Remote Sensing of Environment 331 (2025) 114979

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Neural disaster simulation for transferable building damage assessment

Zhuo Zheng [a], Yanfei Zhong [b,*], Zijing Wan [c], Liangpei Zhang [d], Stefano Ermon [a,*]

[a] Department of Computer Science, Stanford University, Stanford, 94305, CA, USA
[b] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, 430074, China
[c] College of Letters and Science, University of California, Santa Barbara, Santa Barbara, 93106, CA, USA
[d] Aerospace Information Research Institute, Henan Academy of Science, Zhengzhou, 450046, China

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Neural disaster simulation
Building damage assessment
Synthetic data fine-tuning
Deep generative models
Satellite image

ABSTRACT

Timely and reliable building damage assessment is essential for effective disaster response and humanitarian assistance. However, the diversity of disaster types, geographic regions, and data distributions poses significant challenges to transferring building damage assessment models to new disaster scenarios (i.e., target domains). In addition, limited availability of post-disaster training imagery in target domains further hinders progress. Recent approaches, such as single-temporal change adaptation, enable adaptation using only target pre-disaster images by constructing pseudo bitemporal damage samples at an unexplainable embedding level. While effective, these methods produce representations that are difficult for human experts to interpret, inspect for errors, or adjust categorical distributions to ensure reliable model performance. In this paper, we propose Neural Disaster Simulation (NeDS), a deep disaster generative model that synthesizes realistic post-disaster image from pre-event image and customizable disaster information (i.e., disaster types and disaster intensity). Thanks to damage data generation based solely on pre-event imagery, NeDS enables adaptation at any time, effectively bypassing the limitation of post-disaster training image availability. Furthermore, by explicitly modeling disaster effects at the image level, NeDS mitigates distribution shifts between historical training data and unseen disaster events, enhancing both model transferability and visual interpretability. Extensive experiments conducted on both global-scale and local-scale study areas demonstrate that NeDS adaptation outperforms the previous state-of-the-art, achieving a 4.3% improvement in average performance on a global dataset, as well as 3.6%, 7.9%, and 18.5% gains in damage classification performance for the 2025 Eaton fire, the 2025 Palisades fire in Los Angeles, and the 2025 Nigeria flooding, respectively.

1. Introduction

Accurate and timely building damage assessment is crucial for effective disaster response and recovery, as well as advancing the United Nation's Sustainable Development Goal 11 of "Sustainable Cities and Communities", which is nearing its original 2030 deadline. Satellite-based building damage assessment, as one of the most promising approaches, provides a rapid and safe way to generate large-scale building damage maps worldwide (Gupta et al., 2019b; Zheng et al., 2024a), which helps the government to make rescue plans and measure economic loss.

Deep learning-based methods have achieved state-of-the-art results in building damage assessment based on high-resolution bitemporal satellite imagery (Zheng et al., 2021, 2024c). These encouraging advances mostly benefit from architecture improvements (e.g., CNN (Durnov, 2020; Zheng et al., 2021), Transformer (Chen et al., 2022; Zheng et al., 2024c), Mamba (Chen et al., 2024), and their combinations), large-scale pre-training (He et al., 2022; Kirillov et al., 2023), and the availability of sufficient historical disaster event data (Gupta et al., 2019b; Chen et al., 2025) for supervised fine-tuning. Notably, with the emergence of large-scale self-supervised learning in the remote sensing domain, many geospatial foundation models (Ayush et al., 2021; Mall et al., 2023; Cong et al., 2022; Reed et al., 2023; Tang et al., 2024; Noman et al., 2024) have significantly improved model generalization in disaster scenarios. In addition, data augmentation-based approaches (Shen et al., 2021; Sun et al., 2025) have also been explored to improve model generalization by diversifying training data distributions. One of the key challenges in deep learning-based building damage assessment is ensuring transferability across different disaster scenarios (Benson and Ecker, 2020). However, it is important to note that network architectures and pre-training primarily focus on general enhancements and do not incorporate inductive biases tailored to specific disaster events. As a result, they often struggle to effectively address the domain shifts in image data (Candela et al.,

* Corresponding authors.
E-mail addresses: zhuozheng@cs.stanford.edu (Z. Zheng), zhongyanfei@whu.edu.cn (Y. Zhong).

https://doi.org/10.1016/j.rse.2025.114979
Received 23 April 2025; Received in revised form 28 July 2025; Accepted 16 August 2025
Available online 16 September 2025

2009) caused by diverse disaster events and scenarios. In these cross-domain situations, the disparity between new disaster event data and historical training data becomes a critical factor that hinders final accuracy in new disaster events (Liu et al., 2021; Bouchard et al., 2022; Lin et al., 2022).

To reduce this data disparity, transfer learning, especially unsupervised domain adaptation (UDA), is widely used in the generic computer vision community. For building damage assessment, a multi-temporal multi-task problem, (Zheng et al., 2024a) presented a decoupling task modeling to connect generic UDA methods (e.g., AdaptSeg (Tsai et al., 2018) and FDA (Yang and Soatto, 2020)) and transferable building damage assessment. While generic UDA methods reduce the domain gap to some extent, they always need both pre/post-disaster image data for domain adaptation, resulting in a time window problem (Zheng et al., 2024a), i.e., domain adaptation and model inference should be completed in a very limited time period between obtaining post-disaster images and the damage map. The time window problem significantly challenges time-limited disaster response in real-world scenarios. To address this problem, single-temporal change adaptation (STCA) (Zheng et al., 2024a) was proposed, which enables models to achieve better adaptation but with only pre-disaster images in the target domain. This is achieved by constructing pseudo bitemporal damage samples using target pre-disaster images and source post-disaster images.

While STCA shows promising improvements, one important limitation still hinders its reliable application in disaster response: pseudo bitemporal damage samples lack *visual interpretability*. This is because the samples are constructed at the embedding level, making it difficult for human experts to inspect them, identify error sources, and adjust the data distribution to ensure that damage assessment models perform as expected.

In this paper, we propose *Neural Disaster Simulation* (NeDS), a deep disaster generative model, capable of synthesizing realistic post-event images from pre-event images and disaster information. The resulting synthetic bitemporal damage samples are not only visually interpretable but also facilitate model adaptation to new disaster events through synthetic data post-training. NeDS also can avoid the time window problem like STCA, since the damage sample generation process only needs a pre-disaster image and custom disaster information in the target domain. By simulating various disasters on target pre-disaster images, the building damage assessment model can learn more domain-invariant change representations, leading to improved performance in the target domain.

The remainder of this paper is organized as follows. Section 2 specifies the study area and the data. Section 3 describes the details of the neural disaster simulation method. The experimental results and a discussion are provided in Section 4. Section 5 concludes the paper.

2. Data

In this study, we evaluate transferable building damage assessment methods from two perspectives: global-scale statistic evaluation and local-scale real-world case evaluation. We choose 19 disaster events from the xView2 Building Damage Assessment (xBD) dataset (Gupta et al., 2019a) for global-scale statistic evaluation. These disaster events are globally distributed, and their times span from 2011 to 2019. Two recent disaster events, the Libya flooding in 2023 and the Los Angeles wildfire in 2025, are used for case studies, demonstrating the effectiveness of the proposed method in real-world applications.

2.1. Global-scale study area and data

As shown in Fig. 1, globally distributed 19 disaster events were used for global-scale statistic evaluation. These disaster events span four continents: Asia, Europe, North America, and Oceania. There are six disaster types, i.e., flooding, wildfire, volcano, earthquake, and tsunami. To evaluate transferability, we follow the domain split proposed by Zheng et al. (2024a), in which the source domain comprises 9 disaster events, while the target domain consists of the remaining 10 events.

xBD dataset comprises 11,034 bitemporal optical satellite image pairs with 850,736 building instance annotations, covering a total of 45,361.79 km^2. Each image pair contains a pre-disaster image and a post-disaster image, collected from multiple satellite platforms, e.g., GeoEye, WorldView-2, and WorldView-3. Here, Joint Damage Scale (`Non-Damage`, `Minor Damage`, `Major Damage`, `Destroyed`) is used for a unified assessment scale across multiple hazard types, structure categories, and geographical locations. Joint Damage Scale that is based on the HAZUS natural hazard analysis tool (Vickery et al., 2006), the Kelman scale (Kelman, 2003), and the EMS-98 scale (Grünthal, 1998). The Joint Damage Scale was created with the help of the National Aeronautics and Space Administration (NASA), the California Department of Forestry and Fire Protection (CAL FIRE), the Federal Emergency Management Agency (FEMA), and the California Air National Guard.

2.2. Local-scale study area and data

Los Angeles Wildfire (2025): In January 2025, a series of devastating wildfires swept through Southern California, severely impacting the Los Angeles metropolitan area. Exacerbated by drought conditions, low humidity, accumulated vegetation, and hurricane-force Santa Ana winds, the fires spread rapidly, causing widespread destruction. The disaster led to mass evacuations and extensive damage to homes and infrastructure. Among these wildfires, the Eaton Fire and Palisades Fire were the most destructive, ranking among the most severe in California's history.

We acquired satellite imagery covering areas affected by the Eaton and Palisades fires from multiple sources. For the Eaton Fire, post-event imagery was collected by the GeoEye-1 (GE01) satellite on January 10, 2025, with an off-nadir angle of 7.36° and an original spatial resolution of 0.42 m. Pre-event imagery was obtained from Bing Maps. Both pre- and post-event images were upsampled to a spatial resolution of 0.30 m and cropped to a uniform size of 32,276 × 40,832 pixels. For the Palisades Fire, post-event imagery was also collected by the GE01 satellite on January 10, 2025, with an off-nadir angle of 26.1° and a native spatial resolution of 0.50 m. These images were resampled to 0.30 m and cropped to 22,554 × 30,904 pixels. Pre-event imagery was again sourced from Bing Maps and matched in size to the corresponding post-event images. The ground truth data was initially derived from Microsoft's assessments (https://data.humdata.org/dataset/eaton-fire-altadena-damage-assessment-from-1-10) for Eaton Fire and https://data.humdata.org/dataset/palisades-fire-building-damage-assessment for Palisades Fire—and was subsequently refined by domain experts using the Joint Damage Scale.

Nigeria Flooding (2025): The pre-event and post-event imagery were captured by the GE01 satellite on October 9, 2023, and June 2, 2025, respectively, with off-nadir angles of 25.0° and 30.9°, and a native spatial resolution of 0.43 m and 0.53 m. Following the aforementioned pipeline, the images were resampled to a 0.30 m resolution and subsequently cropped to 10,125 × 17,827 pixels to cover Mokwa, Nigeria. The ground truth data was initialized from Google building footprint, UNOSAT's product (https://unosat.org/products/4138), and ChangeOS (Zheng et al., 2021) model predictions. The labels of the main damaged regions were then refined by experts.

