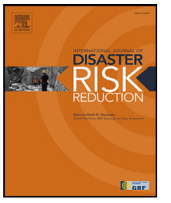
International Journal of Disaster Risk Reduction 128 (2025) 105664

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

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



Benchmarking attention mechanisms and consistency regularization semi-supervised learning for post-flood building damage assessment

Jiaxi Yu [a], Tomohiro Fukuda [a,*], Nobuyoshi Yabuki [b]

[a] Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, The University of Osaka, 1-1 Yamadaoka, Suita, 5650871, Osaka, Japan

[b] Advanced Research Laboratories, Tokyo City University, 1-28-1 Tamazutsumi, Setagaya, 1588557, Tokyo, Japan

ARTICLE INFO

Keywords:
Attention mechanisms
Benchmark performance
Consistency regularization
Flood disaster
Prior knowledge
Semi-supervised learning

ABSTRACT

Rapid and accurate building damage assessment (BDA) following floods is critical for effective disaster response, yet faces challenges from limited labeled data and subtle damage cues in satellite imagery. Existing deep learning change detection (CD) methods may exhibit low recall or misclassify damage inappropriately when transferred directly to the post-flood BDA (Flood-BDA) task. This study addresses these gaps by establishing the first systematic benchmark evaluating both supervised CD model transfer and semi-supervised learning (SSL) specifically for Flood-BDA. This research investigate image-level consistency regularization SSL to combat data scarcity, finding that strategies using pseudo-label derived reference distributions significantly enhance performance (+1.17% avg. Kappa at 5% labels). Notably, pseudo-label outperform ground-truth label strategies in low-label settings (e.g., +4.84% Kappa at 5% labels). Furthermore, confronting the limitations of transferred CD models (low recall, misclassifying 'destroyed' as 'no damage'), this paper proposed a simple prior attention disaster assessment Net (SPADANet), a lightweight U-Net incorporating a simple prior attention module designed for Flood-BDA. SPADANet demonstrably improves recall (+9.22% over best CD baseline) and exhibits more favorable error patterns for Flood-BDA, despite a precision trade-off. This work provides crucial benchmarks, validates the need for recall-driven, DA-specific designs distinct from CD, and demonstrates the potential of prior attention and image-level consistency regularization for post-flood building damage assessment. The code will be available at https://github.com/JX-OctoNeko/Flood_BDA_benchmark.git

1. Introduction

Flood events represent a category of natural disasters that have a widespread impact and seriously threaten civil systems. Due to the continuous expansion of urban areas and the exacerbation of extreme climates, floods are expected to remain a significant factor affecting socio-economic development and the safety of people's lives in the foreseeable future [1]. Implementing pre-event and post-event countermeasures is crucial to minimize the losses and impacts caused by such disasters [2]. Remote sensing information is of great importance in the context of comprehensive disaster management systems, such as facilitating crisis management in disasters [3], vulnerability analysis [4], and disaster assessment (DA) [5].

* Corresponding author.
E-mail addresses: yu@it.see.eng.osaka-u.ac.jp (J. Yu), fukuda.tomohiro.see.eng@osaka-u.ac.jp (T. Fukuda), yabukin@tcu.ac.jp (N. Yabuki).

https://doi.org/10.1016/j.ijdrr.2025.105664
Received 15 February 2025; Received in revised form 13 June 2025; Accepted 22 June 2025
Available online 23 July 2025

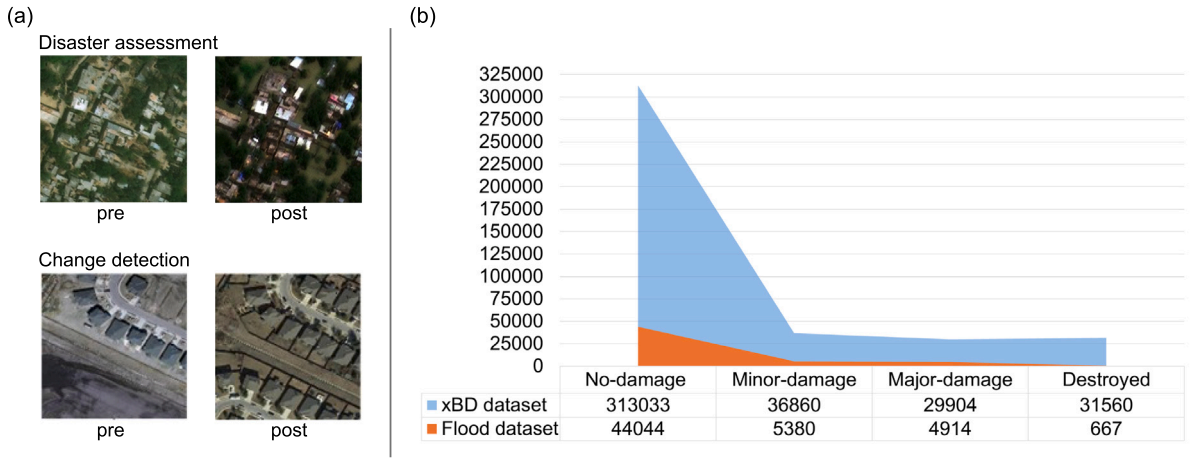


Fig. 1. Characteristics of the flood dataset.(a) Comparison with traditional CD images (b) The sample number of the flood dataset within the overall xBD dataset.

Within this context, the timely assessment of building damage in flood-affected areas is of utmost importance [6,7]. Traditional methods primarily rely on remote sensing images to analyze intensity [8], coherence [9], and polarimetry features [10]. Since these technical approaches have evolved from change detection (CD), building DA is often considered a pre- and post-disaster multi-classification CD issue [11]. However, in the contemporary era, marked by the growing diversity of remote sensing data sources, these handcrafted feature-based methods have faced considerable challenges. As the intricacies of the impacts become more pronounced, the effectiveness of manually extracted features has waned, leading to a decline in detection accuracy [12]. Concurrently, deep learning (DL) has made significant strides in the field of computer vision (CV), prompting the rapid and successful application of DL tools to key issues in remote sensing, such as land use and land cover (LULC) and CD [13].

In traditional research, manually engineered features for CD and DA tasks are identical [2], enabling direct methodological transfer. However, in deep learning, model-extracted features exhibit higher abstraction and complexity than manual features [13]. Given their divergent application scenarios and requirements, for example, empirical observations reveal that in flood building damage assessment (Flood-BDA) scenarios, damaged buildings typically retain overall contours with only subtle edge/texture alterations [14,15], whereas CD predominantly involve complete appearance/disappearance of large structures (Fig. 1 a) [16–18]. Furthermore, task priorities differ fundamentally: CD emphasizes detection precision (minimizing false positives), while DA requires recall rates higher to reduce false negatives – a critical requirement since missing damaged buildings may endanger lives due to misallocated rescue resources [2]. Thus, this paper argues that CD and DA demand differentiated model design paradigms in deep learning frameworks.

However, before developing novel, DA-specific architectures and strategies from scratch, it is crucial to first understand the benchmark performance and limitations of existing, successful DLCD techniques when applied to the Flood-BDA context. Currently, a systematic benchmark exploring how different modules and architectures used in DLCD perform on Flood-BDA tasks is lacking. Establishing such a benchmark is essential to identify specific weaknesses in existing approaches when faced with Flood-BDA data and to provide quantitative grounding for developing targeted improvements.

