in \mathcal{B}^l \cup \mathcal{B}^u$ **do**
  $p_i \leftarrow h \circ f(x_i; \theta^{\text{tea}})$, $\mathcal{H}(p_i) \leftarrow$ entropy of p_i ;
  $z_i^{\text{tea}} \leftarrow g \circ f(x_i; \theta^{\text{tea}})$, $z_i^{\text{stu}} \leftarrow g \circ f(x_i; \theta^{\text{stu}})$;
  for $c \leftarrow 0$ **to** $C-1$ **do**
    if $x_i \in \mathcal{B}^l$ **then**
      $y_i \leftarrow y_i^l$;
      Filter queries: $\mathcal{Q}_c^l \leftarrow \left\{\mathbf{z}_{ij}^{\text{stu}} \mid y_{ij} = c\right\}$;
      Filter negative samples: $\mathcal{N}_c^l \leftarrow \left\{\mathbf{z}_{ij}^{\text{tea}} \mid y_{ij} \neq c\right\}$
    else
      $\hat{y}_i \leftarrow \arg\max p_i$;
      Filter queries: $\mathcal{Q}_c^u \leftarrow \left\{\mathbf{z}_{ij}^{stu} \mid \hat{y}_{ij} = c, \mathcal{H}(p_{ij}) < \delta\right\}$;
      Filter negative samples: $\mathcal{N}_c^u \leftarrow \left\{\mathbf{z}_{ij}^{\text{tea}} \mid \text{argsort}(p_{ij})[c] \geq 2, \mathcal{H}(p_{ij}) \geq \delta\right\}$;
    end
    $\mathcal{Q}_c \leftarrow \mathcal{Q}_c^l \cup \mathcal{Q}_c^u$, $\mathcal{N}_c \leftarrow \mathcal{N}_c^l \cup \mathcal{N}_c^u$;
    Push $\mathcal{N}_c$ into memory bank, and pop oldest ones if necessary ;
    Calculate $\mathcal{L}_{\text{contr}}$ based on $\mathcal{Q}_c$ and $\mathcal{N}_c$;
  end
end

Quantitative analysis of model annotation

Pakistan flooding, Turkey earthquake, and Hurricane Ian are selected as 3 disaster events to display the detailed experimental results. Each of the 3 disasters has a distinct damage scenario and geographical environment and therefore can show the generalization of our approach. The final manual-labeled samples accounted for 6%, 15%, and 6% of the 3 events, respectively, and samples were then divided into the train set and the validation set (60%:40%) by 3 random seeds. In addition to SS-FT, 2 other baseline settings, PT only and supervised FT, served as references to verify the effectiveness of our proposed method. Before optimizing the classification network θ_{cls}, the FT of localization network θ_{loc} was supervised due to its relative simplicity.

As Table 4 shows, the results of 3 disasters all prove the effectiveness of our method in improving the model's performance. When directly applying the PT model to an unseen disaster, the labeling accuracy of damaged classes is not satisfactory. For instance, most of the damage information in Turkey earthquake disaster failed to be recognized because of the scarcity of earthquake-induced damaged samples in xBD, the model's original pretraining material. As the model was fine-tuned on a few manually labeled samples, its performance on the test set improved significantly. In addition, with the contrastive learning loss incorporated into the FT process, the $F1_{\text{dam}}$ scores of Pakistan flooding, Turkey earthquake, and Hurricane Ian reached 62.22%, 41.64%, and 61.25%, respectively. The improvements of SS-FT compared to FT are 4.47%, 5.89%, and 1.27%. Table 5 further

Table 4. Comparison results of model performance under PT only, FT, and SS-FT settings

Dataset	Setting	$F1_{\text{dam}}$(%)	$F1_{\text{dam}}$ (%) per damage level			
			No damage	Minor damage	Major damage	Destroyed
Pakistan flooding	PT	46.15	67.11	46.35	38.63	41.12
	FT	58.99	84.51	59.07	44.41	60.52
	SS-FT	**63.29**	**85.34**	**65.16**	**48.31**	**64.80**
Turkey earthquake	PT	2.70	76.59	0.84	3.94	39.58
	FT	35.45	79.49	22.60	26.98	52.75
	SS-FT	**40.66**	**80.71**	**26.08**	**34.71**	**53.08**
Hurricane Ian	PT	36.53	86.75	32.29	25.23	–
	FT	59.98	87.11	56.55	47.96	–
	SS-FT	**61.25**	**89.19**	**57.68**	**48.94**	–

Hurricane Ian only takes 3 damage levels into accuracy calculation due to the very scarcity of destroyed buildings. Bold values indicate the best performance per damage level of every disaster.

compares the results of FT and SS-FT in different rounds of optimization. It can be seen that the contrastive mechanism brings extra accuracy improvement in different labeled proportions. Besides, the SS-FT gets higher improvements in less proportion of labeled data, which means that the representation distinguishing on unlabeled samples can bring more effect when supervised information is deficient. These quantitative results prove that our proposed SS-FT can activate the damage identification performance of the model with the lowest annotation costs.

The samples from each disaster event were automatically labeled by the iterative workflow. In the last round, the accuracy inspection work was performed on the pseudo labels after automatic OV postprocessing. Together with 3 volcanic eruption events (annotated automatically in our previous work), the 5% sampling inspection results of all 12 disasters are summarized in Fig. 5. The $F1_{loc}$ is higher than 87% on average, implying that the localization model can achieve good performance in disaster-specific scenarios only by FT. Comparatively, the damage classification scores display a large variation among disasters. The $F1_{dam}$ for the subsets of hurricanes and tornadoes that struck North America are all higher than 70%. The scores of 3 volcanic eruption disasters are also satisfactory since the ash-covered damage was additionally identified by a few-shot classification model mentioned in the "Automatic annotation models" section. While for other disaster events with markedly distinct damage characteristics from the xBD dataset, the accuracy of model labeling ranges from 46% to 65%, limited by the inadequacy of the labeled samples.

Visual analysis of model annotation

Figure 6 shows several test images of different model settings, intuitively showcasing the model's performance on sample annotation. FT helps the PT model adapt to the new disaster domain and learn disaster-specific damage information, seeing that some obvious misannotations of damage level are corrected by the FT model. Then, with a pixel-level contrastive learning mechanism, predictions of SS-FT are semantically closer to the ground truth compared to the supervised-only setting. It is also worth noting that the OV strategy can utilize the building contours to improve the intrabuilding damage semantic consistency, which further helps the SS-FT predictions achieve the best visual performance.

In our SS-FT process, the entropy map plays an important role in filtering the queries (pixels with low entropy) and negative samples (pixels with high entropy) into category-wise contrastive learning. Each entropy value represents the certainty of the pixel's predicted category. Figure 7 visualizes the entropy maps of 2 unlabeled samples under different epochs. Initially, the mixed area of buildings and surrounding background showed high entropy, while some in-building foreground pixels were not, indicating that the PT model is unseen to domain-specific building appearance and damage characteristics. After the first several epochs, the background pixels had much lower entropy, and the model tended to absorb more foreground but uncertain-damage pixels into the SSL mechanism. As training proceeded, the high-entropy regions filtered by δ_t narrowed to part of the buildings, and the pixel uncertainty of damaged buildings decreased, showing the model's adaptation to the target domain.