Ground truth quality control: We adopted a two-stage labeling strategy for quality control. First, we initialized the annotations using existing damage assessment products and ChangeOS model predictions to provide a baseline. Then, expert annotators manually reviewed and refined these initial labels by comparing them against high-resolution post-disaster imagery. In cases where the damage level could not be confidently determined, due to image ambiguity, occlusion, or limited visual cues, annotators were instructed to mark the instance as

Fig. 1. Global-scale study area and domain split for transferable building damage assessment. There are 19 disaster events, globally distributed across four continents, spanning the period from 2011 to 2019. The bottom three are local study areas in the 2025 Los Angeles wildfire event and Nigeria flooding event.

uncertain. These ambiguous cases were excluded from quantitative evaluation to ensure that reported performance metrics reflect only reliably labeled samples. This combination of automated initialization and expert-guided refinement helped improve both the efficiency and reliability of the final ground truth labels.

3. Methodology

3.1. Preliminary: diffusion models

Diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) are a kind of generative model, showing great success in generating high-quality images in recent years. The principle is learning to reverse a noising process (also referred to as a forward diffusion process) that progressively perturbs a data sample into its noisy version (Lu and Song, 2025). The noising process always perturbs underlying data distribution to a prior distribution (e.g., standard Gaussian distribution). In this way, diffusion models can generate samples from this prior distribution by iterating the learned denoising function.

3.2. Neural disaster simulation

As shown in Fig. 2, given a pre-event image I_{pre}, its object (e.g., building) mask m_{pre}, disaster type c, and disaster intensity, NeDS generates a post-event image I_{post} with the corresponding post-event damaged mask m_{post}. Following the principle of the generative probabilistic change modeling (Zheng et al., 2025), our NeDS model is built upon three key components: a stochastic damage process, disaster information injection, and a latent diffusion model.

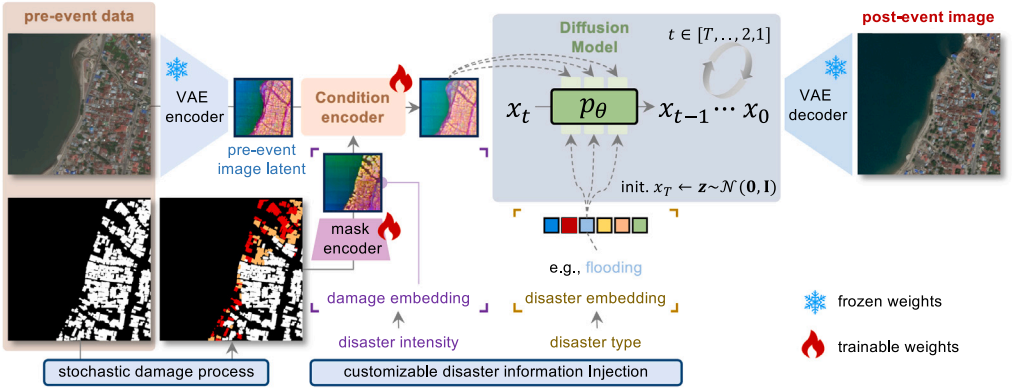

Fig. 2. Our **Neural Disaster Simulation** (NeDS) framework, aims at synthesizing the post-event image from the pre-event image and customizable disaster information (in this case, the intensity and type of the disaster are used). In NeDS, the stochastic damage process is first used to yield a random post-event damage mask from a pre-event building mask. The condition damage generation is then used to edit the pre-event image in latent space using a given post-event damage mask and disaster type, yielding a post-event image with desired damages.

Stochastic damage process: This module samples a post-event damaged mask $m_{\text{post}} \in \{0, 1, 2, 3, 4\}^{h \times w}$ based on a pre-event object mask $m_{\text{pre}} \in \{0, 1\}^{h \times w}$. To generate diverse synthetic damage samples, we introduce stochasticity into the damage process of each object in the pre-event object mask m_{pre}. Specifically, each object is probabilistically assigned one of four damage statuses: `Non-Damage`, `Minor Damage`, `Major Damage`, and `Destroyed` according to a customizable 4-dimensional categorical distribution.

Disaster information injection: To enable fine-grained control over post-event image generation, we leverage two categories of disaster information: damage intensity and disaster type. We represent the damage intensity as a 4-dimensional categorical distribution Categorical ($[p_i]_{i=1}^4$) that describes the probabilities of four damage statuses, where the minimum and maximum intensities are defined as Categorical([1., 0., 0., 0.]) and Categorical([0., 0., 0., 1.]), respectively. The disaster intensity impacts the stochastic damage process, where a larger intensity leads to more damaged objects sampled from the categorical distribution. We represent the disaster type $c \in R^{n_{\text{types}} \times d_{\text{model}}}$ as a set of learnable query embeddings. Here the number of disaster types n_{types} is 6 for the xBD dataset and d_{model} is the embedding dimensionality depending on specific model architecture. These embeddings are added to the time embeddings of the diffusion model as conditioning inputs, enabling disaster-type-controllable post-event image generation.

Conditional latent diffusion model: This module simulates a disaster event on the pre-event image to generate the corresponding post-event image. Here we design a conditional latent diffusion model to achieve this goal. This model consists of a pre-trained variational autoencoder (VAE) to project the image into a latent space, a condition encoder, a mask encoder, and a velocity prediction network p_θ.

Firstly, the condition x_c is derived using a condition encoder that takes as input the pre-event image latent $x_{\text{pre}} \sim F_{\text{enc}}(I_{\text{pre}})$ and the damage embedding x_{dam}, where F_{enc} is a VAE encoder (Kingma et al., 2013; Rombach et al., 2022). The condition encoder consists of a 3×3 convolutional layer that projects the pre-event image latent x_{pre} to channel dimension d_{model}. The projected latent is then concatenated with the damage embedding x_{dam} along the channel axis, followed by a pointwise convolutional layer and two 3×3 convolutional layers to fuse the features and project them back to the channel dimension d_{model}. To obtain the damage embedding x_{dam} with the same spatial resolution and channel dimension as x_{pre}, we adopt a mask encoder consisting of three 3×3 conv layers with a stride of 2 to perform 8× spatial downsampling and expands the channel to d_{model}.

We then employ a diffusion model to generate the post-event image from noise, guided by the aforementioned condition x_c and disaster type c. Specifically, we adopt a standard denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) forward process: $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1-\alpha_t} z, z \in \mathcal{N}(0, I)$, where the constant α_t is the hyperparameter of noise scheduler. Our model learns a reverse process $p_\theta(x_{t-1}|x_t, x_c, c)$ conditioned on x_c and c, enabling the synthesis of realistic, disaster-aware damage patterns specific to each disaster type. Starting from a noise latent $\hat{x}_T \sim \mathcal{N}(0, I)$, the predicted post-event image latent $\hat{x}_0$ is recovered by iteratively applying the learned reverse process over T steps, where T denotes the total number of reverse process iterations. The final post-event image I_{post} is then obtained by decoding the predicted latent $\hat{x}_0$ using a VAE decoder F_{dec}, i.e., $I_{\text{post}} = F_{\text{dec}}(\hat{x}_0)$. The code of NeDS is available at https://github.com/Z-Zheng/pytorch-change-models.

Training objective: We parameterize our diffusion model by v-prediction (Salimans and Ho, 2022) for more stable training; thus we have the velocity vector $v_t = \sqrt{\alpha_t} z - \sqrt{1-\alpha_t} x_0$ as our regression target. We optimize the velocity prediction network by minimizing a mean squared error loss between the target and our velocity prediction $v_\theta(x_t, x_c, c)$:

$$L_\theta := \mathbb{E}_{t,c,x_c} \| v_t - v_\theta(x_t, x_c, c) \|^2 \quad (1)$$

3.3. Training damage assessment model with NeDS

We apply NeDS to transferable building damage assessment in the following three steps:

Training a NeDS model on source domain: For transferable building damage assessment, we train the NeDS model using 6,369 pre- and post-disaster image pairs from nine disaster events in the source domain of the xBD dataset. For faster convergence and improved generation quality, we utilize a pre-trained Stable Diffusion 2.1 (SD2.1) model (Rombach et al., 2022) as the velocity prediction network. To effectively incorporate the condition x_c to this pre-trained diffusion model, we encode this condition using SD2.1's encoder and then add the encoded features into each decoder block of SD2.1, motivated by the successful practice of ControlNet (Zhang et al., 2023). We randomly crop the original image to 512×512 pixels, followed by D4 dihedral group transformations (Dummit et al., 2004) for training data augmentation. We train our model for 400 epochs using the AdamW optimizer with a constant learning rate of 1e-4 and zero weight decay, with a total batch size of 64 distributed across 8 NVIDIA H100 GPUs. We freeze the pre-trained VAE during training due to its exceptional reconstruction capability on satellite images.

Generating a synthetic damage dataset on target domain: We use 2,799 pre-event images from ten disaster events of the target domain in the xBD dataset for subsequent disaster simulation. For each target pre-event image, we first collect its open building footprints as