Therefore, this study first undertakes this critical benchmarking step, focusing specifically on the post-flood domain given its significant societal impact [1] and distinct characteristics [7]. Upon shifting to Flood-BDA tasks, there is a more severe long-tail distribution and data scarcity issues. The scarcity of data can be attributed to the inherent challenges associated with collecting data related to disaster events, which occur with less regularity than conventional CV. If the research scope is narrowed to flood events, the available data further diminishes, leading to a natural predicament of data insufficiency for Flood-BDA tasks. The absence of data is highly detrimental to the model's ability to extract information from remote sensing data and can easily lead to overfitting [13]. Furthermore, due to the low-intensity nature of flood damage to buildings [15], the proportion of buildings falling into the 'destroyed' class is significantly reduced (the original xBD dataset had an even distribution across the three subcategories of damaged, refer to Fig. 1 b), which can easily result in minority class features being overwhelmed by majority class features [19].

Recognizing the specific data landscape of Flood-BDA is crucial for selecting appropriate strategies. While vast amounts of post-disaster satellite imagery are increasingly available due to frequent satellite passes, obtaining high-quality damage labels is the primary bottleneck. Annotating building damage levels accurately requires expert knowledge, adherence to strict damage scales (like the Joint Damage Scale used in xBD [20]), and significant time investment, making labeled datasets inherently small and expensive [17,21]. Conversely, collecting unlabeled imagery is relatively inexpensive. This disparity points strongly towards Semi-supervised learning (SSL) as a highly relevant approach to leverage the abundant unlabeled data alongside limited labeled samples.

Among various SSL paradigms [22], the research specifically investigates image-level consistency regularization [23]. This choice is motivated by the potential to utilize groups of unlabeled images. The paper hypothesizes that within a specific flood event or geographically similar affected areas, the distribution of building damage levels, while unknown a priori, follows certain underlying

patterns influenced by factors like flood intensity and building typology. Image-level consistency regularization allows us to enforce alignment between the model's predicted distribution on unlabeled batches and a reference distribution Q. The core principle [22] suggests that if this reference distribution accurately approximates the true (but unknown) label distribution of the unlabeled data group, the consistency constraint acts as a powerful learning signal, guiding the model towards separating classes effectively in low-density feature space. The experiment explores different ways to construct this reference distribution, aiming to obtain the most suitable strategy for capturing these potential damage patterns from batches of unlabeled images.

Effective SSL, however, still relies on a robust supervised learning component capable of generating meaningful initial predictions and supervised signals. Our benchmark results indicated that attention mechanisms show promise in enhancing feature representation for CD/DA tasks [6,17,24]. While standard self-attention [24] can be computationally intensive and may not be optimal for faint signals, this paper hypothesizes that a simpler prior attention mechanism [25], specifically designed to amplify outlier-like features (corresponding to subtle damage cues) without adding parameters, could be more suitable for Flood-BDA.

Consequently, this paper proposes a simple prior-attention disaster assessment network (SPADANet). This network integrates the parameter-free prior attention module within a classic U-Net structure. SPADANet is designed to be a lightweight yet effective benchmark for Flood-BDA, capable of retaining critical building damaged related information, serving both as a standalone improved model for supervised learning and as the necessary supervised signal generator within our SSL framework.

In summary, this paper makes the following primary contributions:

(1) This study establishes the first deep learning benchmark for Flood-BDA. Through systematic supervised learning controlled experiments, this paper reveals the differences in effectiveness of different modules in Flood-BDA scenarios. By evaluating 36 benchmark configurations combining 8 classic change detection models and the SPADANet with 4 SSL paradigms, this paper quantitatively demonstrates that consistency regularization achieves numerical improvement under 5%–50% annotation ratios as well as SPADANet's adaptability to the Flood-BDA task, clarifying the performance boundaries of attention mechanisms and SSL.

(2) Through designed SSL reference distribution experiments, this study first discovers that pseudo-label distributions (Q_{pseudo}) better approximate unlabeled data truth than ground-truth distributions of limited labeled dataset (Q_{gt}) when annotation ratios $\leq$ 50%. Empirical results show Q_{gt} regularization causes annotation overfitting, while dynamically adjusted Q_{pseudo} implicitly models class transition matrices, attaining 73.53% fully-supervised equivalent performance with only 10% annotations, filling SSL's theoretical validation gap in Flood-BDA.

(3) To address coupled challenges of subtle feature variations and annotation scarcity in Flood-BDA, this research proposes SPADANet—a parameter-free prior attention enhanced UNet architecture. SPADANet counters the trend of advanced CD methods like BIT sacrificing recall for precision; instead, through prior-attention guided outlier neuron features, it achieves 79.10% recall—a 10.83% improvement over its UNet backbone with identical parameters (1.35M), and 1.13% improvement in kappa value than the next-best model. This high-recall performance, which aligns with the critical disaster response principle of "preferring false positives over missed detections". Thus, SPADANet serves as a crucial validation that underexplored mechanisms like prior attention, when selected based on task-specific needs, offer a highly promising direction for developing more effective Flood-BDA models.

2. Related work

2.1. Disaster assessment and change detection

Change detection (CD) in remote sensing involves identifying differences in land use or land cover by comparing multi-temporal images [11]. Post-disaster damage assessment (DA) can be viewed as an extension of CD, transitioning from binary change identification to multi-class semantic segmentation quantifying the extent of damage, particularly to structures like buildings [5]. Historically, especially with traditional feature engineering, methods applied to DA often mirrored those used in CD [2]. This methodological inheritance has largely continued into the deep learning (DL) era.

However, as discussed in the Introduction, the objectives of CD and DA diverge significantly. Specifically, Flood-BDA necessitates a strong emphasis on high recall [2]. While minimizing false positives (precision) is desirable, it is often secondary to ensuring comprehensive identification of all potentially damaged structures in the immediate aftermath of an event. This fundamental difference in task requirements suggests that evaluation metrics and model design philosophies optimal for CD may not directly translate to optimal solutions for DA.

Furthermore, the DL models typically employed for DA have largely evolved alongside advancements in the broader CD research [24,26,27], which benefits from a larger volume of research and more diverse datasets [16–18,20] compared to DA domain [5,6,28]. Consequently, DA research often adopts architectures and modules proven effective in CD, such as fully convolutional networks (FCN) [5,29], Siamese architectures [7,30], U-Net structures with skip connections [31–33], and various attention mechanisms [6,17,21,24,32,34]. In addition, some CD methods that remain unexplored in DA research, for example, Papadomanolaki et al. [31] dedicated to temporal modeling employs techniques like recurrent neural networks (RNNs) to capture long-sequence image features. Lin et al. [26] generated pseudo-videos to incorporate temporal information. Fang et al. [27] explored interaction strategies such as aggregation-distribution and feature exchange using a general MetaChanger architecture. While leveraging these mature CD techniques provides a valuable starting point, their direct applicability and baseline performance within the specific constraints and objectives of Flood-BDA (subtle changes, recall focus, data imbalance) require systematic investigation.