Comprehensive analysis of semiautomatic workflow

Based on the pseudo labels obtained by model labeling and OV postprocessing, the expert inspection postoptimization was then performed to correct the bad cases. For the final labels, damage classification accuracy of each object-level building can be seen

Table 5. Comparison results of $F1_{dam}$ (%) between SS-FT and FT under different labeling proportions

Disaster	Pakistan flooding			Turkey earthquake			Hurricane Ian		
Setting	2%	4%	6%	5%	10%	15%	2%	4%	6%
FT	49.12	55.26	58.99	21.88	32.41	35.45	53.86	57.59	59.98
SS-FT	55.73	59.80	63.29	29.46	39.57	40.66	57.62	58.09	61.25
Improvement	6.61	4.54	4.30	7.58	7.16	5.21	3.76	0.50	1.27

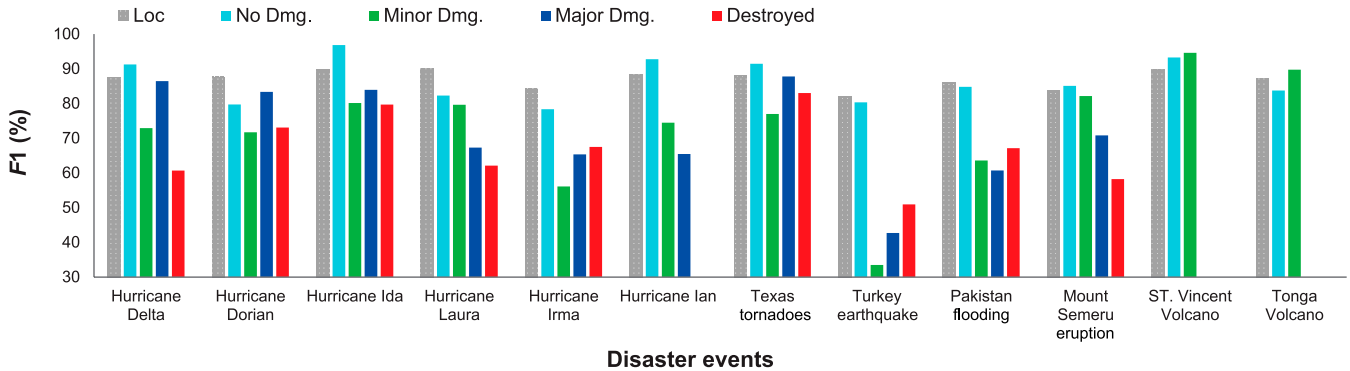

Fig. 5. Accuracy of model-annotated pseudo labels. The inspection set for Hurricane Ian only contains the first 3 damage levels; the inspection set for Tonga Volcano and St. Vincent Volcano only contains the first 2 damage levels.

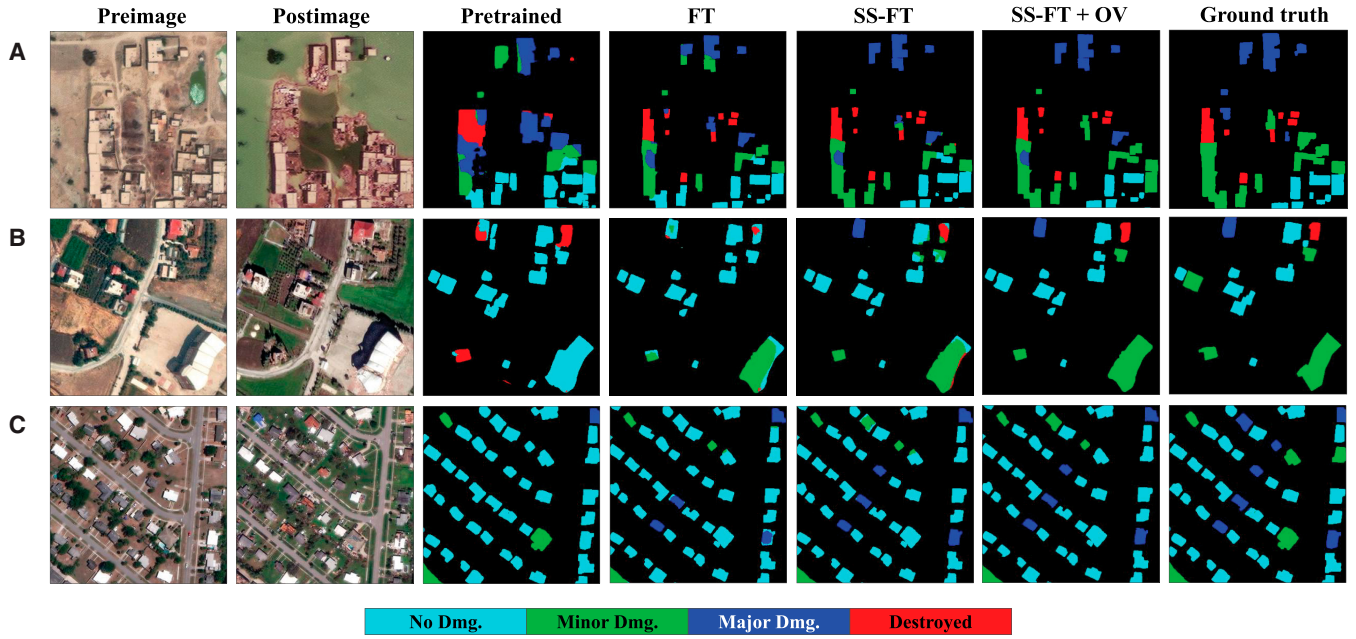

Fig. 6. Visual examples of samples with different annotation strategies. (A) Pakistan flooding, (B) Turkey earthquake, and (C) Hurricane Ian.

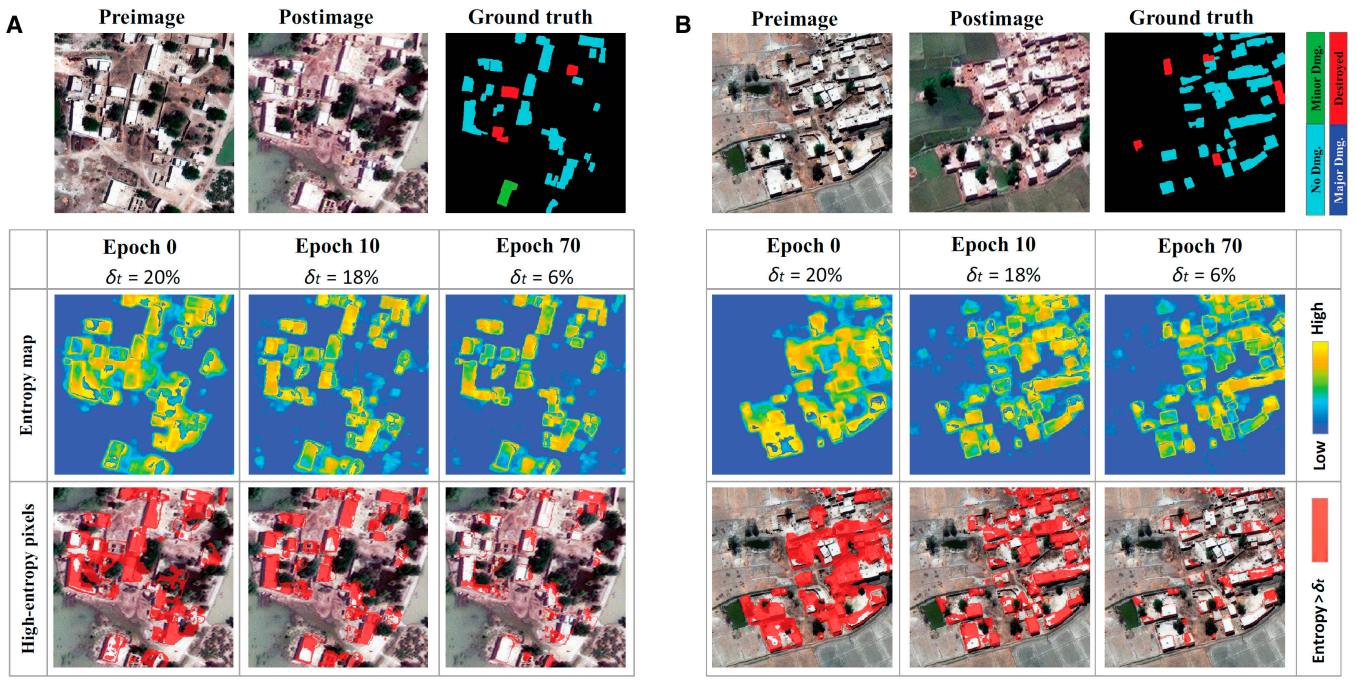

Fig. 7. Visualization of entropy maps during SS-FT process, with δ_t in different epoch calculated by Eq. (11). (A) and (B) are 2 unlabeled samples in Pakistan flooding.