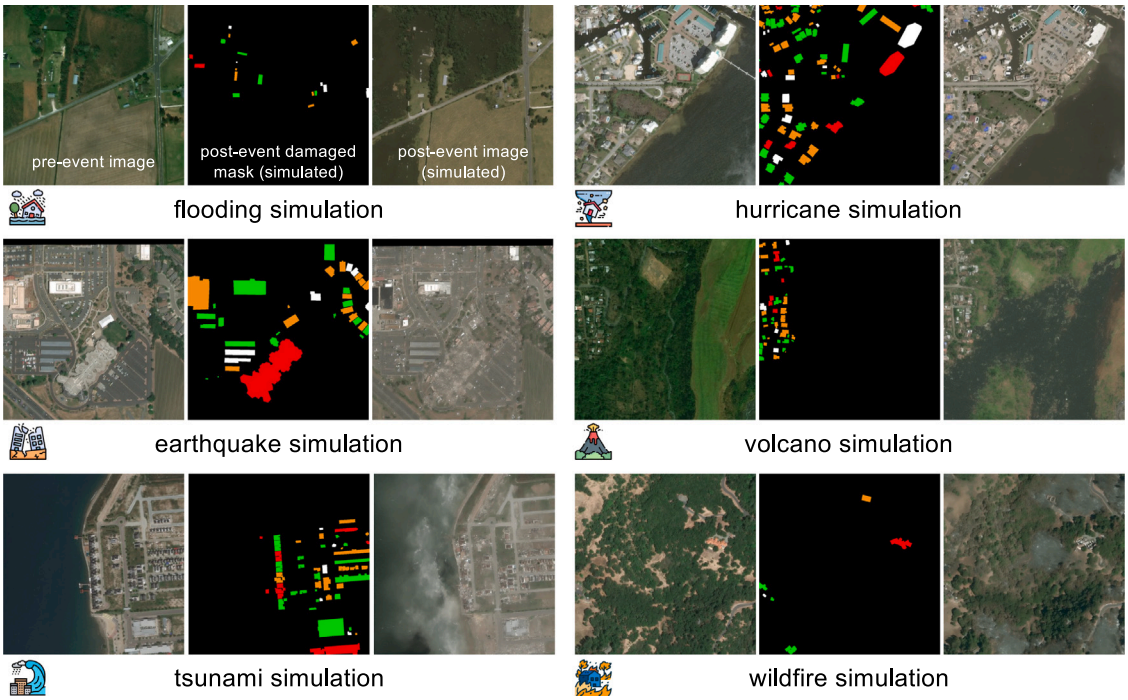

Fig. 3. Multi-hazard disaster simulations using our NeDS model. Each panel shows a different disaster type, including flooding, hurricane, earthquake, volcano, tsunami, and wildfire. From left to right, we present the pre-event satellite image, the simulated post-event damage mask, and the simulated post-event satellite image. Legend: Minor damage, Major damage, Destroyed.

m_{pre}. For those images without open building footprints, we used a pre-trained ChangeOS model (Zheng et al., 2021) to predict their binary building masks. Meanwhile, we randomly assign a disaster type to each target pre-event image to enhance damage data diversity and simulate real-world scenarios. For the stochastic damage process, we set the disaster intensity as Categorical([0.1, 0.4, 0.4, 0.1]) to simulate a higher frequency of minor damages and major damages, which are typically more difficult to recognize for conventional building damage assessment models. Constructing these harder examples is more beneficial for training effective models on synthetic data. To accelerate the sampling process of our NeDS model, we adopt the UniPC sampler (Zhao et al., 2023) with $T = 30$ steps. This results in a synthetic damage dataset comprising 2,799 bitemporal image pairs with corresponding damage labels.

Training a building damage assessment model: To effectively leverage these synthetic damage data, we adopt a mixed training strategy that integrates both target synthetic damage data and source real damage data. For each mini-batch of $2n$ samples, n samples are drawn from the synthetic target damage dataset, and the remaining n samples are taken from the real source damage dataset. To be consistent with the setting of Zheng et al. (2024a) for a fair comparison, we adopt ChangeOS architecture (Zheng et al., 2021) for building damage assessment. For source real damage data, we adopt a combination of binary cross-entropy loss and Tversky loss (Salehi et al., 2017) ($\alpha = 0.9$) for building localization, and a combination of cross-entropy loss and Tversky loss ($\alpha = 0.7$ for non-damage, minor damage, and major damage categories) for damage classification. For target synthetic damage data, we optimize only the damage classification task using a combination of cross-entropy loss and Tversky loss ($\alpha = 0.9$ for non-damage, minor damage, and major damage categories). We train ChangeOS for 40k iterations using the AdamW optimizer with a "poly" ($\gamma = 0.9$) learning rate of 6e-5 and a weight decay of 0.01, with a total batch size of 32 ($n = 16$) distributed across 8 NVIDIA H100 GPUs. We randomly crop each image to 512×512 pixels, followed by D4 dihedral group transformations for data augmentation during training.

3.4. Evaluation metrics

Standard xView2 metrics (Gupta et al., 2019a) are used to evaluate the model performance. The building localization score F_1^{loc}, computed as a standard pixel-based F_1 score, evaluates the accuracy of pre-event building segmentation. The damage classification score F_1^{dam}, defined as a harmonic mean of standard pixel-based F_1 scores across four damage classes, measures post-event damage classification performance. The average score F_1^{avg}, defined as $0.3F_1^{\text{loc}} + 0.7F_1^{\text{dam}}$, evaluates overall performance of the model.

4. Results and discussion

4.1. Multi-hazard disaster simulations

To provide a proof-of-concept of NeDS, we perform multi-hazard simulation experiments across various geographic and environmental contexts. As shown in Fig. 3, we present case-wise simulations of six distinct natural hazards—flooding, hurricane, earthquake, volcanic eruption, tsunami, and wildfire, each applied to separate pre-event satellite images. For each scenario, NeDS first simulates a post-event damage mask and synthesizes a layout-consistent post-event image that reflects hazard-specific patterns. Our model captures inundation zones in flooding, structural destruction in earthquake and hurricane scenarios, vegetative scorching in wildfires, and volcanic ash coverage in eruption-affected regions. These outputs are consistent with known physical and spectral responses of different hazard types in high-resolution satellite imagery, suggesting that the model effectively learns the spatial semantics of disaster impacts across varying land cover and topographic settings. Fig. 4 further investigates the controllability of our NeDS model by conditioning simulations on multiple hazard types at a fixed location. The resulting post-event images exhibit visually distinct transformations, such as increased water turbidity in flooding scenarios, vegetation degradation in wildfire simulations, and soil displacement in earthquake conditions, demonstrating that our NeDS model can generate hazard-specific imagery conditioned on the disaster type.

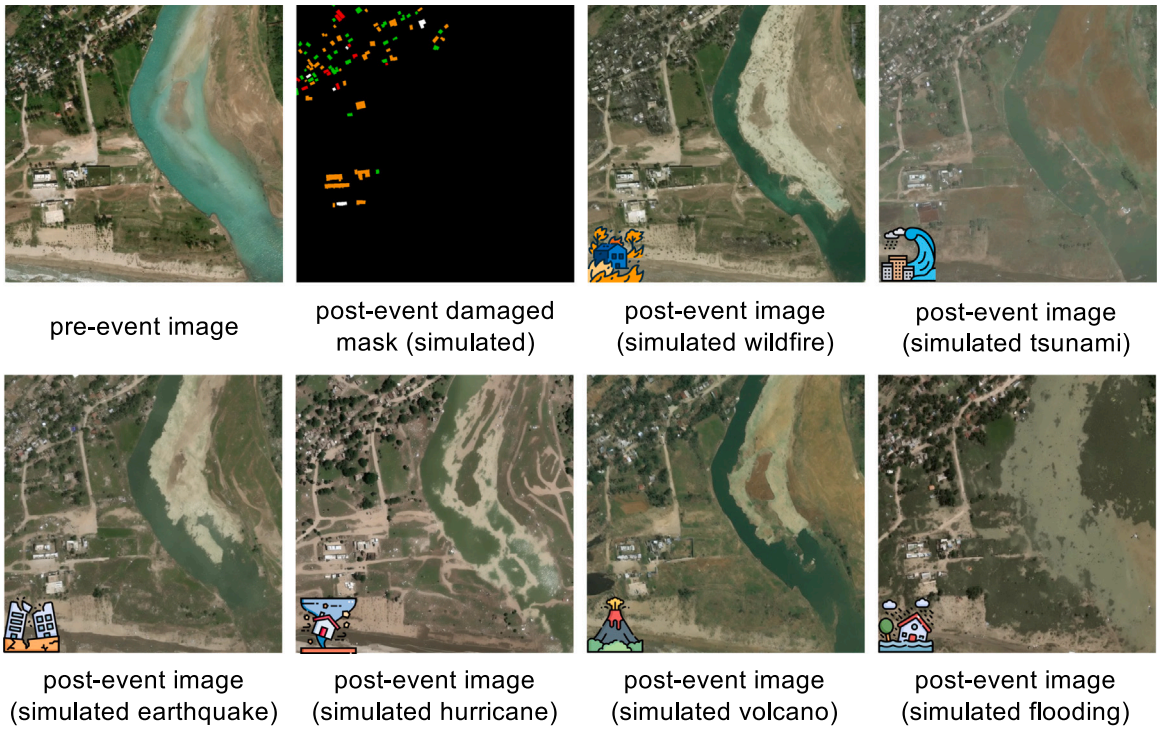

Fig. 4. Controllability of our NeDS model under different disaster types. Starting from a shared pre-event image, our model simulates damage masks and post-event images conditioned on six different disaster types: wildfire, tsunami, earthquake, hurricane, volcano, and flooding. This demonstrates the controllable nature of NeDS in generating diverse and type-specific disaster impacts.

Table 1
Evaluation of post-event image generation. The metrics include Fréchet Inception Distance (FID), Inception Score (IS), Kernel Inception Distance (KID), and Human Expert Score (HES).

Method	FID(↓)	IS(↑)	KID(↓)	HES(↑)
ControlNet	60.64	5.50 ± 0.16	.0261±.0010	40.2%
NeDS (ours)	40.78	6.00 ± 0.29	.0097±.0006	59.8%

Table 1 presents a comparative evaluation of post-event image generation using three widely adopted metrics: Fréchet Inception Distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016), Kernel Inception Distance (KID) (Bińkowski et al., 2018), and our designed Human Expert Score (HES). HES is derived from expert preference voting. For each pair of generated images, human experts vote for the one they perceive as more realistic and higher quality, and the final score is computed as the percentage of votes in favor of a given method. Compared to the ControlNet baseline with same pre-trained SD 2.1, our proposed method NeDS achieves significant improvements across all metrics. This benefits from our better condition mechanism design. Specifically, NeDS achieves a lower FID of 40.78 versus 60.64, indicating better alignment with real post-disaster image distributions. It also attains a higher IS of 6.00 ± 0.29, reflecting enhanced visual diversity and quality, as well as a lower KID of 0.009 ± 0.0006, further confirming improved fidelity. Most notably, NeDS obtains a substantially higher HES of 59.8% compared to 40.2% for ControlNet, demonstrating that domain experts consistently perceive NeDS-generated images as more realistic and semantically accurate.

4.2. Transferable building damage assessment

In this section, we evaluate the building damage assessment performance of models trained on synthetic damage data generated by our NeDS model. To provide performance references, we benchmark two categories of methods: unsupervised change detection-based approaches and domain adaptation-based techniques. The first category of methods is capable of assessing building damage when target post-event training images are unavailable for adaptation. Most domain adaptation-based methods require target-domain bitemporal training images for adaptation, except STCA (Zheng et al., 2024a), which only requires target pre-event images.

Baseline: ChangeOS is a deep object-based semantic change detection framework, which is suitable for assessing building damage because it integrates object-based image analysis for object semantic consistency and deep representation learning. We trained six ChangeOS variants in the source domain of the xView2 dataset and evaluated them on the target domain without any adaptation. The results presented in the source-only section of Table 2 are used as baselines for the subsequent comparative analysis. We use the ResNet-18-based ChangeOS variant as a common reference to illustrate various improvements, such as architectural enhancements and adaptation strategies.