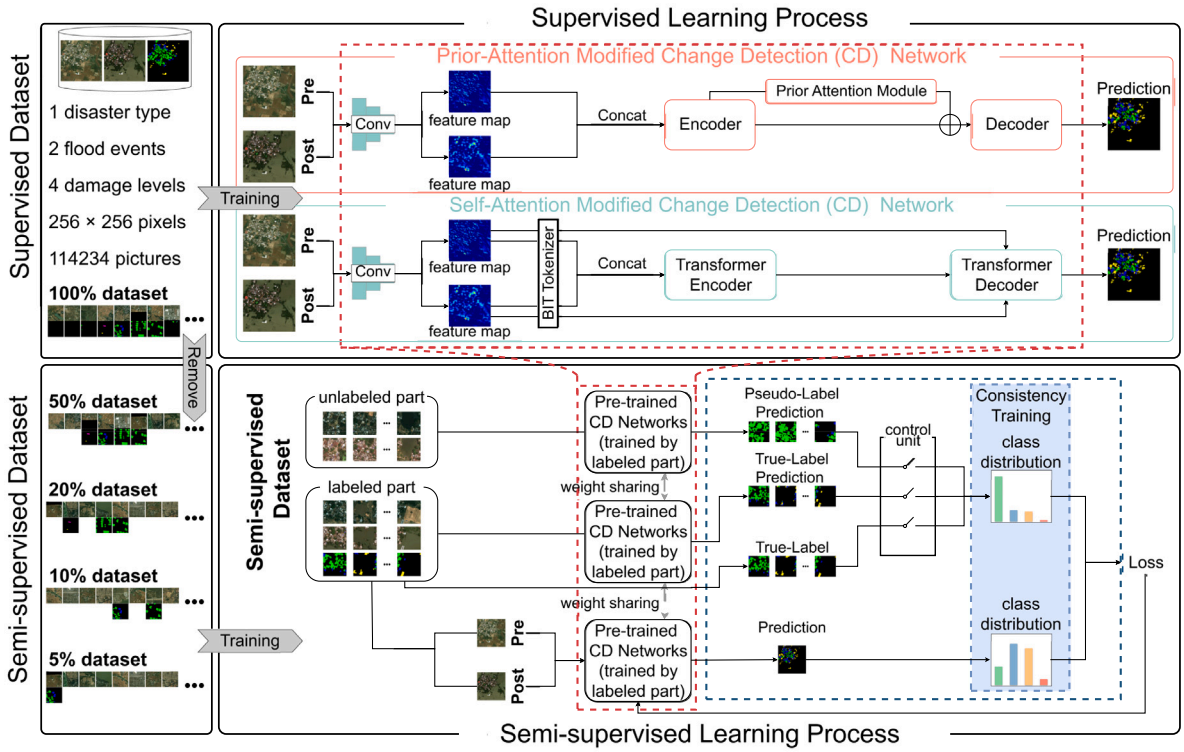


Fig. 2. Overview of the deep learning approach for post-flood building damage assessment. "Pre" and "Post" denote the images before and after the disaster. "Conv" denotes the convolutional layers, "Concat" denotes concatenation.

Therefore, a primary goal of this work is to establish a benchmark performance for representative DLCD techniques when applied to the Flood-BDA task. This research aims to evaluate how different core architectural components developed primarily for CD perform under Flood-BDA conditions. It helps bridge the methodological gap between the two fields by providing benchmark results specific to DA needs.

2.2. Semi-supervised learning

Flood-BDA faces significant challenges related to labeled data scarcity and severe class imbalance (Fig. 1). While various techniques exist to mitigate these issues, such as over/under-sampling [19], transfer learning and domain adaptation [35], or adjusting loss functions to prioritize minority classes [36], these approaches often do not fully leverage the potential information within the large volumes of readily available unlabeled post-disaster imagery. Therefore, this paper argues that semi-supervised learning (SSL), which utilizes limited labeled data alongside large amounts of unlabeled data, represents a potentially powerful approach for DA scenarios.

Despite its potential, the application of SSL specifically for DA tasks remains limited. To understand potential avenues and bridge this methodological gap, this paper investigated existing SSL applications in the related field of CD. While examples utilizing GAN-based approaches [37] and proxy-label/self-training methods [38] exist, this research identified a notable gap: methods based on consistency regularization (CR) [23], a prominent and effective SSL paradigm in general computer vision [22,39], appear largely unexplored in the published CD/DA literature.

Motivated by the desire to rapidly evaluate the potential of this underexplored CR paradigm for Flood-BDA and thereby fill the identified gap, this paper investigated established CR techniques. For instance, Berthelot et al. [40] introduced distribution alignment and augmentation anchoring, focusing on aligning marginal distributions with ground-truth label statistics. Sohn et al. [41] established FixMatch using consistency between weakly and strongly augmented versions of the same input image, and Yang et al. [23] extended this with UniMatch. These influential studies primarily demonstrate CR through intra-image perturbations (i.e., data augmentation) [37]. However, they do not explicitly consider leveraging statistical information aggregated from groups of images sharing similar event characteristics, which this paper believed is a relevant approach for Flood-BDA scenarios where batches of images often come from the same disaster context.

Therefore, this research developed an approach that applies CR by enforcing consistency between the model's predicted class distribution and a reference distribution (Q) derived from the statistics of image groups (either ground truth label or pseudo-label, see Section 4.2).

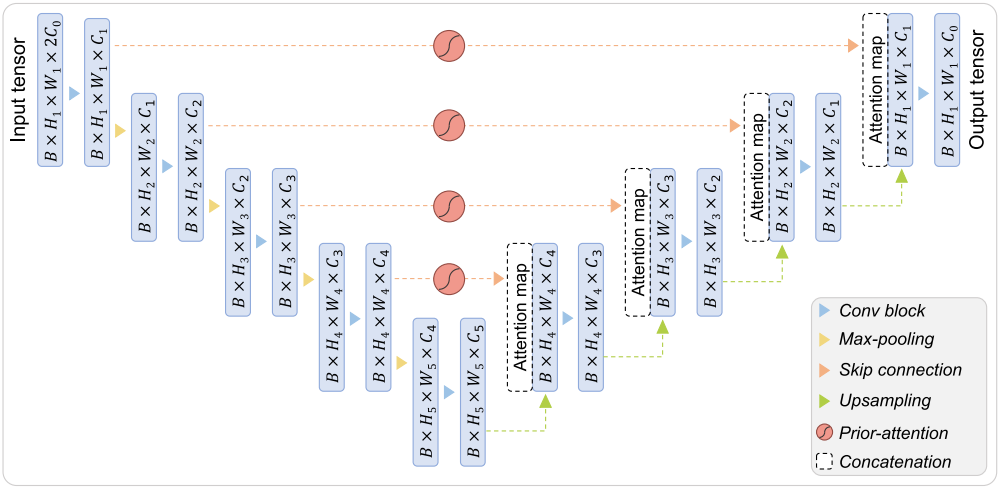


Fig. 3. Illustration of the SPADANet. B denotes batch size, H denotes height, W denotes width, C denotes channel, $Conv$ denotes convolutional layers.

2.3. Attention mechanism

In the benchmark experiments for Flood-BDA, models incorporating attention mechanisms generally exhibited strong performance relative to other architectures 4 [24,26,31]. This finding motivated a deeper investigation into how attention is typically employed in the CD and DA literature.