as 100%, if contour localization accuracy is ignored. Figure 8 gives a typical example to show the combo of model annotation and manual work. A scene from Turkey earthquake is selected here since the $F1_{\text{dam}}$ inspection accuracy of Turkey earthquake was the lowest among all events. The original PT model neglected most of the damage information, especially the minor and major damaged parts. Then, the visual comparison between the pseudo label given by the SS-FT model and the original model highlights the huge leap in automatic annotation quality brought by model optimization. On its basis, the pixels corrected by expert postoptimization accounted for 12.66% of the total nonbackground pixels, involving 23.25% building patches. It is observed that the machine annotator, the intelligent model, still handled most of the annotation work compared to full-scene dense annotation. In addition, for those being corrected building patches, the quick correction based on pseudo labels consumes much less time than building coordinate delineation, as introduced in Fig. 1.

Table 6 shows the time consumption of 3 main stages: manual labeling, model labeling, and manual postoptimization.

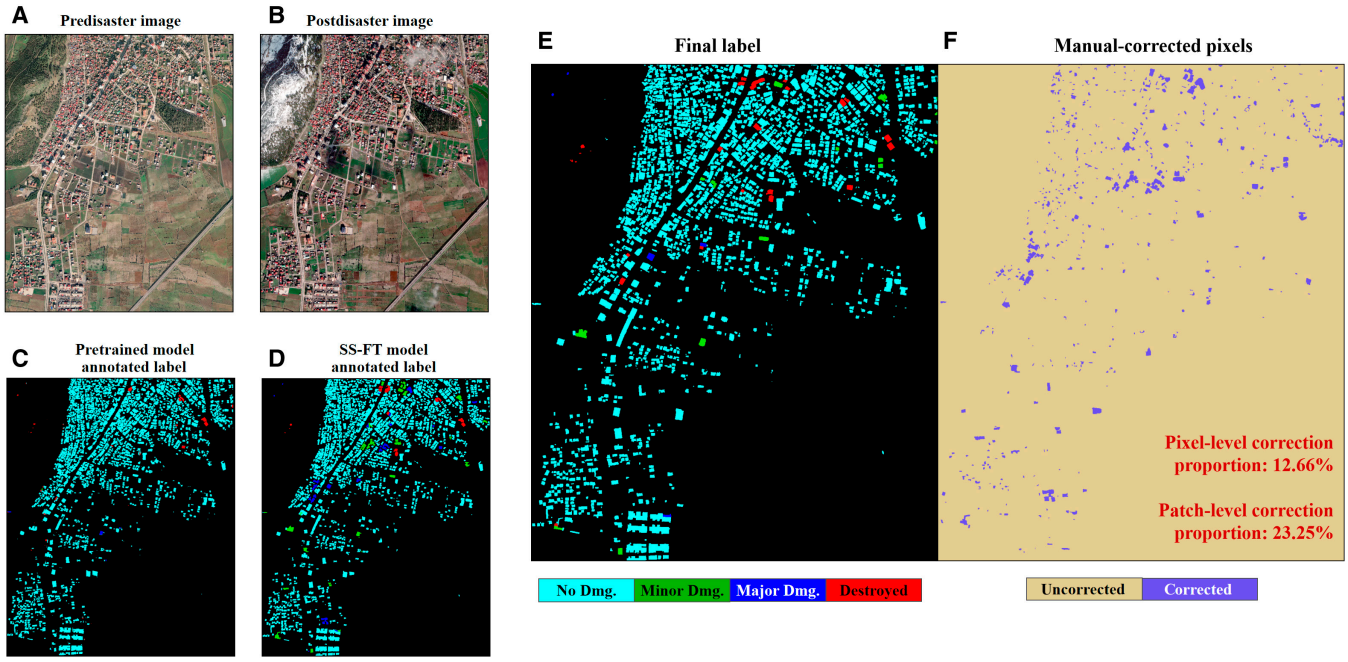

Fig. 8. Empirical study of the manual postoptimization work quantity in a hard-to-label disaster scene (Turkey earthquake). (A) and (B) show pre- and post-disaster images. (C) and (D) show the pseudo label annotated by the PT model and the SS-FT model. (E) and (F) show the final label and the corrected pixels after manual work.

Table 6. Time consumption analysis of EBD's construction

Workflow	Stage	Sample proportion	Average time (minutes per sample)	Time (hours)
Traditional	Manual labeling	100% (18,000)	10	3,000
Ours	Manual labeling	10% (1,800)	10	300
	Model labeling	90% (16,200)	0.01	2.7
	Manual postoptimization		0.1–1	27–270
	(Total)	100% (18,000)	1.1–1.9	330–573

Since each EBD sample contains more than 10 buildings on average, its dense annotation including building outline delineation and damage level rating practically costs no less than 10 min. In our work, however, the manual correction based on pseudo labels takes much less time, which fluctuates from 0.1 to 1 min. In addition, the manual work can be shortened by the higher accuracy of pseudo labels. To achieve this, around 10% samples of each disaster are manually labeled for model FT and validation. In total, EBD's construction time saved by our semiautomatic workflow was more than 80% of the fully manual annotation.

EBD: A new multidisaster dataset for BDA

Covering 12 disaster events and multiple building damage types, our proposed EBD dataset has promising applications in BDA tasks. EBD's broad building damage information enables itself to serve as an independent dataset for emergency mapping, which is statistically analyzed in the "Internal analysis" section and practically tested in the "Emergency mapping application" section. Meanwhile, EBD's compatibility with the xBD dataset makes itself a vital complementary resource to xBD's existing disaster materials, which is experimentally elaborated in the "External evaluation on xBD" section.

Internal analysis

The EBD dataset has over 18,000 samples with over 175,000 connected components, which can be counted as the minimal building quantity. As Fig. 9 shows, 18.40% of the total buildings are damaged, and the intercategory distribution varies greatly across different disaster events.

Covering multiple disaster events, EBD is expected to provide universal building damage characteristics. Figure 10A displays bitemporal images and their red–green–blue (RGB) histograms from different disasters, which showcase the distinct characteristics of building damage and geographic environments. Down to each pair of images, the modification of bitemporal histograms is also specific to its disaster context. Furthermore, we mapped the representations of all nonbackground pixels into