Comparison to unsupervised change detection-based damage assessment methods: We implemented this category of methods through a pipeline of unsupervised change detection and post-event image-based damage assessment. The unsupervised change detection methods adopt visual foundation model-based deep change vector analysis (DCVA), specifically using DINOv2 (Oquab et al., 2023) and SAM (Kirillov et al., 2023), which have been demonstrated as state-of-the-art unsupervised change detectors (Zheng et al., 2024b). The post-event image-based damage assessment is performed using the damage classification network f_{dam} of ChangeOS. The results presented in the first two sections of Table 2 indicate that this category of methods is unsuitable for transferable building damage assessment. Specifically, using f_{dam} of ChangeOS yields 34.4% F_1^{avg}, dropping 3.7% from the baseline (38.1% F_1^{avg}). After applying unsupervised change detection, the performances are further reduced by 4.6% and 4.1%. This suggests that these unsupervised change detection methods struggle to capture building damage, resulting in an increase in intermediate errors. Compared with this category of methods, our NeDS model

Table 2
Benchmark comparison on the xView2 `holdout` split. source domain (`tier3`)→ target domain (`train`).

Method	Backbone	F_1^{avg}(%)	F_1^{loc}(%)	F_1^{dam}(%)	Damage F_1 per class (%)				#Params Trainable/All	Requirement Target-domain
					Regular	Minor	Major	Destroyed		
source-only, post-event image-based damage assessment										
f_{dam} of ChangeOS	ResNet-18	34.4	83.0	13.6	71.3	12.5	5.8	36.3	17.9M	-
source-only, unsupervised change detection with post-event image-based damage assessment										
+ DINOv2-DCVA	ResNet-18	29.8	83.3	6.9	77.4	4.5	3.1	38.3	17.9M/321.2M	-
+ SAM-DCVA	ResNet-18	30.3	83.3	7.6	76.0	4.6	3.7	37.9	17.9M/107.6 M	-
source-only										
ChangeOS	ResNet-18	38.1	82.4	19.1	77.4	13.3	9.2	72.4	24.8M	-
ChangeOS	ResNet-50	39.4	82.9	20.8	78.2	15.7	9.8	70.2	39.0M	-
ChangeOS	Swin-T	42.8	83.3	25.4	77.3	15.6	15.1	71.7	41.4M	-
ChangeOS	MiT-B1	43.1	84.0	25.6	79.0	19.5	12.7	73.9	26.8M	-
ChangeOS	ConvNext-T	43.4	83.7	26.1	77.6	18.6	13.8	71.6	41.7M	-
w/ foundation model										
ChangeOS	X-Scale MAE/ViT-L	42.1	80.0	25.8	73.3	16.7	15.2	63.1	323.5M	-
ChangeOS	CACo/ResNet-50	43.1	80.7	26.9	78.6	18.4	14.9	70.4	39.0M	-
ChangeOS	Scale-MAE/ViT-L	43.2	75.4	29.4	70.4	14.8	26.8	58.4	323.5M	-
ChangeOS	GASSL/ResNet-50	44.3	81.3	28.5	79.3	16.3	18.9	73.5	39.0M	-
ChangeOS	SeCo/ResNet-50	44.4	79.1	29.5	74.6	18.3	18.9	68.0	39.0M	-
ChangeOS	SAM/ViT-B	46.7	84.4	30.5	78.8	20.5	17.9	74.2	17.0M/105.9M	-
ChangeOS	SatMAE++/ViT-L	47.7	81.9	33.1	76.4	19.2	24.2	69.2	323.5M	-
ChangeOS	SatMAE/ViT-L	48.6	83.8	33.5	79.2	18.2	26.6	70.9	323.5M	-
w/ adaptation		Reference: ChangeOS with ResNet-18 (source-only)								
+ DDC	ResNet-18	37.9(↓0.2)	82.7(↑0.3)	18.7(↓0.4)	77.0	12.4	9.4	71.1	24.8M	pre. and post.
+ DAN	ResNet-18	38.0(↓0.1)	82.7(↑0.3)	18.8(↓0.3)	77.1	13.1	9.2	71.1	24.8M	pre. and post.
+ DAFormer	MiT-B1	38.8(↑0.7)	84.8(↑2.4)	19.1	82.3	10.4	11.4	73.3	54.5M	pre. and post.
+ TransNorm	ResNet-18	39.1(↑1.0)	82.5(↑0.1)	20.5(↑1.4)	76.1	12.5	11.3	69.4	28.0M	pre. and post.
+ FADA	ResNet-18	39.6(↑1.5)	82.7(↑0.3)	21.2(↑2.1)	77.2	15.8	10.1	71.0	27.6M	pre. and post.
+ AdaptSeg	ResNet-18	39.9(↑1.8)	82.8(↑0.4)	21.6(↑2.5)	77.2	16.5	10.2	71.1	28.0M	pre. and post.
+ FDA	ResNet-18	41.8(↑3.6)	82.2(↓0.2)	24.6(↑5.5)	77.5	15.5	13.9	72.3	24.8M	pre. and post.
+ GDA ($r = 32$)	SatMAE/ViT-L	46.7(↑8.6)	85.1(↑2.7)	30.2(↑11.1)	81.2	18.8	19.1	69.9	32.0M/336.1M	pre. and post.
+ GDA ($r = 224$)	SatMAE/ViT-L	50.4(↑12.3)	84.0(↑1.6)	36.0(↑16.9)	79.7	18.5	33.4	69.1	107.4M/411.6M	pre. and post.
+ STCA	ResNet-18	46.5(↑8.4)	81.7(↓0.7)	31.4(↑12.3)	74.6	22.4	18.2	69.3	24.8M	pre.
+ STCA	SAM/ViT-B	47.6 (↑9.5)	85.0(↑2.6)	31.5(↑12.4)	73.5	22.7	18.0	72.5	17.0M/105.9M	pre.
+ STCA	ConvNext-T	48.8(↑10.7)	84.1(↑1.7)	33.7(↑14.6)	74.8	27.7	17.9	73.9	41.7M	pre.
+ STCA	MiT-B1	49.5(↑11.4)	83.9(↑1.5)	34.8(↑15.7)	77.5	24.1	21.5	71.6	26.8M	pre.
+ STCA	Swin-T	50.1(↑12.0)	82.8(↑0.4)	36.1(↑17.0)	78.0	26.3	21.6	73.6	41.4M	pre.
+ ControlNet (SD)	SAM/ViT-B	51.3(↑13.2)	85.3(↑2.9)	36.6(↑17.5)	79.1	20.1	30.5	72.2	104.1M	pre.
+ NeDS (ours)	Swin-T	50.7(↑12.6)	86.4(↑4.0)	35.3(↑16.2)	73.4	22.6	24.6	69.1	41.4M	pre.
+ NeDS (ours)	SAM/ViT-B	54.4(↑16.3)	87.8(↑5.4)	40.1(↑21.0)	81.9	22.3	33.9	74.6	104.1M	pre.

improves the model transferability via synthetic damage data post-training. The resulting model remains end-to-end, with no intermediate errors introduced.

Comparison to foundation model-based methods: Table 2 shows that replacing conventional CNN backbones with self-supervised geospatial foundation models, GASSL (Ayush et al., 2021), SeCo (Manas et al., 2021), CACo (Mall et al., 2023), SatMAE (Cong et al., 2022), ScaleMAE (Reed et al., 2023), Cross-Scale MAE (Tang et al., 2024), SatMAE++ (Cong et al., 2022), already yields noticeable gains: the best plain ChangeOS variant with SatMAE/ViT-L reaches an 48.6% F_1^{avg}, but stays below 50%. Although applying geospatial domain adaptation (GDA) (Scheibenreif et al., 2024), the performance only gains to 50.4% F_1^{avg}. Applying NeDS on top of foundation model breaks this ceiling. With the same SAM/ViT-B backbone, NeDS lifts F_1^{avg} from 46.7% to 54.4% and F_1^{dam} from 30.5% to 40.1%, while also improving class-wise performance. Within our adaptation framework, we additionally implemented a strong ControlNet baseline based on pre-trained Stable Diffusion to enable a fine-grained comparison. The results suggest that our designed conditioning mechanism is better suited for disaster-scenario imagery.

Comparison to other adaptation methods: For comparison, we extend seven advanced transfer learning methods (DDC (Tzeng et al., 2014), DAN (Long et al., 2015), DAFormer (Hoyer et al., 2022), TransNorm (Wang et al., 2019), FADA (Wang et al., 2020), AdaptSeg (Tsai et al., 2018), FDA (Yang and Soatto, 2020)) to ChangeOS by decoupled task modeling (Zheng et al., 2024a). We also compare with STCA, a state-of-the-art tailored domain adaptive building damage assessment method. STCA utilizes target pre-disaster imagery and source post-disaster imagery to simulate the change process, thereby producing training data, i.e., pseudo bitemporal damage samples that confuse the source and target domains. The results from the *w/ adaptation* section of Table 2 suggest that adaptation through synthetic damage data generated by our NeDS model achieves better F_1^{avg} than previous state-of-the-art methods. Under the same backbone settings (Swin-T and ViT-B from SAM), NeDS achieves performance gains of 0.6% and 4.6% in F_1^{avg} over STCA, respectively. There is a consistent observation that a larger visual model gains more from synthetic data (Zheng et al., 2025). In addition, unlike the pseudo-bitemporal damage samples of STCA, the synthetic bitemporal samples generated by NeDS are visually interpretable, enhancing transparency and trustworthiness in real-world disaster response scenarios. Our optimal adaptation configuration (NeDS and SAM/ViT-B) improves over the source-only baseline by 14.1% F_1^{avg}. More specifically, compared to STCA, NeDS achieves notable improvements in the non-damage (regular), major damage, and destroyed classes, with gains of 9.6%, 7.7%, and 2.5%, respectively. This advantage comes from the ability of NeDS to generate diverse damage samples in underrepresented classes and disaster scenarios, facilitating the learning of more balanced and disaster-invariant representations.

Relation between pixel-based and instance-based assessment: In real-world decision-making scenarios, instance-level assessment is more meaningful. Although the pixel prediction can be converted to instance prediction via object-based postprocessing, the relation between pixel-based and instance-based metrics is still unclear. Here we compute this instance-based F_1 score over 11 top-performing models and plot the relation in Fig. 5. We observe a strong linear correlation ($R^2 = 0.81$, $p < 0.001$) between pixel-based and instance-based F_1 scores across different models. This suggests that pixel-based evaluation can serve as a useful proxy for instance-level performance in most cases.