Existing CD and DA research highlight several ways attention mechanisms are utilized. For example, Xing et al. [34] added a self-attention module to U-Net to assess the flood vulnerability of buildings. Fang et al. [32] applied the UNet++ model to change detection and employed a channel attention module to focus the model on channel information and avoid redundancy. Chen and Shi [17] designed a pyramid spatial–temporal attention module (PAM) to utilize acquired spatial information. Zhang et al. [21] used spatial and channel attention as part of the deep supervision information added to the network. These designs aim to expand the model's RF to obtain more information directly but do not consider that extending the RF also increases the redundancy of the model's input. More advanced research builds on this by transforming image information into important semantic information for learning, such as the BIT [24] established by Chen et al. which treats information as tokens for positioning and uses a convolutional neural network to embed input images before employing a Transformer module for change detection.

In summary, this paper discovers that these CD methods utilize the understanding of CD tasks to inject priors knowledge into the model and improve the model by capturing correlations. This led us to explore prior attention mechanisms, an approach where attention patterns are guided not just by the input data itself, but also by incorporating pre-defined structural assumptions or priors about what information is likely to be important. This class of attention mechanism appears less explored in the specific context of CD/DA compared to self-attention variants.

Various forms of prior attention exist, often tailored to specific tasks by incorporating assumptions about global context or relationships. For instance, Hou et al. [42] used an Interaction-Aggregation-Update (IAU) module incorporating global spatial, temporal, and channel context information for pedestrian recognition tasks. Zhang et al. [43] proposed an effective Relation-Aware Global Attention (RGA) module for capturing global spatial and channel structure information for attention learning. Yang et al. [25] established an energy function based on the activity of neurons, allowing the model to process all dimensions of image information according to the spatial inhibitory characteristics of neurons. This study believes that the prior attention module established by Yang et al. can amplify the distance between outlier neurons and other neurons, emphasizing small change pixels. Therefore, this study adopted a parameter-free prior attention based on spatial inhibition [25].

3. Preliminaries

3.1. A representative self-attention CD network: BIT

As discussed in Section 2.3, this benchmark experiments highlighted the general effectiveness of attention mechanisms for Flood-BDA tasks. Among the methods evaluated, the BIT network, proposed by Chen et al. [24], demonstrated particularly strong performance. Understanding its architecture is valuable for two reasons: (1) It serves as a high-performing baseline representative of advanced CD methods found in the literature. (2) Its reliance on self-attention provides a clear point of contrast to the prior attention approach employed in our proposed SPADANet (Section 3.2). The key steps are:

Feature Extraction: Like many CD methods, BIT first uses a standard CNN backbone to extract deep semantic features $(F_{pre}, F_{post}) \in \mathbb{R}^{H \times W \times C}$ from the pre- and post-disaster images $(x_{pre}, x_{post}) \in \mathbb{R}^{H \times W \times C_0}$, where H, W represent the height, width of the images, C

and C_0 represent the predefined feature map channel dimension and original image channel dimension. In this paper, C_0 is set to 4, representing four types of label classes. C is a predefined value of 32.

Semantic Tokenizer: A convolution step in BIT is to distill the rich spatial information within the feature maps (F_{pre}, F_{post}) into a small, fixed number of "semantic tokens" $(T_{pre}, T_{post}) \in \mathbb{R}^{L' \times C}$, where $L' \ll (H \times W)$.

Transformer Encoder: The concatenated tokens $(T_{sum} = Concat(T_{pre}, T_{post}))$ are fed into a Transformer encoder. This encoder utilizes Multi-Head Self-Attention (MSA) layers. After the transformer encoder, the dense context information $T_{new} \in \mathbb{R}^{2L' \times C}$. The MSA formula is,

$$\begin{aligned} \mathrm{MSA}(T_{\mathrm{sum}}) &= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W_O, \\ \mathrm{head}_j &= \mathrm{Attention}(T_{\mathrm{sum}} W_j^q, T_{\mathrm{sum}} W_j^k, T_{\mathrm{sum}} W_j^v) \end{aligned} \tag{1}$$

where h is the number of attention heads, j represents the jth head. $\mathbf{W}_j^q, \mathbf{W}_j^k, \mathbf{W}_j^v \in \mathbb{R}^{C \times d}$, $W_O \in \mathbb{R}^{h \times d \times C}$ are linear projection matrices that transform the encoded information back into the original tensor size of T_{sum}.

Transformer Decoder: This decoder uses Multi-Head Cross-Attention (MA). Here, the original spatial feature maps (F_{pre}, F_{post}) act as queries, while the context-enriched tokens T_{new} is split into two sets $(T_{new\_pre}, T_{new\_post})$ serve as keys and values. This allows the decoder produce enhanced feature maps $(F_{new\_pre}, F_{new\_post})$. The MA formula is,

$$\begin{aligned} \mathrm{MA}(F_i, T_i) &= \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W_O, \\ \mathrm{head}_j &= \mathrm{Attention}(F_i W_j^q, T_i W_j^k, T_i W_j^v) \end{aligned} \tag{2}$$

where F_i represents F_{new_pre} or F_{new_post}, T_i represents T_{new_pre} or T_{new_post}, $W_j^q, W_j^k, W_j^v \in \mathbb{R}^{C \times d}$, $W_O \in \mathbb{R}^{h \times d \times c}$ are linear projection matrices.

Prediction Head: Finally, the difference between the refined feature maps $(F_{new\_pre}, F_{new\_post})$ is computed, and a simple prediction head (e.g., a few convolutional layers) processes this difference map to generate the final pixel-wise change probability map.

In essence, BIT leverages the power of self-attention to model long-range spatio-temporal dependencies via a compact token representation. Its effectiveness, demonstrated in our benchmarks (Table 4) and the original work [24], establishes it as a strong baseline using attention mechanism. This architecture serves as an important reference point when evaluating the SPADANet, which adopts a fundamentally different, parameter-free prior attention strategy designed specifically for the subtle change characteristics of Flood-BDA (For detailed formulations of the attention mechanisms within BIT, please refer to the original publication [24]).

3.2. Simple prior-attention DA network (SPADANet)

Since the application of UNet in medical imaging [33], the U-shaped architecture has become an effective structure for sampling and preserving image features in many tasks, such as Diffusion models [44], Pix2Pix [45].

Therefore, in designing the prior-attention CD model, this paper considered that the change features in disaster scenarios are subtle and prone to loss during downsampling. Therefore, a UNet capable of retaining shallow information was used as the base model, with the addition of the prior-attention module as shown in Fig. 3. The input bitemporal images $(x_{\mathrm{pre}}, x_{\mathrm{post}}) \in \mathbb{R}^{H_1 \times W_1 \times C_0}$ are concatenated to obtain $X \in \mathbb{R}^{H_1 \times W_1 \times 2C_0}$. Through five convolutional units in UNet, generating downsampled feature maps $X_f^m \in \mathbb{R}^{H_m \times W_m \times C_m} (m = 1, 2, 3, 4, 5)$. In the feature map X_f^m, a target neuron t is selected, and all other neurons are denoted as X_i. The attention maps for the corresponding layer are obtained by inputting both t and x_i into the prior-attention module (5)(6), denoted as $X_{new}^n (n = 1, 2, 3, 4)$.