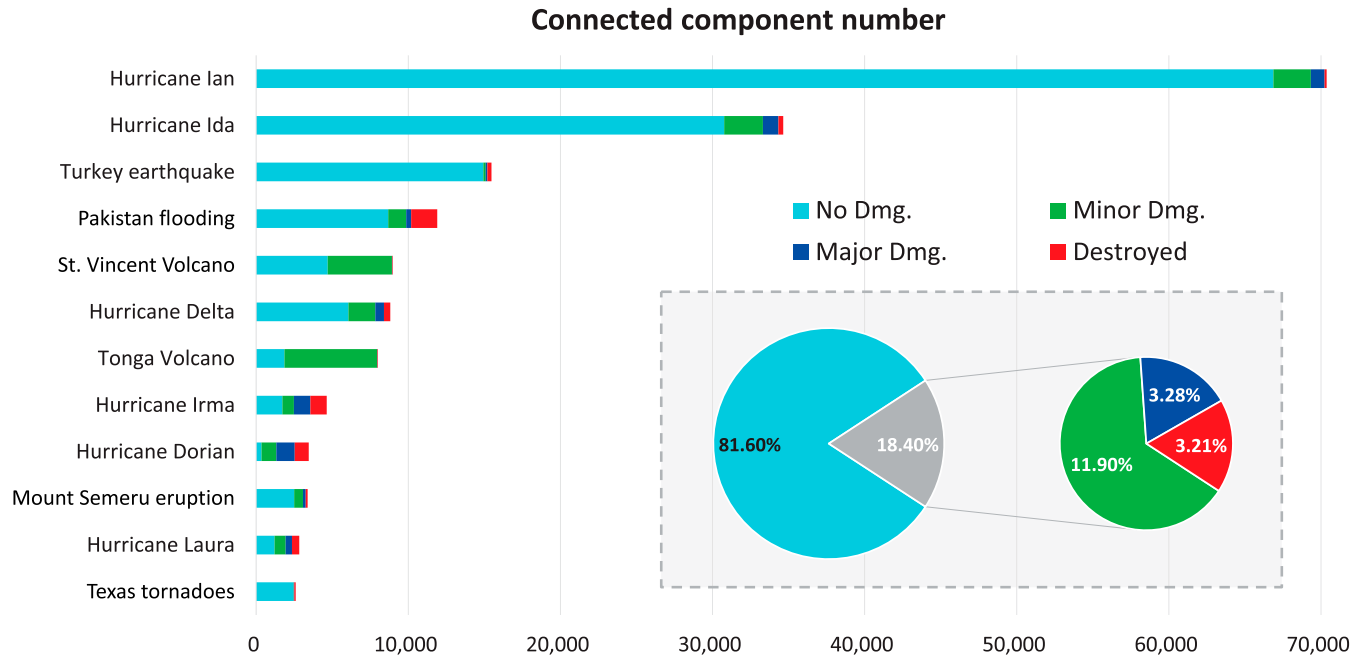

Fig. 9. Building number statistics for all disasters and damage levels.

2-dimensional space by *t*-distributed stochastic neighbor embedding (t-SNE) method [37] and marked them as the mean positions of each damage level. As Fig. 10B shows, apart from the clustering trend of the same disaster, the undamaged pixels across all disasters are distributed relatively close, while the "major damaged" and "destroyed" features are scattered all around, indicating their heterogeneous characteristics.

In addition to diversity, the intradataset consistency of labeled information is also crucial for EBD's application in the BDA tasks as a unified dataset. Here, we adopted a 5-fold cross-validation experiment on EBD, and the structure of the base model is the same as the annotation model. Figure 10C shows the statistical accuracy performance of the model trained on EBD (named model-EBD). The value of $F1_{\text{loc}}$ and overall $F1_{\text{dam}}$ scores are $88.31_{\pm 0.88}$% and $74.36_{\pm 0.90}$%, respectively; also, the standard deviations of $F1_{\text{dam}}$ in 4 damage levels are 0.14%, 3.00%, 3.32%, and 0.81%, respectively. The results demonstrate the high intraconsistency of EBD, proving that the semiautomatic construction process has successfully made the dataset information conform to expert knowledge.

Emergency mapping application

Having established that EBD encompasses diverse and reliable building damage information, we select an out-of-EBD disaster, Libya flooding (2023 September), to test the ability of EBD as a knowledge base being referred to in real emergency response. Figure 11 shows the BDA results of the AOI by model-EBD. The major damaged and destroyed buildings are mainly distributed in the estuary and embankment areas along the Derna River. This highly conforms to the severe flooding impacts shown in the post-disaster image, which reduced building blocks to mud and debris. Constrained by model-EBD unseen to Libya flooding, some visually minor damaged buildings in low-lying flood-prone areas were unrecognized. Still, the overall satisfactory results showcase the potential of EBD to aid emergency mapping and disaster rescue.

External evaluation on xBD

Sharing the same damage classification criteria and WorldView satellite data source with xBD, EBD is also a vital supplement to the existing materials of xBD. To prove this, we explored the domain difference and the generalization between EBD and xBD. Experimental results of 4 typical xBD disaster events, categorized into groups I and II, are shown in Table 7. The model-EBD performed significantly better in group I than that in group II. The results indicate that the damage characteristics of wildfire and tsunami scenarios display larger domain gaps to EBD than flooding and hurricane scenarios, especially under "minor damaged" and "major damaged" classes.

For these 4 xBD disasters, we also compared the performances of models trained from scratch (TFS), pretrained on EBD only (PT), fine-tuned from model-EBD (FT), and semisupervised fine-tuned from model-EBD (SS-FT); the training set of each disaster was randomly sampled. Results show that the performance of TFS models is far from the PT or FT setting. Comparison results have verified the EBD's generalization as a benchmark for new disasters, e.g., the model of Nepal flooding improves 33.07% $F1_{\text{dam}}$ when being pretrained on EBD, and the result of Portugal wildfires improves 29.56% $F1_{\text{dam}}$. Moreover, the model's adaptation performance was further enhanced when the remaining samples in the target domain were incorporated into contrastive learning.

Figure 12 shows representative cases of model-EBD's prediction on xBD samples. Figure 12A to C is cases where the PT model can already identify building damage information basically consistent with the ground truth; Figure 12D to F is cases where the EBD model fails to predict the damage information accurately but significantly improves its performance through semisupervised optimization on a few labeled samples of the target disaster. This verifies that even if EBD does not encompass all building damage characteristics, its utility can be generalized to xBD as complementary materials or extended to new disaster scenarios when serving as pretraining materials.