4.3. Method analysis

We adapt the building damage assessment models using synthetic damage data generated by our proposed NeDS. This adaptation process consists of two main steps: synthetic damage data generation and subsequent training with synthetic data. To better understand this procedure, we analyze two critical hyperparameter choices: (i) the categorical distribution used in synthetic damage data generation, and (ii) the ratio of simulated to real data within each training mini-batch under the mixed training strategy. We also report the model's training and inference efficiency for disaster response.

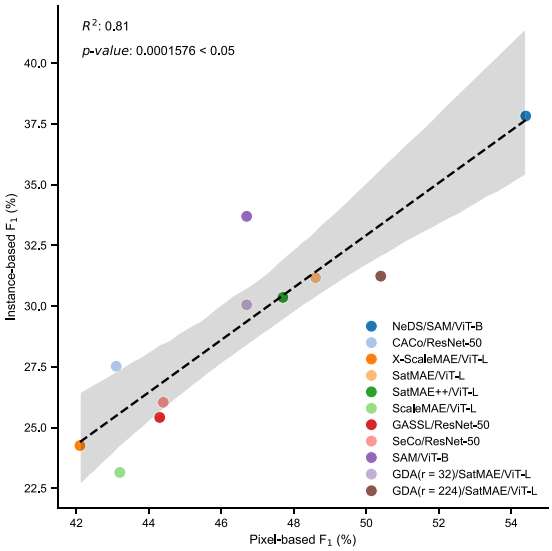

Fig. 5. Correlation between pixel-based and instance-based F_1 scores across different models. Each dot represents a model, color-coded by its architecture (see legend). The black dashed line shows the linear regression fit ($R^2 = 0.81$, $p < 0.001$), with the shaded region indicating the 95% confidence interval. A strong positive correlation suggests that models achieving higher pixel-level performance generally also perform better at the instance level.

Categorical distribution: We incorporate disaster intensity information by adjusting the categorical distribution, which can be customized for specific disaster events. Our primary objective here is to identify an effective default setting that significantly improves the transferability of the building damage assessment model for most cases. We discuss three representative options. The first is a uniform distribution, assigning equal probabilities across all building statuses. The second emphasizes three specific damage levels (i.e., minor damage, major damage, and destroyed), assigning each an equal probability ($p = 0.3$). The third option focuses specifically on the two damage classes (minor and major damage) that are most challenging to recognize, i.e., 0.4 for minor and major damage, 0.1 for the remaining two. Fig. 6 suggests that the second categorical distribution with probabilities [0.1, 0.3, 0.3, 0.3] clearly yields the highest overall average F_1^{avg} (54.38%), outperforming both the uniform distribution (50.82%) and the Categorical([0.1, 0.4, 0.4, 0.1]) distribution (52.31%). Thus, we recommend Categorical([0.1, 0.3, 0.3, 0.3] as the default setting. All options surpass the previous state-of-the-art (STCA: 47.6%), indicating the effectiveness of our NeDS adaptation. We also draw several insights from the per-class performances. All choices improve significantly over the baseline across four classes. They improve the baseline from 78.8% to 81.1%~82.1% for the regular class. Minor damage is a hard-to-recognize category. A simple uniform distribution brings no performance gain for this class. However, when shifting the focus more toward three damage classes, NeDS adaptation delivers non-trivial improvements, achieving performance comparable to the previous state-of-the-art. The major damage class shows the most significant performance improvement among the four categories, up to 16%. This suggests that equal emphasis on all three damage levels is critical for modeling this visually subtle, hard-to-detect class. For destroyed class, uniform sampling achieves the best score (76.0), slightly above other two categorical variants (~74.7). This is because destroyed buildings exhibit large, unambiguous visual cues, making over-emphasis unnecessary, and simple class balancing is sufficient.

Ratio of simulated to real data: We guide the adaptation training process by adjusting the ratio of simulated to real data within each training mini-batch. Here, we focus on three representative cases: simulated-to-real data ratios of 1:2, 1:1, and 2:1. Fig. 7 shows that a balanced 1:1 mix of simulated and real data yields the best overall performance, peaking at 52.31% F_1^{avg}, 87.75% F_1^{loc}, and 37.12% F_1^{dam}, while per-class results reveal that regular and minor damage classes also maximize at this ratio. Interestingly, the major damage class benefits most from a lower simulation ratio (1:2), rising to 27.24%, whereas the destroyed class improves with more synthetic data, reaching 75.45% at a 2:1 ratio. These findings suggest that, for a general default, a 1:1 simulated-to-real data ratio optimally balances overall and per-class performance. If one aims to target the major-damage class specifically, a lower simulation ratio (e.g., 0.5) is beneficial; for destroyed buildings, higher simulation ratios (e.g., 2.0) yield marginal gains.

Model training and inference efficiency: The training of the NeDS model required 14.5 h on 8 NVIDIA H100 GPUs. Generating the synthetic damage datasets took 18, 6, and 3 min for the Eaton Fire, Palisades Fire, and Nigeria Flooding events, respectively, using 10 NVIDIA A4000 GPUs. The adaptation training for the ChangeOS variant of SAM/ViT-B took 1.8, 0.4, and 3.5 h for the Eaton Fire, Palisades Fire, and Nigeria Flooding events, respectively, on 8 NVIDIA A4000 GPUs. Generating synthetic damage samples and performing adaptation training is highly time-efficient for each local disaster event and can be further accelerated using higher-end GPUs (e.g., H100). Moreover, all training and inference processes can be completed prior to the disaster event. These results confirm the practicality of NeDS for real-world disaster response.

Theoretical analysis: To delve into the effectiveness of our NeDS from a theoretical perspective, we first leverage the classical generalization bound (Ben-David et al., 2006, 2010) for UDA decomposes the target-domain error $\epsilon_T(h)$ of a hypothesis $h \in \mathcal{H}$.

$$\epsilon_T(h) \leq \epsilon_S(h) + d_{\mathcal{H}}(P_S, P_T) + C \tag{2}$$

where C is a constant for the complexity of hypothesis space and the risk of an ideal hypothesis for both domains. $d_{\mathcal{H}}(P_S, P_T)$ is the $\mathcal{H}$-divergence between source domain P_S and target domain P_T. $\epsilon_S(h)$ denotes the source-domain error. NeDS reduces the target-domain error $\epsilon_T(h)$ primarily by lowering the domain divergence $d_{\mathcal{H}}(P_S, P_T)$. Because the original source images and damage labels remain unchanged and are still used during adaptation, the source-domain error $\epsilon_S(h)$ does not deteriorate NeDS synthesizes post-disaster images and their corresponding labels, conditioned on real pre-disaster target imagery, creating a synthetic target distribution $\tilde{P}_T$. Training on the combined set of source data and $\tilde{P}_T$ shifts the overall training distribution toward the true target domain P_T, making it harder for a domain discriminator to distinguish the two domains and thereby reducing $d_{\mathcal{H}}(P_S, P_T)$. Nevertheless, NeDS is not without limitations. If the simulated post-disaster imagery fails to capture the true physical effects of a particular hazard, the synthetic target distribution $\tilde{P}_T$ may drift away from both P_S and P_T, potentially increasing the divergence term $d_{\mathcal{H}}(P_S, P_T)$. When the conditional label distributions differ substantially, divergence term will likewise increase. This is why the categorical distribution affects the final performance more (see Fig. 6). These failure modes highlight the need for higher-fidelity simulators, diverser label distribution, complementary strategies (such as physics-informed priors or multi-modal conditioning) to keep $d_{\mathcal{H}}(P_S, P_T)$ low and maintain NeDS's effectiveness across diverse disaster scenarios.

4.4. Application on local-scale study areas

To demonstrate the effectiveness of our NeDS in real-world disaster scenarios, we conducted transferable building damage assessment applications on the Los Angeles wildfire of 2025, which comprised two major events: the Eaton fire and the Palisades fire, which are now among the most destructive wildfires in California history.

Image preprocessing: Post-event satellite imagery used in the case studies was acquired from the GeoEye-1 (GE01) satellite, while the corresponding pre-event imagery was obtained from Bing Maps. All

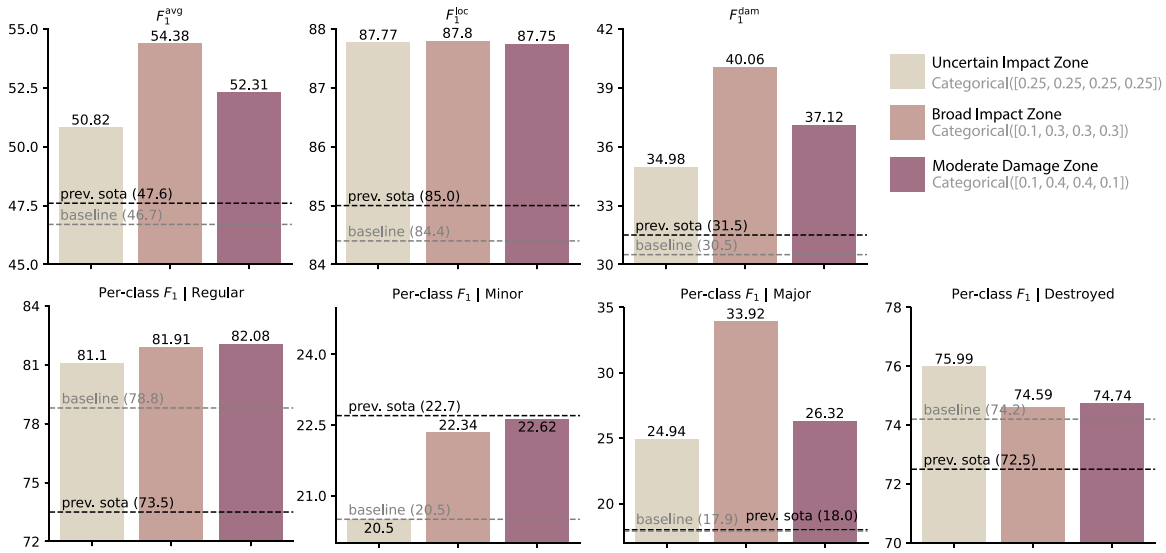

Fig. 6. Ablation study: Categorical distribution. The baseline model is the source-only ChangeOS. The previous state-of-the-art (prev. sota) is ChangeOS with STCA. The backbones used in all these methods are SAM/ViT-B for consistency. In some per-class sub-figures, the previous sota performs worse than the baseline. This is because the sota is determined based on the overall performance F_1^{avg} rather than individual class performance.