Prior-attention Module: This paper adopts the prior-attention module proposed by Yang et al. [25], which is based on the spatial inhibitory properties of neurons. This work designs a simplified energy formula to calculate the importance of each neuron in a neural network. With $x_i \in X_i$ computing the mean $\mu = \frac{1}{S} \sum_{i=1}^{S} x_i$ and variance $\sigma^2 = \frac{1}{S} \sum_{i=1}^{S} (x_i - \mu)^2$, where i is the index of the spatial dimension. In this formula, $S = H \times W$ represents the number of neurons in a channel. The energy function is calculated as follows:

$$e_t^* = \frac{m(\sigma^2 + \lambda)}{(t - \mu)^2 + n\sigma^2 + n\lambda} \tag{3}$$

The greater the difference between the target neuron t and its surrounding neurons x_i, the lower the value of e_t^*, indicating the greater importance of t. Therefore, the importance of each neuron can be obtained by calculating $\frac{1}{e_t^*}$ (where m and n serve as hyperparameters to adjust the energy of individual neurons) The ratio $\frac{n}{m}$ of used in this paper is 0.5. Therefore, the importance of each neuron in the feature map is calculated by sequentially computing e_t^*, obtaining $e \in E$ representing the overall attention of all dimensions, and the formula for calculating the attention maps x_{new}^n is as follows:

$$X_{new}^n = \mathrm{sigmoid}\left(\frac{1}{e}\right) \odot X_f^n \tag{4}$$

The generated attention maps X_{new}^n are then concatenated with the convolutional layers in the decoder and ultimately projected back into the pixel space to obtain the change maps.

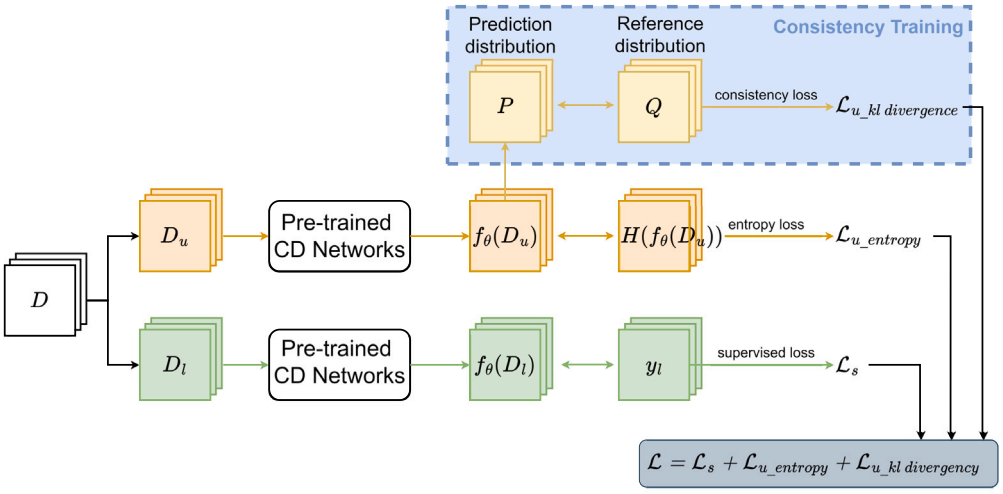


Fig. 4. Overall framework of the proposed semi-supervised learning. A pre-trained CD network f_θ processes labeled D_l and unlabeled D_u data. Supervised loss $\mathcal{L}_s$: Calculated using labeled predictions $f_\theta(D_l)$ and ground-truth y_l (refer to Eq. (5)). Unsupervised Losses: Based on unlabeled predictions $f_\theta(D_u)$: $\mathcal{L}_{u\_entropy}$ (refer to Eq. (6)): Minimizes prediction entropy to encourage confidence. $\mathcal{L}_{u\_kl\ divergence}$ (refer to Eq. (8)): Aligns the batch's predicted class distribution P with a target Reference Distribution Q (refer to Fig. 5). The total loss $\mathcal{L}$ optimizes f_θ.

4. Proposed method

4.1. Overview of the proposed method

Fig. 2 illustrates the experimental workflow for achieving benchmark performance in attention modules and image-level consistency regularization. This workflow comprises a fully supervised process and a semi-supervised process. The fully supervised learning process tests the performance of CD networks with prior and self-attention modules in post-DA tasks, comparing the advantages of the two mechanisms under different tasks. The SSL process evaluates the impact of various image-level consistency regularization perturbations and compares whether different reference distributions affect the amount of information learned from unlabeled data.

The proposed method in this paper adopts a two-stage improvement approach:

(1) The improvements in the red dashed box in Fig. 2 are specific to the model architecture. The purpose of this experiment is to determine which of the temporal relationships [26,31], self-attention [24], and spatial inhibition prior attention [25] is more suitable for DA tasks. Additionally, it provides supervised signals for semi-supervised experiments.
(2) The SSL strategy employs a comprehensive approach involving consistency regularization and entropy minimization for unlabeled data. The improvements are shown in the blue dashed box in Fig. 2. This paper assumes four image-level reference distributions most likely to approximate the ground-truth label classification distribution of unlabeled data as practical perturbations: (1) Pseudo-label predictions generated by a well-trained model on unlabeled data. (2) Ground-truth label predictions generated by a well-trained model on labeled data. (3) Ground-truth label group from the same dataset. (4) A combined distribution formed by combining the above three reference distributions. This assumption is based on the idea that when the preset reference distribution is closer to unlabeled data's ground-truth label classification distribution, the SSL method trained with it will produce the most accurate predictions.

The following sections will describe the above improvements in detail.

4.2. Semi-supervised learning framework

The SSL method framework utilizing consistency constraint and proxy-label method is depicted in Fig. 4. The semi-supervised dataset $D_l = \{(x^i_{l\_pre}, x^i_{l\_post}), y_l\}^M_{i=1}$ and the unlabeled dataset $D_u = \{(x^i_{u\_pre}, x^i_{u\_post})\}^N_{i=1}$. Here, (x_{u_pre}, x_{u_post}) and (x_{l_pre}, x_{l_post}) represent image pairs before and after the disaster, respectively, with M and N denoting the number of image pairs. y_l is a length L vector, where each element is drawn from the set $\{0, 1, 2, 3\}$ representing the four possible classes.