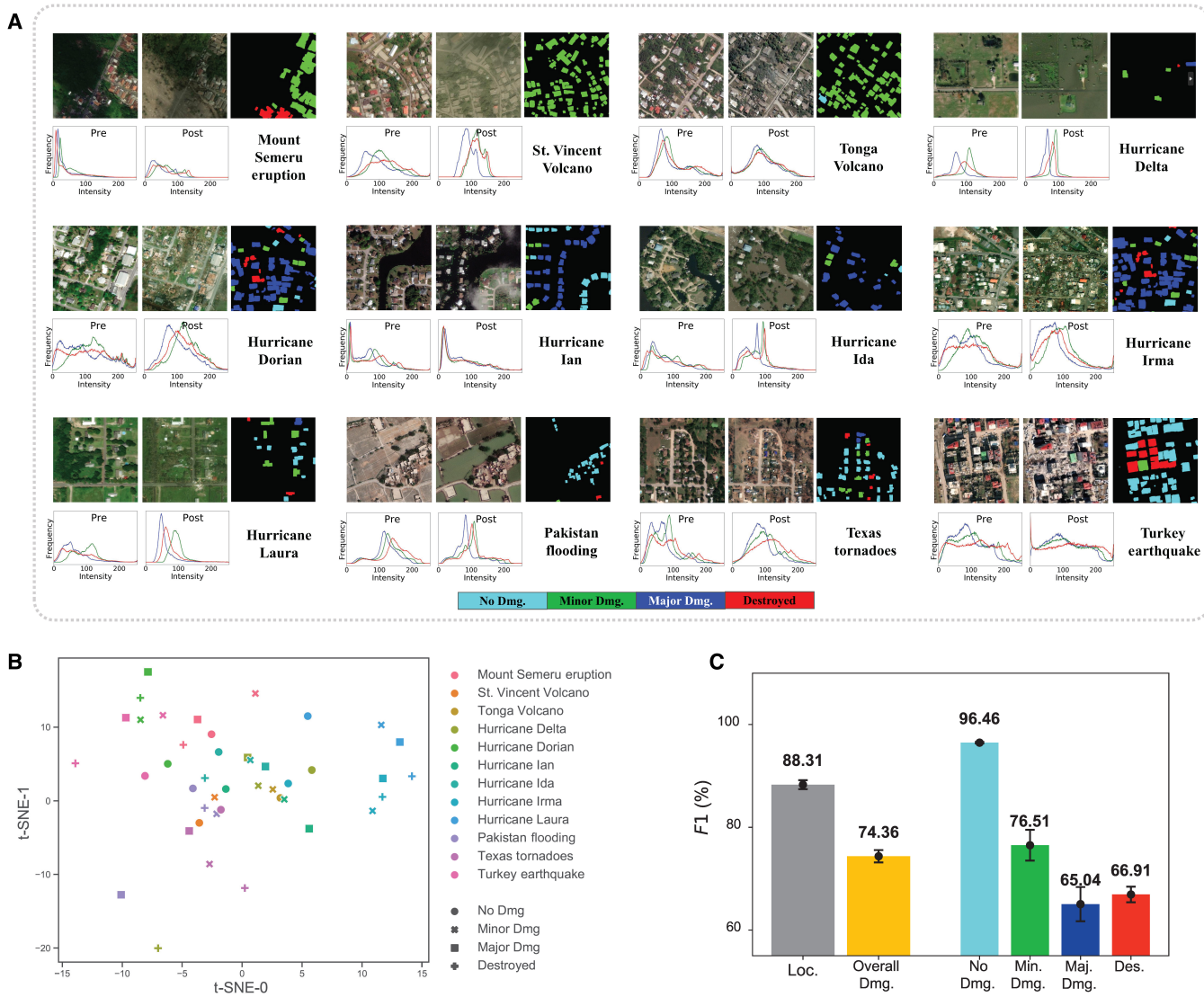

Fig. 10. Internal analysis of the EBD dataset. (A) shows sample diversity: For each disaster's instance sample, the first row contains the pre-disaster image (left), post-disaster image (middle), and building damage label (right); the second row contains the RGB color histograms of bitemporal images. (B) shows feature distribution: Each 2-dimensional point is the center of pixel-level representations categorized by the union of disaster events and damage levels. (C) shows 5-fold cross-validation results of EBD.

Discussion

Data contribution of EBD

In the fine-grained BDA research field, the one and only benchmark, the xBD dataset, was introduced in 2019 by many disaster response agencies (including California Air National Guard, NASA, and Federal Emergency Management Agency) and has remained static since then. To this insufficiency, our constructed EBD dataset comes as a powerful complement to natural disasters that happened in recent years, covering 12 events with multiple building damage characteristics. Figure 13 shows the disaster distribution of EBD and xBD. Notably, 5 countries (Pakistan, Turkey, St. Vincent, Bahamas, and Tonga) are newly included in our EBD, and all of them are developing countries. In addition, the samples of volcanic eruption, earthquake, and flooding in EBD have brought huge knowledge enrichment to the less-distributed disaster types of xBD. In both research and application aspects, the enrichment of disaster data by EBD can improve the equity of intelligent disaster response globally. Note that the knowledge base of EBD combined with xBD can be further expanded at a low cost. This prospect is ensured by the machine-driven annotation experience from EBD, as only auxiliary expert supervision rather than intensive manual labeling was required for new disaster annotation.

Inspirations and limitations of EBD's construction

Apart from the data contribution, we also presented a roadmap for efficiently constructing an expert-knowledge-requiring RS dataset. Earlier efforts aiming at constructing BDA datasets, as listed in Table 1, all rely on manual work, either expert labeling or crowd-sourced labeling. Meanwhile, recent research has applied generative models in giving fake building damage samples [38,39], but there is a remarkable gap between its focus on "binary change" information and our "multilevel damage" information. Overall, our practice is the first to utilize the machine expert, a DL model equipped with prior disaster knowledge, to construct a large-scale and high-accuracy BDA dataset.

In terms of making the machine expert function well in a new disaster scenario, the difficulty lies in 2 aspects. One is only limited labeling available for supervision, and the other is the severe sample quantity imbalance between categories. Motivated by studies of SSL for semantic segmentation [32,33], RS image classification [40,41], and RS species classification [42], we designed a pixel-level contrastive learning module into BDA tasks. Our idea is straightforward in incorporating large part of unlabeled samples into feature discrimination among different damage levels. Moreover, to balance the negative samples of each category limited by the batch size, the category-wise memory bank mechanism was adopted to store and reuse category features. According

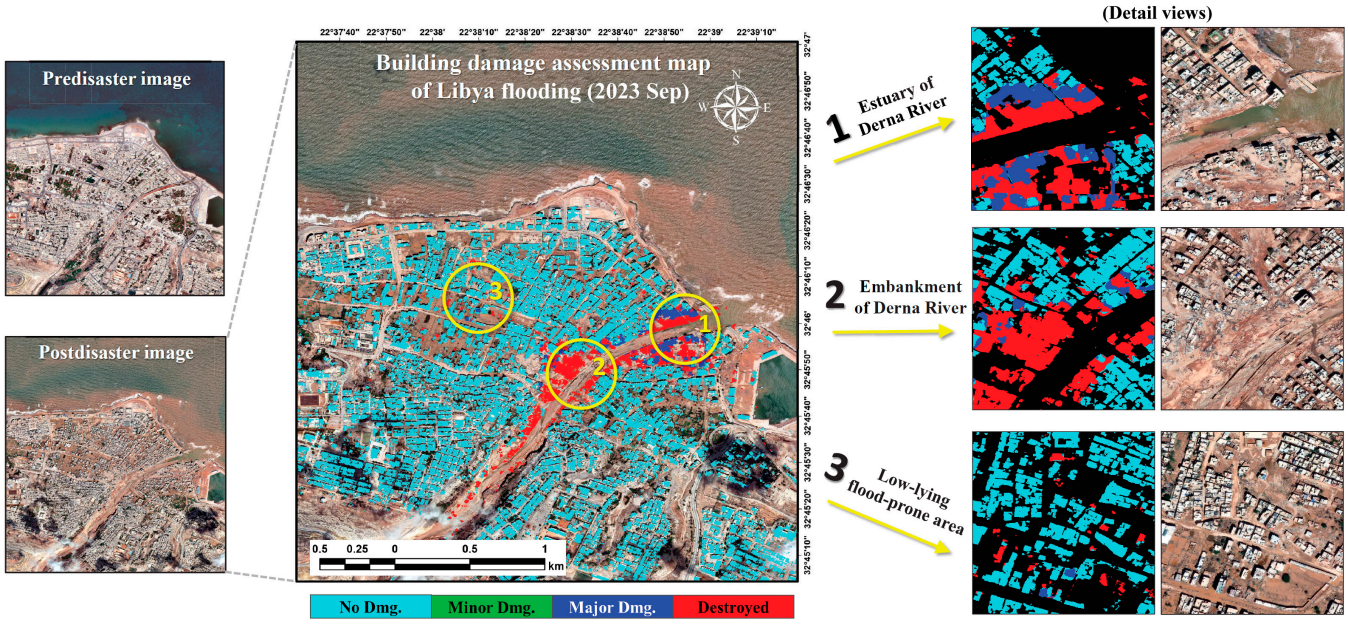

Fig. 11. Building damage mapping of Libya flooding (2023 September) by the model trained on EBD.

Table 7. Quantative results of the model pretrained on the EBD dataset and validated on disasters in the xBD dataset

	Target disaster		Setting				$F1_{dam}$ (%) per damage level			
Group[a]	Name	Training set for L_{sup}	Pretrained	L_{sup}	L_{contra}	$F1_{dam}$ (%)	No damage	Minor damage	Major damage	Destroyed
I	Nepal flooding	10% (54)	✓			39.46	92.11	26.50	46.29	32.08
				✓		17.49	93.78	26.76	58.02	6.17
			✓	✓		50.56	93.40	44.51	58.58	34.59
			✓	✓	✓	**51.62**	**94.02**	**46.16**	**60.25**	**34.98**
	Matthew Hurricane	10% (35)	✓			32.85	42.76	34.54	24.01	36.01
				✓		20.89	21.99	28.23	25.16	14.12
			✓	✓		35.31	46.83	**48.02**	22.61	37.20
			✓	✓	✓	**38.97**	**50.49**	47.81	**25.99**	**42.64**
II	Portugal Wildfire	5% (82)	✓			17.97	98.09	8.36	20.11	23.27
				✓		–	98.11	–	6.03	43.47
			✓	✓		29.56	98.12	14.94	25.66	**52.05**
			✓	✓	✓	**36.46**	**98.39**	**23.55**	**26.99**	49.91
	Palu Tsunami	20% (33)	✓			5.25	93.43	1.69	6.91	63.60
				✓		–	89.16	–	13.06	61.46
			✓	✓		15.15	94.42	4.96	26.62	69.36
			✓	✓	✓	**15.78**	**95.43**	**5.12**	**29.82**	**70.68**