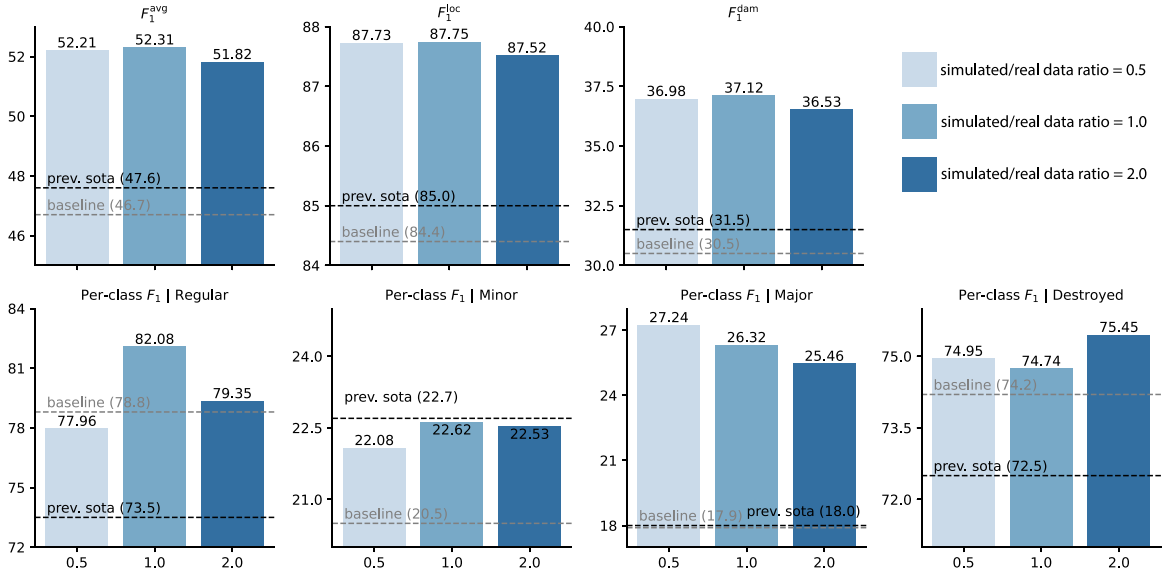

Fig. 7. Ablation study: Ratio of simulated and real data in each training mini-batch under the mixed training strategy. The baseline model is the source-only ChangeOS. The previous state-of-the-art (prev. sota) is ChangeOS with STCA. The backbones used in all these methods are SAM/ViT-B for consistency. In some per-class sub-figures, the previous sota performs worse than the baseline. This is because the sota is determined based on the overall performance F_1^{avg} rather than individual class performance.

images were preprocessed using atmospheric compensation, dynamic range adjustment, and pan-sharpening. To account for differences in spatial resolution, all images were resampled to 0.30 m, ensuring spatial alignment between pre- and post-event image pairs. No additional image enhancement was applied to minimize processing time, which is critical in time-sensitive disaster response scenarios.

Model training, adaptation, and inference: We first trained a ChangeOS (Zheng et al., 2021) model variant with a SAM/ViT-B (Kirillov et al., 2023) as the backbone network, using xView2 `train` and `tier3` sets. This choice is based on our empirical observation that the variant with SAM/ViT-B benefits the most from the synthetic damage data generated from NeDS. The obtained model can be seen as a source model. We then leverage a well-trained NeDS model to generate post-event images and damage labels from pre-disaster satellite images of the Los Angeles wildfire event. We adopt our proposed mixed training strategy to fine-tune the source model on synthetic damage data and xView2 datasets as adaptation training, reducing the domain gap between disaster events in xView2 and the Los Angeles wildfire event. Similarly to STCA, this adaptation can be performed at any time, as it also does not require post-disaster imagery for adaptation training. After completing the model adaptation, the target model was directly applied to assess building damage using the pre/post-disaster image pair as soon as the post-disaster image became available. Pre- and post-event satellite imagery products are typically large in size, resulting in high memory demands that can exceed the capacity of conventional GPUs, including the H100 with 80 GB of memory. To address this issue, we adopt an overlapped sliding window inference strategy, where the window has a size of 1,024 × 1,024 with a stride of 512.

Accuracy assessment: The accuracy of this application is measured using the standard xView2 metric based on manually annotated

Table 3
Building damage assessment performance comparison between ChangeOS, ChangeOS + STCA, and ChangeOS + our NeDS for Eaton fire, Palisades fire, and Nigeria flooding datasets. We focus solely on evaluating damage classification performance, as pre-disaster building footprints are widely available.

Method	F_1^{dam}(%)	Damage F_1 per class (%)				Requirement
		Regular	Minor	Major	Destroyed	Target-domain
Eaton Fire (2025)						
ChangeOS (source-only)	27.9	99.4	14.2	19.0	96.7	-
w/ adaptation						
+ STCA	65.7	99.7	47.7	34.0	97.5	pre.
+ NeDS (ours)	69.3	99.8	64.7	45.5	98.0	pre.
Palisades Fire (2025)						
ChangeOS (source-only)	70.6	97.4	56.1	54.5	97.7	-
w/ adaptation						
+ STCA	77.1	98.7	73.0	56.2	97.9	pre.
+ NeDS (ours)	85.0	99.1	81.4	68.8	98.5	pre.
Nigeria Flooding (2025)						
ChangeOS (source-only)	0.0	32.8	-	0.0	2.2	-
w/ adaptation						
+ STCA	13.2	76.8	-	5.3	39.5	pre.
+ NeDS (ours)	31.7	78.1	-	16.1	50.3	pre.

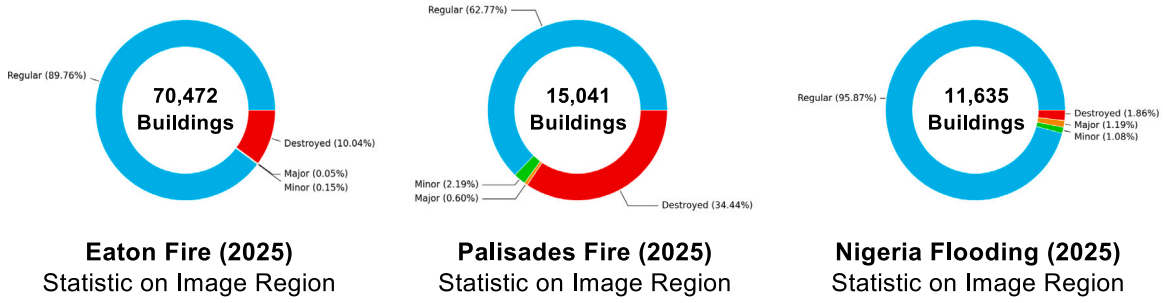

Fig. 8. Statistic of building damage assessment based on our predictions for the image region of Eaton fire, Palisades fire, and Nigeria flooding events.

and corrected ground truth. Given the public availability of pre-event building footprints in Los Angeles, we did not assess building localization accuracy in this case and instead focused solely on damage classification accuracy. All raw pixel-wise predictions were converted to instance-level predictions using pre-event building footprints and object-based image analysis. The evaluation metrics were then computed at the building-instance level. There are 70,472 and 15,041 building instances for the study areas of the Eaton and Palisades fires, respectively.

Results: Table 3 presents a comparative evaluation of building damage assessment performance across the Eaton Fire (2025), Palisades Fire (2025), and Nigeria Flooding (2025) datasets. The baseline model, ChangeOS (source-only), is compared against two adaptation methods: STCA and our proposed NeDS. On the Eaton Fire dataset, the source-only model performs poorly on the minor and major damage classes (14.2% and 19.0%, respectively), despite achieving high accuracy on regular and destroyed classes. Incorporating STCA significantly boosts the overall F_1^{dam} from 27.9% to 65.7%. However, NeDS further improves this to 69.3%, with substantial gains in the minor and major categories (64.7% and 45.5%, respectively), highlighting its effectiveness in addressing class imbalance and domain shift. Similarly, on the Palisades Fire dataset, adaptation yields notable improvements over the source-only model. Although STCA increases F_1^{dam} from 70.6% to 77.1%, NeDS achieves the highest score of 85.0%, along with superior performance in the minor damage (81.4%) and major damage (68.8%) classes. On the Nigeria flooding dataset, the source-only model fails to recognize damage status because of geographic differences and the abnormal surface reflections produced by the severe flood-related weather. Applying NeDS markedly reduces these distribution disparities, delivering substantial improvements over both the source-only baseline and the STCA method. These results demonstrate that NeDS not only generalizes better across unseen target domains but also provides more balanced performance across damage severity levels. Importantly, both adaptation methods operate under a realistic setting that only requires access to pre-disaster target-domain imagery, making them applicable in practical disaster response scenarios.

Based on our building damage assessment products, we provide a statistical summary in Fig. 8. For the Nigeria flooding (see Fig. 11), there are 1.86% buildings destroyed, 1.19% buildings suffered major damage, and 1.08% sustained minor damage. For the Eaton fire, there are 10.04% (7,075) buildings destroyed, 0.15% (109) sustained minor damage, and 0.05% (33) buildings suffered major damage. Fig. 9 demonstrates widespread destruction across residential neighborhoods, with most buildings in the examined zones sustaining severe damage from the wildfire that spread from nearby mountainous terrain to the urban-wildland interface. For the Palisades fire, there are 34.44% (5,180) buildings destroyed, 2.19% (330) sustained minor damage, and 0.60% (90) buildings suffered major damage. Fig. 10 shows the fire appears to have spread from the mountainous/hillside areas down toward the coastal communities. The neighborhoods located at the wildland-urban interface (where development meets natural areas) experienced the highest concentration of damage, particularly in zones Figs. 10(a), (b), and (c) that are highlighted. From this map product, we also find that (i) some coastal areas seem to have been less affected, possibly due to differences in vegetation, topography, or firefighting efforts prioritizing these areas; (ii) the damage is not uniform across neighborhoods, suggesting varying fire behavior or defensive measures. Based on our results, the total number of damaged buildings is 12,817, which is comparable to the officially reported figure of “more than 12,300 structures” (https://www.census.gov/topics/preparedness/events/wildfires/2025-los-angeles.html). This confirms that adaptation through NeDS is effective in real-world cases.