First, the labeled dataset D_l is used to train the CD model, resulting in a pre-trained CD model f_θ. By inputting labeled images x_l into the pre-trained model f_θ, segmentation predictions $\hat{y}_l$ are obtained. Utilizing $\hat{y}_l$ and ground truth y_l, the model is optimized through supervised loss $\mathcal{L}_s$, by using Eq. (5). For the unlabeled dataset D_u, this paper employs a comprehensive approach combining proxy-label and consistency training. The essence of the proxy-label method is to minimize the model's entropy in low-density

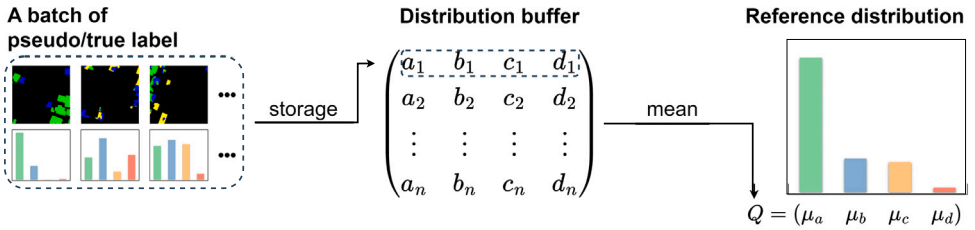


Fig. 5. Process of obtaining reference distribution. (a, b, c, d) denotes the sum of pixel statistics for each of the four categories in each image, and μ denotes the average value of pixels for each category.

regions [46] by pre-training model f_θ to generate predictions $\hat{y}_u$ for unlabeled data x_u and then minimizing the information entropy $\mathcal{L}_{u\_entropy}$ of $\hat{y}_u$ to optimize the model using Eq. (6).

The semi-supervised learning framework in this study leverages consistency regularization to utilize unlabeled data. The core principle behind consistency regularization relies on the low-density separation assumption [22]: the decision boundary between classes should ideally lie in regions where data points are sparse. Enforcing consistency – requiring that the model's predictions are robust to certain perturbations or that the predicted distribution aligns with a reference distribution – encourages the model to place decision boundaries away from high-density data clusters.

In our image-level approach, consistency is enforced by minimizing the KL divergence between the model's predicted class distribution P for a batch of labeled/unlabeled data reference distribution Q. The effectiveness of this loss term hinges on how well Q represents the 'true' target distribution for the unlabeled data. This research explores four different strategies for constructing Q, each based on different assumptions about what constitutes a reliable reference distribution:

(1) Strategy 1 (unlabeled pseudo-label distribution) is derived from pseudo-labels generated by the current model on the unlabeled data itself. If the current model is accurate, its predictions on the unlabeled data (pseudo-labels) reflect the underlying structure and class proportions of that data. Enforcing consistency towards this distribution helps the model refine its boundaries based on the unlabeled data structure itself, potentially improving separation in low-density regions populated by unlabeled points. This assumes the pseudo-labels capture the true unlabeled distribution better than alternatives, especially when the model is reliable.
(2) Strategy 2 (labeled prediction distribution) is derived from the model's predictions on the labeled data. The model's output distribution on labeled data represents a potentially 'smoothed' or 'noisy' version of the ground-truth distribution. Enforcing consistency towards this might help stabilize training or propagate knowledge from labeled to unlabeled data, assuming the model's outputs on labeled data is a beneficial target.
(3) Strategy 3 (labeled ground-truth distribution) is the distribution of the ground-truth labels in the labeled set. The labeled and unlabeled data are drawn from the same underlying distribution (i.e., they belong to similar events/conditions). Therefore, the class proportions observed in the labeled set are assumed to be representative of the unlabeled set. Enforcing consistency towards this fixed distribution encourages the model's predictions on unlabeled data to match the known global statistics of the labeled data.
(4) Strategy 4 (combined distribution) is effectively an average or combination of the distributions used in Strategies 1, 2, and 3. A combination of targets – leveraging the adaptability of pseudo-labels (1), the stability of model behavior on known data (2), and the anchor of ground-truth statistics (3) – might provide the most robust overall learning signal, balancing different potential biases.

By comparing these strategies, this research aim to understand which assumption regarding the reference distribution Q leads to the most effective application of the consistency regularization principle for improving model performance on Flood-BDA tasks using limited labeled data.

As shown in Fig. 2 blue dashed box, a control unit chooses different reference distributions Q based on the adopted strategies. The predictions $\hat{y}_u$ from unlabeled data are transformed into class distributions P, and calculating KL divergence $\mathcal{L}_{u\_kl\ divergence}$ by using Eq. (8).

The process of generating the reference distribution Q is depicted in Fig. 5. In this process, batches of generated labels are used to count the elements in each class and store them in a distribution buffer. A buffer size n is set, and after a certain number of labels are accumulated in the buffer, the average is calculated to obtain the reference distribution Q.

The four different reference distributions generated will serve as different optimization directions for semi-supervised learning, and their effectiveness will be compared through experiments to determine their superiority or inferiority.

4.3. Loss function

Our proposed method involves two training processes, utilizing three loss functions for training models. The loss function in the supervised training process is consistent with the supervised part of the semi-supervised training process, both being $\mathcal{L}_s$. Semi-supervised loss consists of minimizing entropy loss and KL divergence.

4.3.1. Supervised loss

In this experiment, which is a multi-classification task, the weighted cross-entropy loss is used to optimize the prediction probability distribution in supervised learning. This loss function only acts on labeled data and can be expressed as

$$\mathcal{L}_s = -\frac{1}{M}\sum_{i=1}^{M}\sum_{c=1}^{4} w^c y_i^c \log(p_i^c) \tag{5}$$

where $\mathcal{L}_s$ represents the supervised loss, M denotes the number of samples, i is the sample index, c is the category index, ranging from 1 to 4, representing the four damage categories. w^c represents the weight of the category, y_i^c is the ground truth label of sample i for category c, and p_i^c is the model's predicted probability distribution for sample i in category c.

4.3.2. Entropy minimization loss

Generally, models tend to produce low-certainty, high-entropy predictions for unlabeled data. The pseudo-labeling method is based on the assumption that similar data in low-density regions are separable [46]. Therefore, minimizing the information entropy of pseudo-labels during training is necessary to prevent the decision boundary from approaching data points. The information entropy loss term used in this experiment is defined as

$$\mathcal{L}_{u\_entropy} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} p_i^c \log p_i^c \tag{6}$$

where $\mathcal{L}_{u\_entropy}$ represents the loss function for entropy minimization, N denotes the number of samples, c is the category index, and p_i^c is the probability that sample i belongs to category c.

4.3.3. Kullback–Leibler divergence

Since there are no original labels, this research assumes the ground-truth label distribution of unlabeled data as the reference distribution Q. By learning, this method aims to make the predicted probability distribution of categories P approach the reference distribution. This paper uses the standard KL divergence to measure these two distributions' differences. The definition of KL divergence is as follows

$$D_{KL}(P \parallel Q) = \sum_i P(i)\log\frac{P(i)}{Q(i)} \tag{7}$$

In this equation, $P(i)$ and $Q(i)$ represent the probabilities of distribution P and Q at the ith element, respectively. Based on Eq. (7), the loss function is defined as

$$\mathcal{L}_{u\_kl\ divergency} = \frac{1}{N}\sum_{i=1}^{N} D_{KL}(P,Q) \tag{8}$$

where $\mathcal{L}_{u\_kl\ divergence}$ denotes the loss function for KL divergence, N is the number of samples, $D_{KL}(P,Q)$ represents the KL divergence between the model's predicted distribution and the reference distribution of unlabeled data, with P representing the model's predicted distribution and Q representing the reference distribution of unlabeled data.

This loss function aims to minimize the KL divergence, thereby making the model's prediction distribution as close as possible to the ground-truth label classification distribution of labeled data, improving the model's performance on unlabeled data.