[a]The distinction between groups I and II is based on whether the disaster type is included in EBD. The symbol "–" denotes no pixels correctly identified or the 4-class $F1_{dam}$ cannot be calculated. Bold values indicate the best performance per damage level of every disaster.

to the experimental results from Table 4 (model-xBD applied to EBD events) and Table 7 (model-EBD applied to xBD events), SS-FT has been proven as an effective method to enhance the PT model's performance in new annotation-scarce disaster contexts. Unlike previous research that followed consistency regulation in semisupervised CD and BDA tasks [25–27], this study offers a contrastive-based perspective on RS image self-supervised interpretation, especially for bi- or multitemporal RS images. SS-FT does not require image-level perturbations (such as transformations of color and contrast) or feature-level perturbations (such as feature drop and cutout), thereby avoiding confusion with the real building damage information conveyed by images.

Despite the effectiveness of SS-FT, there are still some limitations. Since RS images from different disasters exhibit obvious

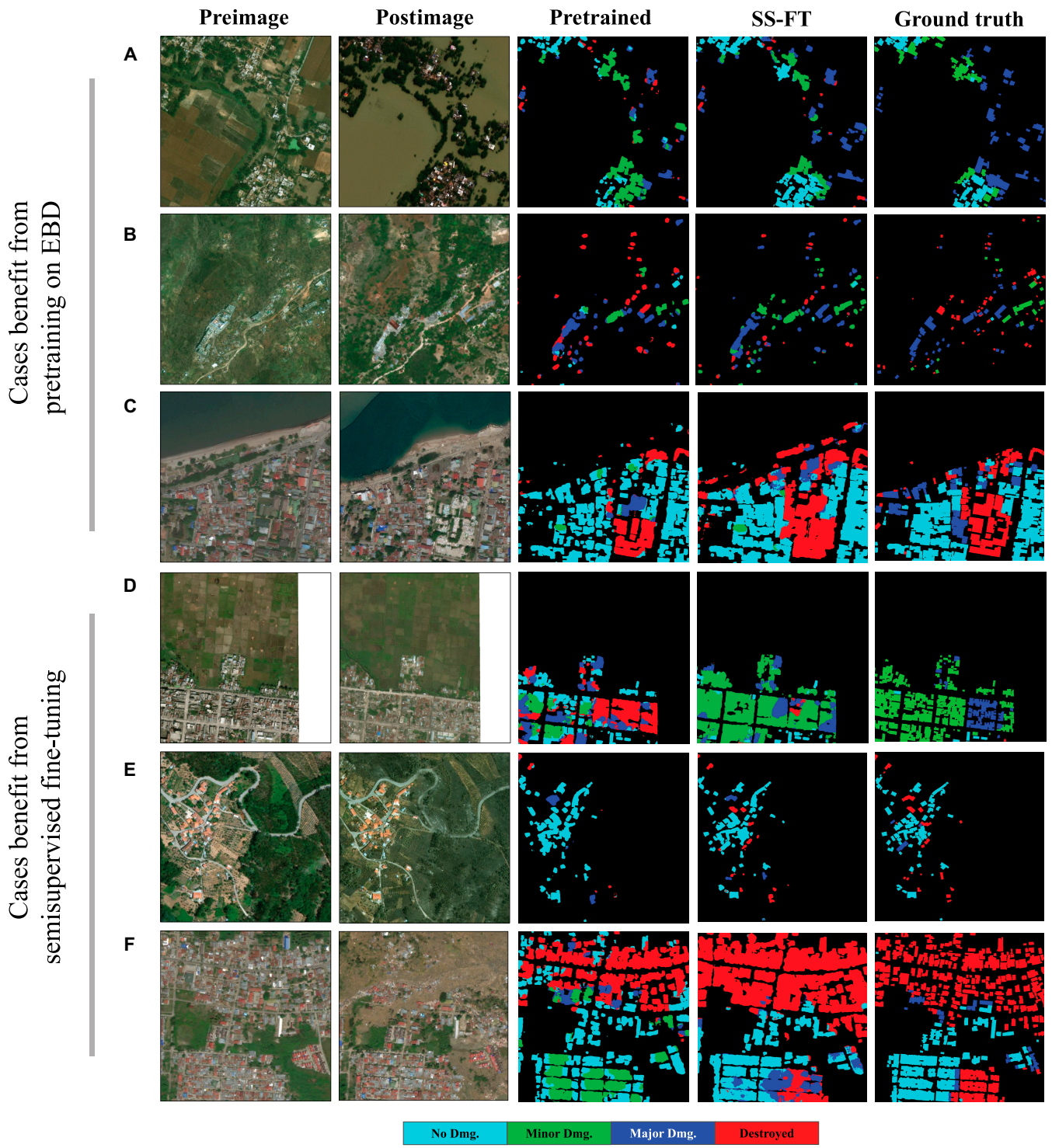

Fig. 12. Visualization of model-EBD's generalization on xBD samples. (A) is from Nepal flooding. (B) and (D) are from Matthew Hurricane. (C) and (F) are from Palu Tsunami. (E) is from Portugal Wildfire.