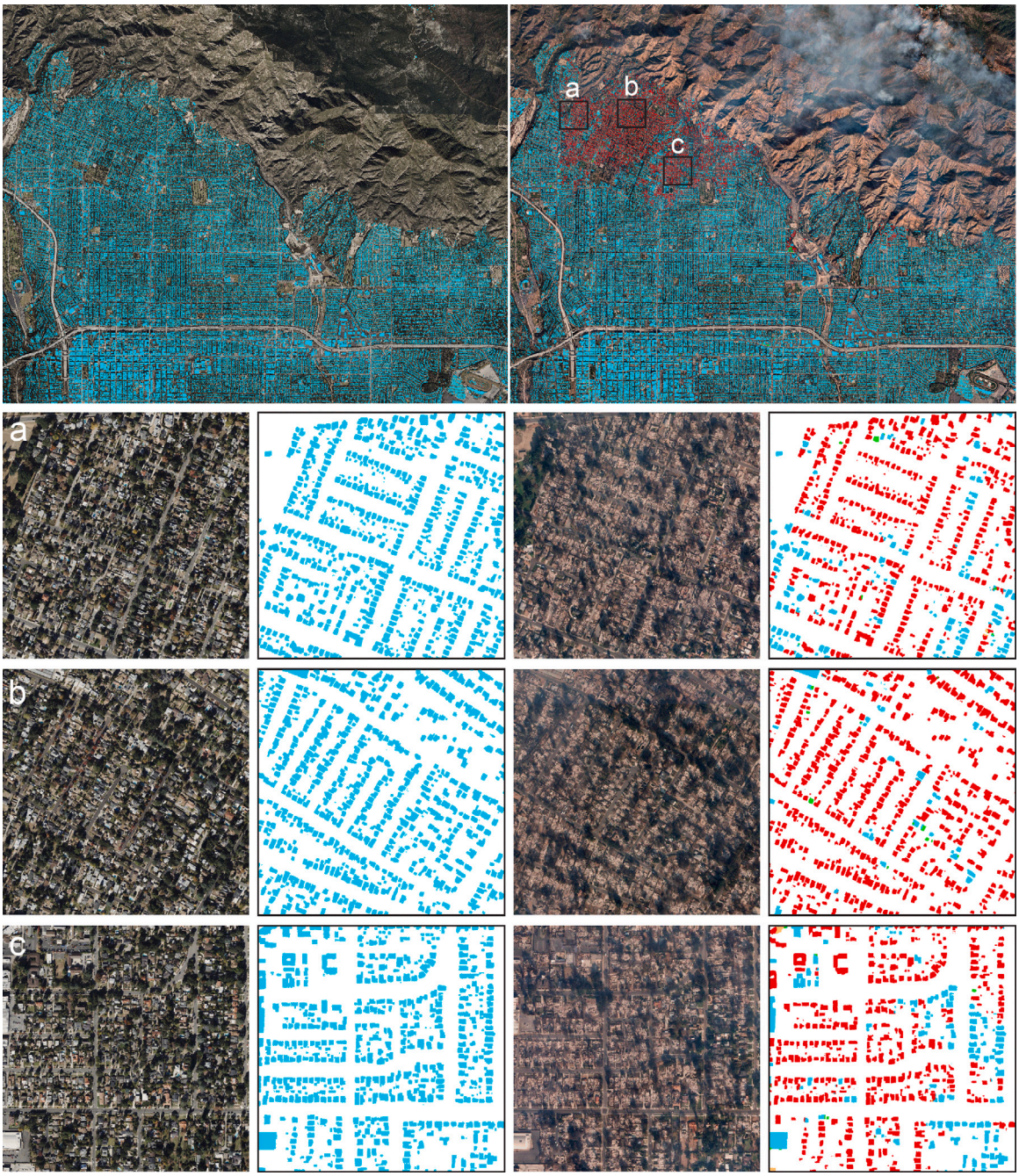

Fig. 9. Building damage assessment of Eaton fire in Los Angeles. Each image has a size of 40,832 × 32,276 pixels. The top row displays pre-disaster (left) and post-disaster (right) satellite images of the entire area, with three zones of interest (a, b, c) marked for detailed assessment. These are raw predictions directly from our building damage assessment model.

4.5. Potential limitations

High-resolution disaster simulation remains a challenging task. Our work, NeDS, approaches this problem in a purely data-driven manner. However, there are several limitations, particularly regarding the integration of geospatial priors and the modeling of disaster-specific physical processes. Topographic features such as elevation, slope, and hydrological flow play an important role in shaping disaster impacts, especially in cases like flooding, landslides, and debris flows. While integrating such information could enhance model performance and generalizability, current global DEM products (e.g., SRTM or ASTER GDEM at ~30 m resolution) are poorly aligned with the sub-meter resolution of input imagery (e.g., WorldView at 0.3 m). This resolution gap has so far limited the practical utility of terrain data. Future work may explore higher-resolution sources or multi-scale fusion strategies to overcome this mismatch. In addition, NeDS does not explicitly model the physical dynamics of specific disasters, such as flood propagation or seismic ground motion. This omission may limit the physical plausibility of the generated post-disaster imagery. Incorporating domain-specific physical processes into the generative framework could improve both the realism and interpretability of simulation results. Addressing these aspects would significantly enhance the quality of synthetic data and improve its alignment with the distribution and patterns observed in real-world disaster events.

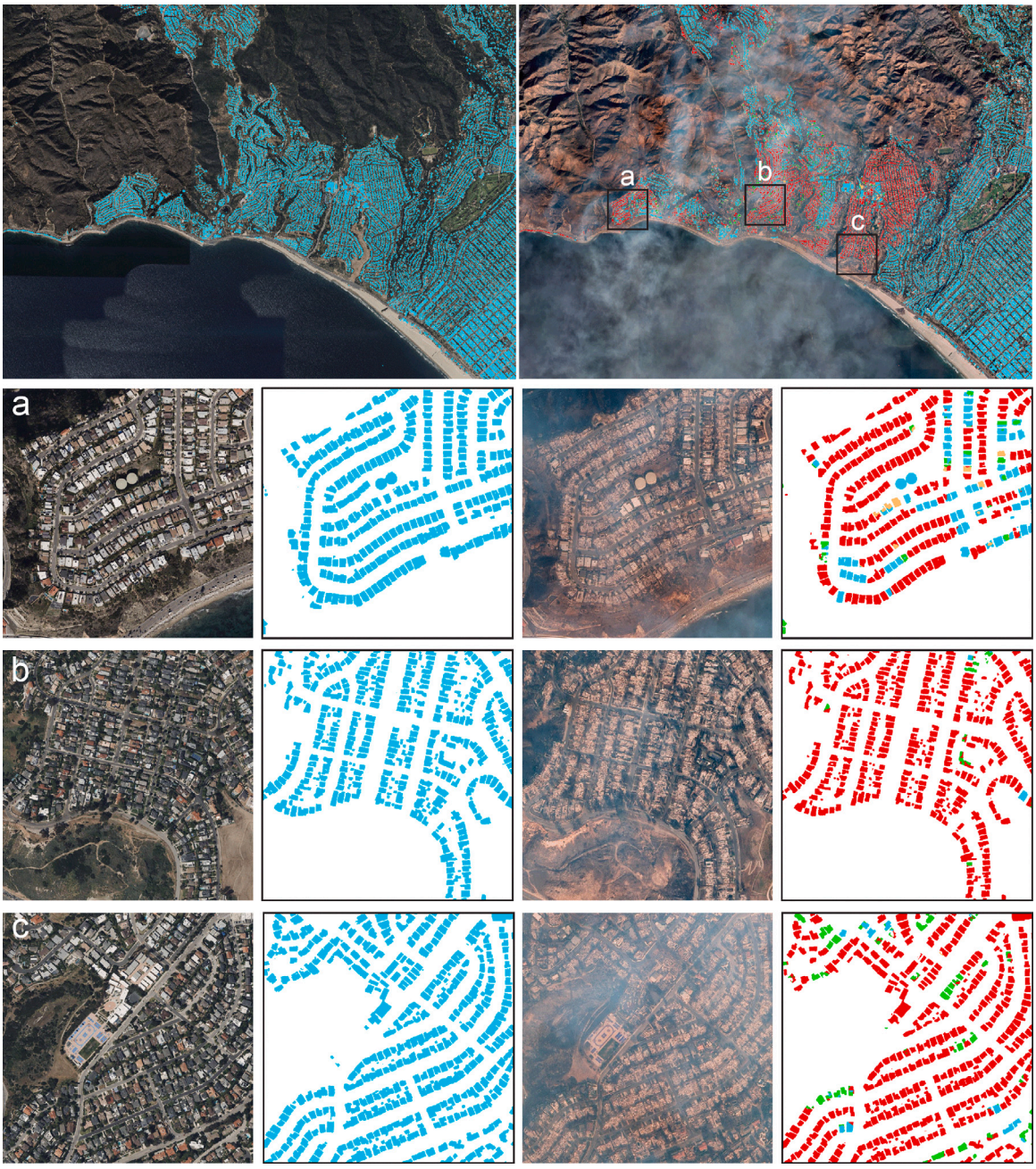

Fig. 10. Building damage assessment of Palisades fire in Los Angeles. Each image has a size of 30,904 × 22,554 pixels. The top row displays pre-disaster (left) and post-disaster (right) satellite images of the entire area, with three zones of interest (a, b, c) marked for detailed assessment. These are raw predictions directly from our building damage assessment model.

5. Conclusion

This work advances transferable building damage assessment by introducing visually interpretable pseudo bi-temporal damage samples. To achieve this, we propose Neural Disaster Simulation (NeDS), a deep disaster generative model capable of synthesizing post-disaster images conditioned on pre-disaster imagery and controllable disaster information such as disaster type and intensity. NeDS is first trained on source-domain damage samples and subsequently used to generate target-domain damage samples from pre-disaster images alone. These synthetic samples are then mixed with real source-domain data to fine-tune building damage assessment models for effective domain adaptation. Thanks to its reliance solely on pre-disaster imagery, NeDS enables adaptation at any time, circumventing the limitations imposed by the unavailability of post-disaster training data. Furthermore, by generating pseudo bitemporal samples at the image level, NeDS inherently offers visual interpretability, making it easier for human experts to inspect and validate. Extensive experiments, including two recent real-world wildfire cases, demonstrate the effectiveness and superiority of NeDS in enhancing the transferability of building damage assessment models. Future work will explore the integration of additional disaster-related information using text-to-image paradigms. In particular, incorporating textual disaster reports, social media signals into the NeDS framework could provide richer event context and further enhance the realism, fidelity, and interpretability of simulated post-disaster imagery.