5. Experiments and results

5.1. Description of dataset

The xBD dataset is the largest building damage assessment dataset [20], consisting of images sourced from the Maxar/DigitalGlobe Open Data Program. This dataset encompasses over 453,610 square kilometers and includes 850,736 building instances. xBD provides building polygons, labels of damage levels, and high-spatial-resolution (HSR) bitemporal optical satellite images with dimensions of 1024 × 1024 pixels and a ground sample distance (GSD) of less than 0.8 m, capturing scenes before and after various disaster events. To assess building damage across multiple disaster types, xBD employs the joint damage scale, developed with the assistance 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. The joint damage scale comprises four discrete damage levels: no damage, minor damage, major damage, and destroyed, serving as the criteria for damage classification.

This study focused on pure flood events and manually selected relevant samples from the xBD dataset that were classified as flood. Specifically, the data originates from two major disaster events: the US Midwest floods (occurring January–May, 2019) and the India/Nepal floods (occurring July–September, 2017). The approximate geographic sampling locations for these events are illustrated in Fig. 6(a) and 6(b), respectively.

The flood dataset comprises 1064 pairs of high-spatial-resolution (HSR) remote sensing images. To ensure robust evaluation and prevent geographical bias in the splits, the image pairs were randomly divided into training (60%), validation (20%), and test (20%) sets. Subsequently, all images were cropped into 256 × 256-pixel blocks. To augment the training data, image cropping on

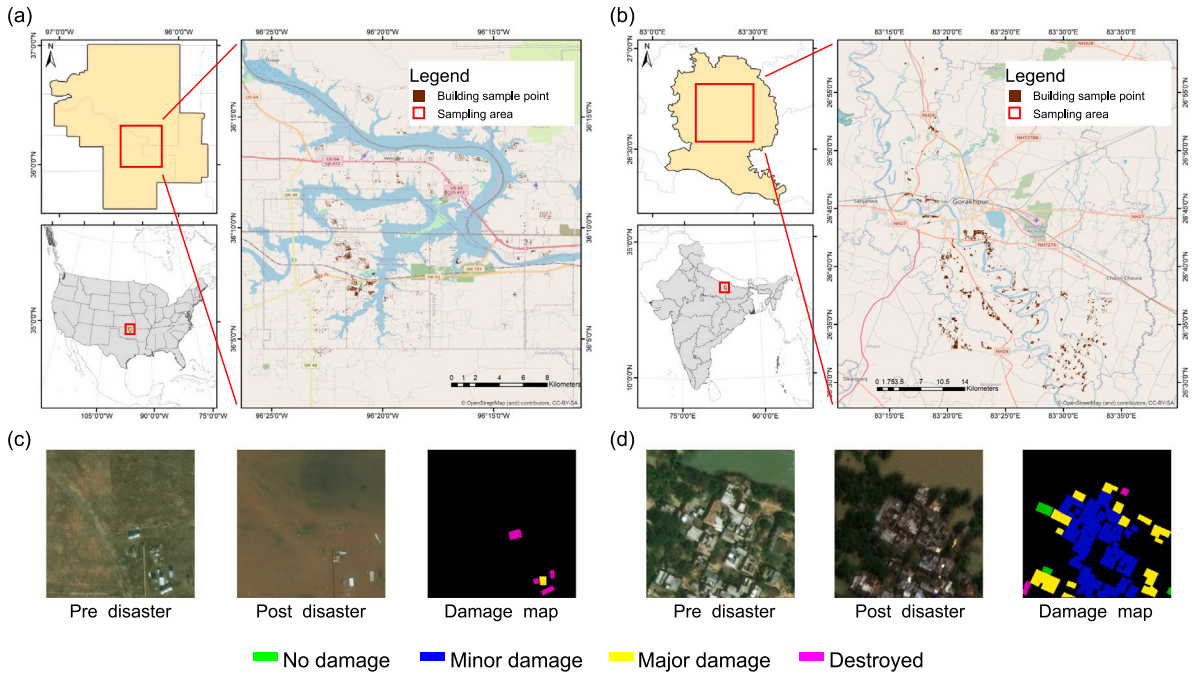


Fig. 6. Geographic locations and building damage examples from the flood dataset used in this study. (a) Affected areas for the US Midwest flood events. (b) Affected area for the Nepal flood event. (c) Examples of building damage classifications (No Damage, Minor, Major, Destroyed - Pre/Post/Label) from the US Midwest events. (d) Examples of building damage classifications from the Nepal event.

Table 1
Proportion of building pixels per damage class for all events and flood events in the xBD dataset.

Included disaster events	All events	Flood events
Original sample number	22 068	2128
No-damage class ratio	87.21%	91.68%
Minor-damage class ratio	4.57%	4.28%
Major-damage class ratio	5.48%	3.71%
Destroyed class ratio	2.74%	0.33%

the training set was performed with a stride of 128 pixels, while a stride of 256 pixels (no overlap) was used for the validation and test sets. These operations resulted in 93,786/10,224/10,224 image patches for the respective sets.

Visual examples illustrating the challenges inherent in Flood-BDA are shown in Fig. 6(c) and (d). Example (c), from the US Midwest event, highlights difficulties such as the small proportion of pixels corresponding to damaged labels within the overall image patch and the close spatial proximity of visually similar but distinct damage classes (e.g., 'major damage' adjacent to 'destroyed'), making classification ambiguous. Example (d), from the India event, showcases another key challenge: the subtle nature of visual changes between pre- and post-disaster imagery. For many buildings, damage is not immediately apparent through visual inspection alone, making it difficult to distinguish between different damage levels based purely on observable pixel differences. These visual difficulties are compounded by the statistical rarity of severe damage classes ('major damage' and 'destroyed') in flood scenarios, which, as shown by the pixel ratios in Table 1, are even less frequent compared to their prevalence across all disaster types in the original xBD dataset.

5.2. Implementation details

The experiment utilized the PyTorch framework and trained on a single NVIDIA RTX 4090 GPU. The Adam optimizer was selected.

Fundamental hyperparameters, including the initial learning rate and batch size, were determined following established practices where early-stage training dynamics serve as a reliable indicator for tuning [47]. Through systematic observation across initial experimental trials, the initial learning rate was set to 0.00003 and the batch size was configured at 24. The learning rate was subsequently reduced by 80% every 60 iterations using a step decay schedule. All models were trained for a total of 150 epochs.

Regarding the proposed SPADANet, the prior attention module incorporates an energy function influenced by hyperparameters m and n. The ratio $\frac{n}{m}$ was set to 0.5 (only the ratio affects the outcome). The energy function bias term λ for the prior attention

module is set to 0.0001 in the supervised learning setup. The size of the convolution kernel is set to 3 × 3, and the number of kernels in each convolution unit for the basic UNet is set to {16, 32, 64, 128, 256}.

In the semi-supervised learning setup, the sampling ratios for the dataset were set to {5%, 10%, 20%, 50%}. The weights α and β for the three loss functions were set to 0.001. In strategy 4, since three different reference distributions are used to calculate the loss terms, there are three KL loss weights, $\beta_1, \beta_2, \beta_3$, all set to 0.001. The buffer size n for the distribution buffer used to store the reference distributions was set to 10.