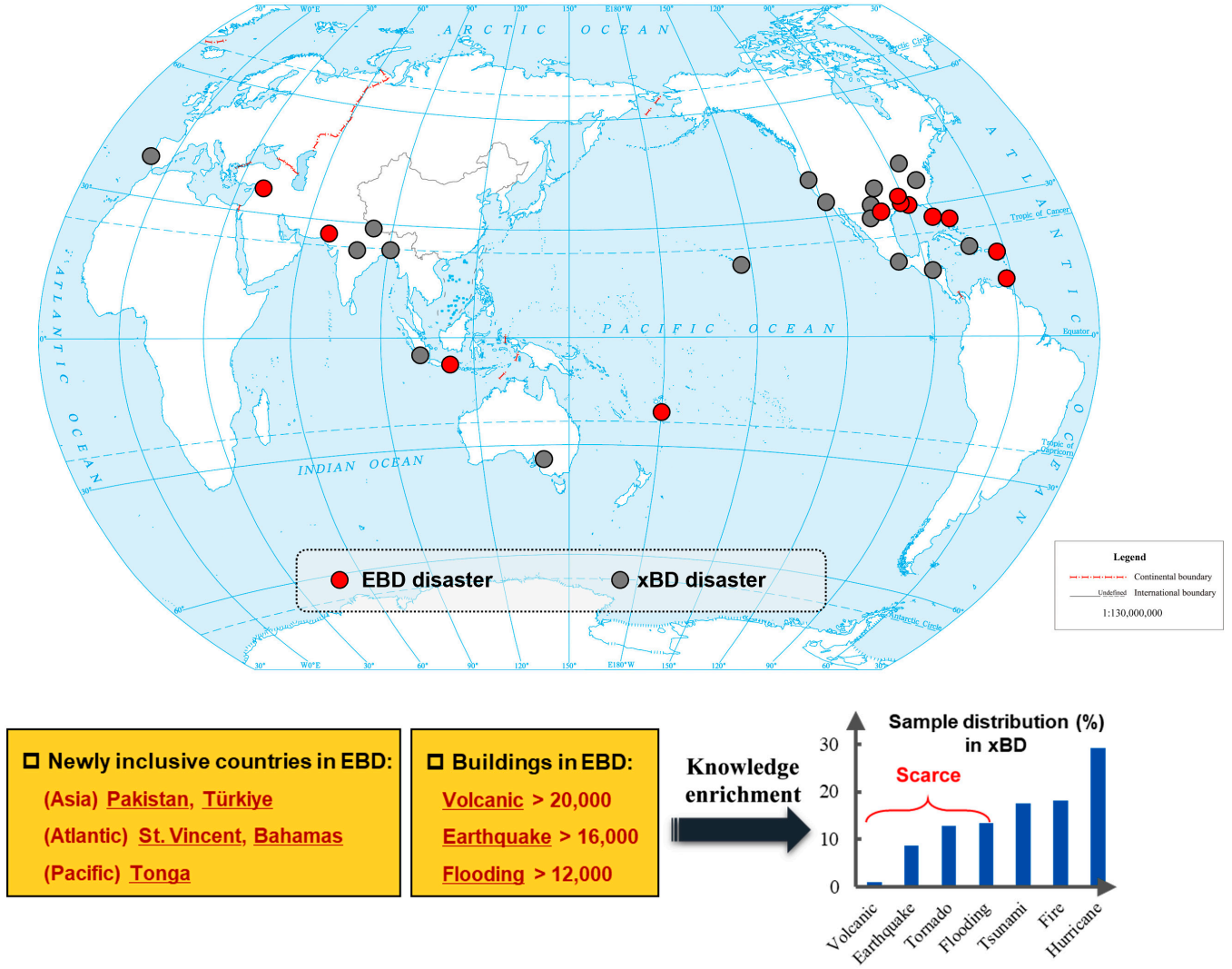

Fig. 13. Disaster knowledge enrichment brought by EBD to the original xBD.

differences in image style and damage characteristics, we optimized task-specific models instead of a general model for all disasters. The potential cross-domain problem was avoided, but it brings repeated training and validation work for all disasters. In addition, the effectiveness of our SS-FT relies on the feature space of one single domain. This limits the collaborative learning of building damage knowledge from multiple disaster events. Some studies have designed mechanisms to learn a shared decision boundary across different domains, such as the representation alignment method for cross-domain landslide extraction [43] and the knowledge aggregation module for global-scale building mapping [44]. When the scenario is for natural disasters, a more delicate design of SS-FT is required to remove the domain shift of different disaster types and geographic contexts. Incorporating other TL strategies, such as feature-based domain adaptation [11,45,46] or data-based generative adversarial networks [20,47], is expected to address this bottleneck.

Conclusion

To bridge the insufficiency in the renewal of disaster samples, this paper introduced a new multidisaster building damage dataset named EBD and a machine-driven labeling solution. The use of the machine annotator has reduced dataset construction time by 80% compared to full manual labeling. The DL model was trained as the machine annotator by first learning prior knowledge from historical disaster samples and then transferring this knowledge to new disasters through SS-FT. With very limited manual labeling in new disasters, the main idea of SS-FT is to utilize both labeled and unlabeled samples to help the DL model best distinguish positive and negative pixel representations among different levels of building damage. Experimental results of 3 demonstration disasters, i.e. Pakistan flooding, Turkey earthquake, and Hurricane Ian, show that the model has gained at least 17.14%, 39.82%, and 24.72% $F1_{dam}$ improvements under SS-FT compared to the PT-only setting and 4.30%, 5.21%, and 1.27% improvements compared to the FT setting, more so with less manually labeled samples.

The final EBD dataset is constructed as a large-scale and EBD dataset. The large scale is reflected in its pixel-level annotations of over 175,000 buildings from 12 disaster events; the extensibility is demonstrated by the semiautomatic construction process, which can alleviate most of the manual labor. Internel-EBD analysis work has proven EBD's overall labeling

consistency and sample feature diversity across different geographical contexts and disaster types. External-EBD experiments have proven EBD's practical application in emergency mapping as an independent knowledge base and generalization yet domain difference to the existing BDA benchmark, xBD. In the future, we expect the EBD dataset to contribute to open science and facilitate disaster response technology equity on a global scale, especially for developing countries.

Acknowledgments

We gratefully acknowledge Maxar/DigitalGlobe for data sharing through their Open Data Program (https://www.maxar.com/open-data).
Funding: This work was supported by the National Key Research and Development Program of China under grant 2019YFE0127400 and Kakenhi under grant 25K03145.
Author contributions: Z.W.: Methodology, software, writing—original draft, formal analysis, visualization, and writing—review and editing. C.W.: Methodology and software. F.Z.: Conceptualization, writing—review and editing, supervision, and project administration. J.X.: Supervision and writing—review and editing.
Competing interests: The authors declare that they have no competing interests.

Data Availability

Our constructed EBD dataset is publicly available at https://doi.org/10.6084/m9.figshare.25285009.v2. The codes of annotation optimization by SS-FT are publicly available at https://github.com/Zeoyoon/SSFT-main.

References

1. Rentschler J, Salhab M, Jafino BA. Flood exposure and poverty in 188 countries. *Nat Commun*. 2022;13(1):3527.
2. Coburn AW, Spence RJ, Pomonis A. Factors determining human casualty levels in earthquakes: Mortality prediction in building collapse. In: *Proceedings of the tenth world conference on earthquake engineering*. Rotterdam: Balkema; 1992. Vol. 10. p. 5989–5894.
3. Schweier C, Markus M. Classification of collapsed buildings for fast damage and loss assessment. *Bull Earthq Eng*. 2006;4:177–192.
4. Boccardo P, Giulio TF. Remote sensing role in emergency mapping for disaster response. In: *Engineering geology for society and territory-volume 5: Urban geology, sustainable planning and landscape exploitation*. Cham: Springer; 2015. p. 17–24.
5. Adriano B, Yokoya N, Xia J, Miura H, Liu W, Matsuoka M, Koshimura S. Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS J Photogramm Remote Sens*. 2021;175:132–143.
6. Gupta R, Goodman B, Patel N, Hosfelt R, Sajeev S, Heim E, Doshi J, Lucas K, Choset H, Gaston M. Creating xBD: A dataset for assessing building damage from satellite imagery. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*. Long Beach (CA): IEEE; 2019. p. 10–17.
7. Ji M, Liu L, Du R, Buchroithner MF. A comparative study of texture and convolutional neural network features for detecting collapsed buildings after earthquakes using pre-and post-event satellite imagery. *Remote Sens*. 2019;11(10):1202.
8. Wu C, Zhang F, Xia J, Xu Y, Li G, Xie J, Du Z, Liu R. Building damage detection using U-Net with attention mechanism from pre- and post-disaster remote sensing datasets. *Remote Sens*. 2021;13(5):905.
9. Shen Y, Zhu S, Yang T, Chen C, Pan D, Chen J, Xiao L, Du Q. Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. *IEEE Trans Geosci Remote Sens*. 2021;60:1–14.
10. Chen H, Nemni E, Vallecorsa S, Li X, Wu C, Bromley L. Dual-tasks siamese transformer framework for building damage assessment. In: *IGARSS 2022-2022 IEEE international geoscience and remote sensing symposium*. Kuala Lumpur (Malaysia): IEEE; 2022. p. 1600–1603.
11. Da Y, Ji Z, Zhou Y. Building damage assessment based on siamese hierarchical transformer framework. *Mathematics*. 2022;10(11):Article 1898.
12. Kaur N, Lee CC, Mostafavi A, Mahdavi-Amiri A. Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *Comput Aided Civ Infrastruct Eng*. 2023;38(15):2072–2091.
13. Zhang Y, Liu P, Chen L, Xu M, Guo X, Zhao L. A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: GF-FloodNet. *Int J Digit Earth*. 2023;16(1):2522–2554.
14. Shi Q, Zhu J, Liu Z, Guo H, Gao S, Liu M, Liu Z, Liu X. The last puzzle of global building footprints—Mapping 280 million buildings in East Asia based on VHR images. *J Remote Sens*. 2024;4:Article 0138.
15. Bonafilia D, Tellman B, Anderson T, Issenberg E. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. Seattle (WA): IEEE; 2020. p. 210–211.
16. Xu Y, Zhou J, Zhang Z. A new Bayesian semi-supervised active learning framework for large-scale crop mapping using Sentinel-2 imagery. *ISPRS J Photogramm Remote Sens*. 2024;209:17–34.
17. Osco LP, Wu Q, De Lemos EL, Gonçalves WN, Ramos AP, Li J, Junior JM. The segment anything model (sam) for remote sensing applications: From zero to one shot. *Int J Appl Earth Obs Geoinf*. 2023;124:Article 103540.
18. He Y, Wang J, Zhang Y, Liao C. An efficient urban flood mapping framework towards disaster response driven by weakly supervised semantic segmentation with decoupled training samples. *ISPRS J Photogramm Remote Sens*. 2024;207:338–358.
19. Shao J, Tang L, Liu M, Shao G, Sun L, Qiu Q. BDD-Net: A general protocol for mapping buildings damaged by a wide range of disasters based on satellite imagery. *Remote Sens*. 2020;12(10):Article 1670.
20. Hu Y, Tang H. On the generalization ability of a global model for rapid building mapping from heterogeneous satellite images of multiple natural disaster scenarios. *Remote Sens*. 2021;13(5):Article 984.
21. Ma Y, Chen S, Ermon S, Lobell DB. Transfer learning in environmental remote sensing. *Remote Sens Environ*. 2024;301:Article 113924.
22. Valentijn T, Margutti J, Van den Homberg M, Laaksonen J. Multi-hazard and spatial transferability of a CNN for automated building damage assessment. *Remote Sens*. 2020;12(17):Article 2839.