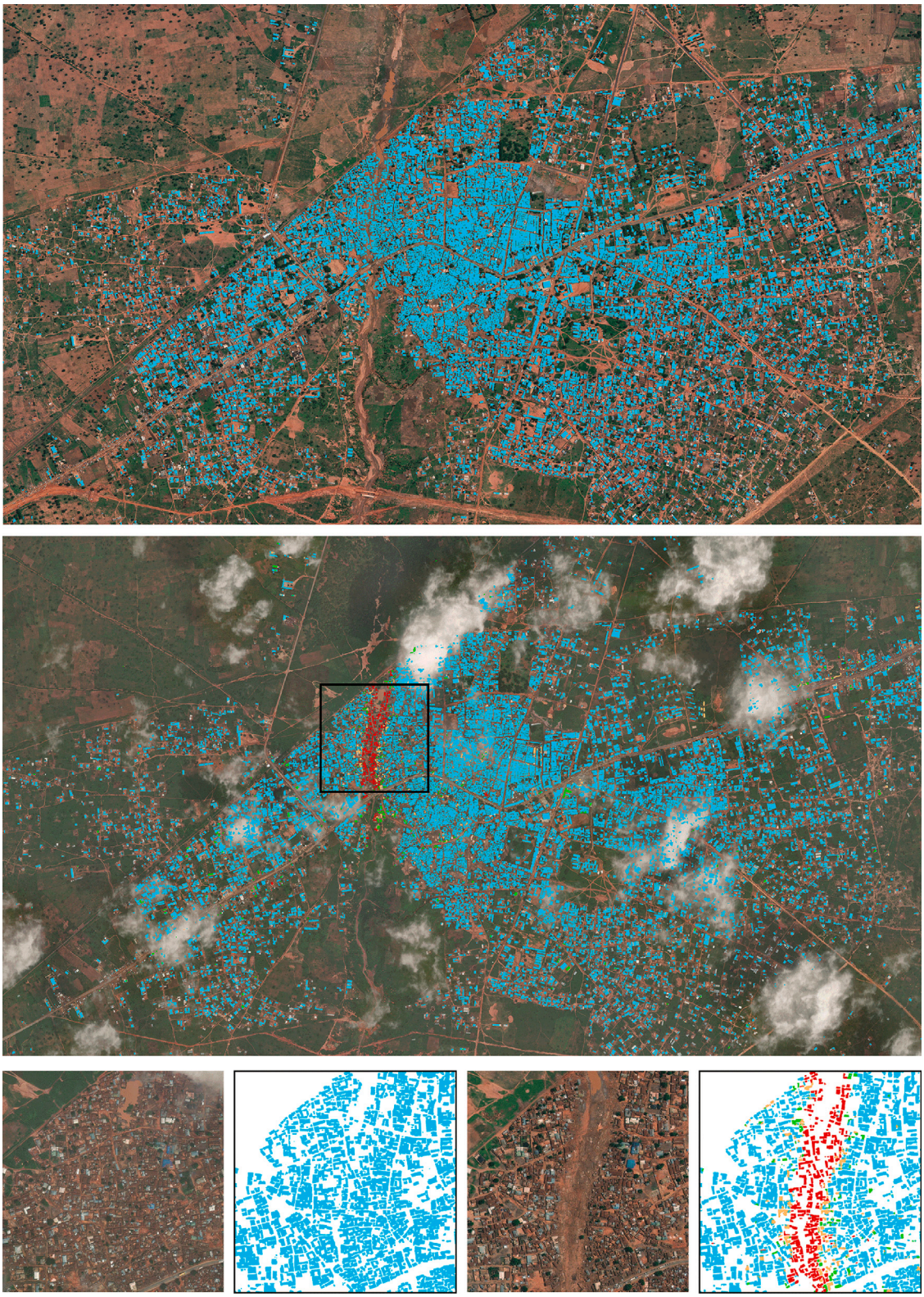

Fig. 11. Building damage assessment of flooding in Nigeria. Each image has a size of 10,125 × 17,827 pixels. The top row displays pre-disaster (left) and post-disaster (right) satellite images of the entire area, with one zone of interest marked for detailed assessment. These are raw predictions directly from our building damage assessment model.

CRediT authorship contribution statement

Zhuo Zheng: Writing – original draft, Methodology, Data curation, Conceptualization. **Yanfei Zhong:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Zijing Wan:** Writing – review & editing, Investigation, Data curation. **Liangpei Zhang:** Writing – review & editing, Supervision, Resources. **Stefano Ermon:** Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the Editor, Associate Editor, anonymous reviewers for their helpful comments and suggestions that improved this article. This work was supported in part by ARO, United States (W911NF-21-1-0125), ONR, United States (N00014-23-1-2159), the CZ Biohub, and the National Natural Science Foundation of China under Grant No. 42325105.

Data availability

Data will be made available on request.

References

Ayush, K., Uzkent, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S., 2021. Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10190.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Mach. Learn. 79 (1), 151–175.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., 2006. Analysis of representations for domain adaptation. Adv. Neural Inf. Process. Syst. 19.

Benson, V., Ecker, A., 2020. Assessing out-of-domain generalization for robust building damage detection. arXiv preprint arXiv:2011.10328.

Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A., 2018. Demystifying MMD GANs. In: International Conference on Learning Representations.

Bouchard, I., Rancourt, M.È., Aloise, D., Kalaitzis, F., 2022. On transfer learning for building damage assessment from satellite imagery in emergency contexts. Remote. Sens. 14 (11), 2532.

Candela, J.Q., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., 2009. Dataset Shift in Machine Learning, vol. 1, MIT Press, p. 5.

Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L., 2022. Dual-tasks siamese transformer framework for building damage assessment. In: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1600–1603.

Chen, H., Song, J., Dietrich, O., Broni-Bediako, C., Xuan, W., Wang, J., Shao, X., Wei, Y., Xia, J., Lan, C., et al., 2025. BRIGHT: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. arXiv preprint arXiv:2501.06019.

Chen, H., Song, J., Han, C., Xia, J., Yokoya, N., 2024. Changemamba: Remote sensing change detection with spatio-temporal state space model. IEEE Trans. Geosci. Remote Sens..

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S., 2022. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In: Advances in Neural Information Processing Systems, vol. 35, pp. 197–211.

Dummit, D.S., Foote, R.M., et al., 2004. Abstract Algebra, vol. 3, Wiley Hoboken.

Durnov, V., 2020. xView2 first place solution. https://github.com/DIUx-xView/xView2_first_place.

Grünthal, G., 1998. European Macroseismic Scale 1998. Technical Report, European Seismological Commission (ESC).

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019a. Creating xBD: A dataset for assessing building damage from satellite imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 10–17.

Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E., Choset, H., Gaston, M., 2019b. xBD: A dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851.

Hoyer, L., Dai, D., Van Gool, L., 2022. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9924–9935.

Kelman, I., 2003. Physical Flood Vulnerability of Residential Properties in Coastal, Eastern England (Ph.D. thesis). University of Cambridge.

Kingma, D.P., Welling, M., et al., 2013. Auto-encoding variational bayes.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R., 2023. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026.

Lin, Q., Ci, T., Wang, L., Mondal, S.K., Yin, H., Wang, Y., 2022. Transfer learning for improving seismic building damage assessment. Remote. Sens. 14 (1), 201.

Liu, C., Ge, L., Sepasgozar, S.M., 2021. Post-disaster classification of building damage using transfer learning. In: 2021 IEEE International Geoscience and Remote Sensing Symposium. IGARSS, IEEE, pp. 2194–2197.

Long, M., Cao, Y., Wang, J., Jordan, M., 2015. Learning transferable features with deep adaptation networks. In: International Conference on Machine Learning. PMLR, pp. 97–105.

Lu, C., Song, Y., 2025. Simplifying, stabilizing and scaling continuous-time consistency models. In: The Thirteenth International Conference on Learning Representations.

Mall, U., Hariharan, B., Bala, K., 2023. Change-aware sampling and contrastive learning for satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5270.

Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P., 2021. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9414–9423.

Noman, M., Naseer, M., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.S., 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27811–27819.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning robust visual features without supervision.

Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T., 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4088–4099.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 379–387.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems, vol. 29.

Salimans, T., Ho, J., 2022. Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations.

Scheibenreif, L., Mommert, M., Borth, D., 2024. Parameter efficient self-supervised geospatial domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27841–27851.

Shen, Y., Zhu, S., Yang, T., Chen, C., Pan, D., Chen, J., Xiao, L., Du, Q., 2021. Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pmlr, pp. 2256–2265.

Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. Adv. Neural Inf. Process. Syst. 32.

Sun, C., Ming, D., Xu, L., Xie, S., Liu, R., Ling, X., 2025. A hybrid damaged building sample generation method based on cross-scale fusion generative model for destroyed building detection after earthquake. IEEE Trans. Geosci. Remote Sens..

Tang, M., Cozma, A., Georgiou, K., Qi, H., 2024. Cross-scale MAE: A tale of multiscale exploitation in remote sensing. In: Advances in Neural Information Processing Systems, vol. 36.

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7472–7481.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T., 2014. Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.

Vickery, P.J., Skerlj, P.F., Lin, J., Twisdale, Jr., L.A., Young, M.A., Lavelle, F.M., 2006. HAZUS-MH hurricane model methodology. II: Damage and loss estimation. Nat. Hazards Rev. 7 (2), 94–103.

Wang, X., Jin, Y., Long, M., Wang, J., Jordan, M.I., 2019. Transferable normalization: Towards improving transferability of deep neural networks. In: Advances in Neural Information Processing Systems, vol. 32.

Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T., 2020. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: European Conference on Computer Vision. Springer, pp. 642–659.

Yang, Y., Soatto, S., 2020. FDA: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4085–4095.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847.

Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J., 2023. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In: Advances in Neural Information Processing Systems, vol. 36, pp. 49842–49869.

Zheng, Z., Ermon, S., Kim, D., Zhang, L., Zhong, Y., 2025. Changen2: Multi-temporal remote sensing generative change foundation model. IEEE Trans. Pattern Anal. Mach. Intell. 47 (2), 725–741. http://dx.doi.org/10.1109/TPAMI.2024.3475824.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. Remote Sens. Environ. 265, 112636.

Zheng, Z., Zhong, Y., Zhang, L., Burke, M., Lobell, D.B., Ermon, S., 2024a. Towards transferable building damage assessment via unsupervised single-temporal change adaptation. Remote Sens. Environ. 315, 114416.

Zheng, Z., Zhong, Y., Zhang, L., Ermon, S., 2024b. Segment any change. In: The Thirty-Eighth Annual Conference on Neural Information Processing Systems.

Zheng, Z., Zhong, Y., Zhao, J., Ma, A., Zhang, L., 2024c. Unifying remote sensing change detection via deep probabilistic change models: From principles, models to applications. ISPRS J. Photogramm. Remote Sens. 215, 239–255.