23. Zheng Z, Zhong Y, Wang J, Ma A, Zhang L. 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*. 2021;265:Article 112636.
24. Yang W, Zhang X, Luo P. Transferability of convolutional neural network models for identifying damaged buildings due to earthquake. *Remote Sens*. 2021;13(3):Article 504.
25. Zhang X, Huang X, Li J. Joint self-training and rebalanced consistency learning for semi-supervised change detection. *IEEE Trans Geosci Remote Sens*. 2023;61:1–3.
26. He Y, Wang J, Liao C, Zhou X, Shan B. MS4D-Net: Multitask-based semi-supervised semantic segmentation framework with perturbed dual mean teachers for building damage assessment from high-resolution remote sensing imagery. *Remote Sens*. 2023;15(2):Article 478.
27. Lee J, Xu JZ, Sohn K, Lu W, Berthelot D, Gur I, Khaitan P, Huang KW, Koupparis K, Kowatsch B. Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques. ArXiv. 2020. https://doi.org/10.48550/arXiv.2011.14004
28. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Seoul (Korea): IEEE; 2019. p. 6023–6032.
29. Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv Neural Inf Proces Syst*. 2020;33:596–608.
30. Liu Y, Tian Y, Chen Y, Liu F, Belagiannis V, Carneiro G. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans (LA): IEEE; 2022. p. 4258–4267.
31. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. ArXiv. 2018. https://doi.org/10.48550/arXiv.1807.03748
32. Alonso I, Sabater A, Ferstl D, Montesano L, Murillo AC. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Montreal (Canada): IEEE; 2021. p. 8219–8228.
33. Wang Y, Wang H, Shen Y, Shen Y, Fei J, Li W, Jin G, Wu L, Zhao R, Le X. Semi-supervised semantic segmentation using unreliable pseudo-labels. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans (LA): IEEE; 2022. p. 4248–4257.
34. Pesaresi M, Gerhardinger A, Haag F. Rapid damage assessment of built-up structures using VHR satellite data in tsunami-affected areas. *Int J Remote Sens*. 2007;28:3013–3036.
35. Brunner D, Lemoine G, Bruzzone L. Earthquake damage assessment of buildings using VHR optical and SAR imagery. *IEEE Trans Geosci Remote Sens*. 2010;48(5):2403–2420.
36. Wang Z, Zhang F, Wu C, Xia J. Rapid mapping of volcanic eruption building damage: A model based on prior knowledge and few-shot fine-tuning. *Int J Appl Earth Obs Geoinf*. 2024;126:Article 103622.
37. Van derMaaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–2605.
38. Zheng Z, Tian S, Ma A, Zhang L, Zhong Y. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In: *Proceedings of the IEEE/CVF international conference on computer vision*. Paris (France): IEEE; 2023. p. 21818–21827.
39. Zheng Z, Ermon S, Kim D, Zhang L, Zhong Y. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Trans Geosci Remote Sens*. 2024;47(2):725–741.
40. Zeng Q, Geng J. Task-specific contrastive learning for few-shot remote sensing image scene classification. *ISPRS J Photogramm Remote Sens*. 2022;191:143–154.
41. Huang W, Shi Y, Xiong Z, Wang Q, Zhu XX. Semi-supervised bidirectional alignment for remote sensing cross-domain scene classification. *ISPRS J Photogramm Remote Sens*. 2023;195:192–203.
42. Chen L, Wu J, Xie Y, Chen E, Zhang X. Discriminative feature constraints via supervised contrastive learning for few-shot forest tree species classification using airborne hyperspectral images. *Remote Sens Environ*. 2023;295:Article 113710.
43. Zhang X, Yu W, Pun MO, Shi W. Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS J Photogramm Remote Sens*. 2023;197:1–17.
44. Zhong Y, Yan B, Yi J, Yang R, Xu M, Su Y, Zheng Z, Zhang. Global urban high-resolution land-use mapping: From benchmarks to multi-megacity applications. *Remote Sens Environ*. 2023;298:Article 113758.
45. Li Y, Lin C, Li H, Hu W, Dong H, Liu Y. Unsupervised domain adaptation with self attention for post-disaster building damage detection. *Neurocomputing*. 2020;415:27–39.
46. Zheng Z, Zhong Y, Zhang L, Burke M, Lobell DB, Ermon S. Towards transferable building damage assessment via unsupervised single-temporal change adaptation. *Remote Sens Environ*. 2024;315:Article 114416.
47. Ge J, Tang H, Yang N, Hu Y. Rapid identification of damaged buildings using incremental learning with transferred data from historical natural disaster cases. *ISPRS J Photogramm Remote Sens*. 2023;195:105–128.
48. Fujita A, 590 Sakurada K, Imaizumi T, Ito R, Hikosaka S, Nakamura R. Damage detection from aerial images via convolutional neural networks. In: *2017 fifteenth IAPR international conference on machine vision applications (MVA)*. Nagoya (Japan): IEEE; 2017. p. 5–8.
49. Ci T, Liu Z, Wang Y. Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression. *Remote Sens*. 2019;11(23):Article 2858.
50. Zhu X, Liang J, Hauptmann A. Msnet: A multilevel instance segmentation network for natural disaster damage assessment in aerial videos. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. Virtual: IEEE; 2021. p. 2023–2032.
51. Chen SA, Escay A, Haberland C, Schneider T, Staneva V, Choe Y. Benchmark dataset for automatic damaged building detection from post-hurricane remotely sensed imagery. ArXiv. 2018. https://doi.org/10.48550/arXiv.1812.05581
52. Rudner TG, Rußwurm M, Fil J, Pelich R, Bischke B, Kopačková V, Biliński P. Multi3net: Segmenting flooded buildings via fusion of multiresolution, multisensor, and multitemporal satellite imagery. *Proc AAAI Conf Artif Intell*. 2019;33:702–709.
53. Hänsch R, Arndt J, Lunga D, Gibb M, Pedelose T, Boedihardjo A, Petrie D, Bacastow TM. Spacenet 8-The

detection of flooded roads and buildings. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. New Orleans (LA): IEEE; 2022. p. 1472–1480.
54. Rahnemoonfar M, Chowdhury T, Sarkar A, Varshney D, Yari M, Murphy RR. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*. 2021;9:89644–89654.
55. Rahnemoonfar M, Chowdhury T, Murphy R. RescueNet: A high resolution UAV semantic segmentation dataset for natural disaster damage assessment. *Sci Data*. 2023;10:Article 913.