5.3. Evaluation metrics

This research employed overall accuracy (OA), precision, recall, F1 score, Intersection over Union (IoU) and kappa as evaluation metrics. The definitions of these metrics are as follows:

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1\ Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{13}$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \tag{14}$$

$$p_o = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$p_e = \frac{(TP + FP) \cdot (TP + FN)}{(TP + TN + FP + FN)^2} + \frac{(FP + TN) \cdot (FN + TN)}{(TP + TN + FP + FN)^2} \tag{16}$$

In these formulas, TP (True Positive) represents the number of positive samples correctly predicted by the model, FP (False Positive) signifies the number of negative samples incorrectly predicted as positive, and FN (False Negative) indicates the number of positive samples incorrectly predicted as negative. p_o is the observed classification consistency, while p_e is the expected classification consistency. Note that higher F1 score, OA, and kappa indicate better overall performance.

In this study, the meaning of the metrics is as follows:

- Precision: A metric that represents the proportion of true positive samples among those predicted as positive by a classification model. A higher value indicates that the model can more accurately identify damaged buildings. Still, it may also imply that the number of positive samples found is very low.
- Recall: A metric that represents the proportion of true positive samples among all actual positive samples. A higher value suggests that the model can detect more damaged buildings. Still, it may also result in many undamaged buildings being incorrectly classified as damaged or more severe misclassification of damage levels.
- F1-score: This is a comprehensive evaluation metric used when Precision and Recall have varying performances. It balances the trade-off between the two.
- IoU: This parameter serves as an indicator of the model's performance on individual patches, with higher values indicating more accurate predictions.
- Overall Accuracy (OA): This metric represents the proportion of correctly classified samples among the total samples. However, in cases of data imbalance, the majority class has a more significant influence on this metric. The undamaged class is more critical in our experiment than the damaged class.
- Kappa: This metric measures the agreement between the model's predictions and the ground truth, correcting for the agreement that would be expected purely by chance. Unlike OA, which can be misleadingly high on imbalanced datasets if the model simply predicts the majority class, Kappa explicitly accounts for the possibility of correct predictions occurring randomly based on the marginal distributions of predicted and actual classes.
- Parameters (Params): The size of the model's parameters indicates the ease of deployment and the time required for computational inference. Lower parameter values offer more significant advantages in terms of computational efficiency.
- Giga Floating-point Operations Per Second (GFLOPs): This parameter measures the computational complexity of the model per forward pass. Higher GFLOPs indicate a more computationally intensive model, typically requiring greater resources and often leading to longer processing times per iteration during training and inference.

It is noteworthy that, despite the use of numerous evaluation metrics to measure model performance, in DA tasks, this paper considers the metrics for the damaged classes (minor damage, major damage, destroyed) to be more important than those for the undamaged class. When overall metrics are similar, recall is considered more important than precision, as a more comprehensive and rapid identification of damaged buildings during the post-disaster rapid rescue phase can provide strong support for rescue operations.

5.4. Comparative methods

To validate the effectiveness of the proposed method, this research conducted a comparative analysis using some representative CD methods. These methods were deliberately chosen to represent a diverse range of foundational architectural principles, allowing for a systematic evaluation of their suitability for the DA task. All methods were trained under identical settings (e.g., learning rate, optimizer, number of epochs) for a fair comparison. The selected models and the architectural principle each represents are as follows:

(1) UNet [33]: Represents the classic encoder–decoder architecture with skip connections. Its proven effectiveness in high-resolution semantic segmentation makes it a fundamental baseline for any pixel-level prediction task.
(2) CDNet [29]: Represents early fully convolutional network (FCN) designs applied to change detection. It serves as a baseline for non-UNet, deconvolutional approaches.
(3) FC-siam-conc & FC-siam-diff [30]: Represent the Siamese network paradigm, a common and effective structure for CD tasks. These models extend the FCN approach by using twin networks to process bi-temporal images, testing two common feature fusion strategies (concatenation vs. differencing).
(4) LUNet [31]: Represents explicit temporal modeling using Recurrent Neural Networks (RNNs). By incorporating LSTM blocks, LUNet is designed to capture sequential relationships between temporal images, a distinct approach from purely convolutional or attention-based methods.
(5) SNUNet [32]: Represents the integration of nested, dense skip connections (UNet++) and a self-attention mechanism. This allows us to evaluate the impact of more complex U-Net variants and the general effectiveness of self-attention in a hybrid architecture.
(6) BIT [24]: Represents a purely Transformer-based self-attention approach. In contrast to hybrid models like SNUNet, BIT treats image features as semantic tokens, allowing us to assess the performance of a vision transformer architecture when applied to the Flood-BDA context.
(7) P2V [26]: Represents an alternative temporal modeling strategy based on pseudo-video generation. This method converts the bi-temporal CD problem into a video understanding task, offering a different perspective on capturing change over time.

5.5. Results and analysis

This section will present part of the results of the SSL experiment in Section 5.5.1 (the complete raw results can be referred to in supplementary material Table S3 to S22). In this subsection, it will be explained that the image-level consistency regularization method adopted in this paper can obtain additional information from unlabeled data to enhance model performance. Moreover, by comparing pseudo-label predictions, ground-truth label predictions, and ground-truth labels to form different image-level reference distributions, the results show that using the reference distribution composed of pseudo-label predictions can produce better improvements. Furthermore, in Section 5.5.2, this paper conducts a supervised learning experiment. A comparative experiment was carried out in the model that provides supervisory signals for the SSL experiment, aiming to explore whether the design of post-flood DA tasks should use different prior methods to capture relevance compared to CD task models. The numerical experimental results indicate that under the background of using DL models, Flood-BDA tasks should adopt different model design strategies due to the differences in application scenarios. Specifically, the specific focus on the recall should be an additional concern for the design of DA task models.

5.5.1. Semi-supervised learning results

This paper applied four consistency regularization strategies to modify nine networks, including SPADANet. Due to the imbalance of the dataset used in this study, OA as a comprehensive evaluation metric was not adopted. Instead, F1 score, and kappa value were chosen as the evaluation metrics. Considering that all four levels of building damage in DA tasks are equally important, this study used the average of the F1 scores of all four categories to evaluate the model's performance. This choice reflects the importance of accurately classifying buildings with different damage levels in DA and ensures that the model's performance on all categories is fairly and evenly considered.

It was challenging to present complete numerical results in the main text (full numerical results can be found in supplementary material Table S3 to S22).

Fig. 7 presents a boxplot of 160 kappa values obtained from experiments with different base methods (excluding SNUNet, which failed to learn effective features, as detailed in the supplementary material) combined with various image-level consistency regularization strategies across different label ratios (32 results for base methods and 128 for improved methods). The boxplot visually demonstrates the positive and negative impacts of SSL on each model.

Statistical analysis shows that among the improved methods, Strategy 1 resulted in 21 positive impacts and 11 negative impacts, Strategy 2 in 15 positive and 17 negative impacts, Strategy 3 in 15 positive and 17 negative impacts, and Strategy 4 in 20 positive and 12 negative impacts. These results indicate that SSL generally improves model performance, and Strategies 1 and 4 are more likely to yield positive impacts compared to Strategies 2 and 3. This demonstrates that image-level consistency regularization can effectively extract useful information from unlabeled data.

It is noteworthy that from Fig. 7, it can be observed that the effects produced by different strategies at various label ratios are complex and varied for each model. When examining models with upper quartile and median values above the boxplot, SPADANet

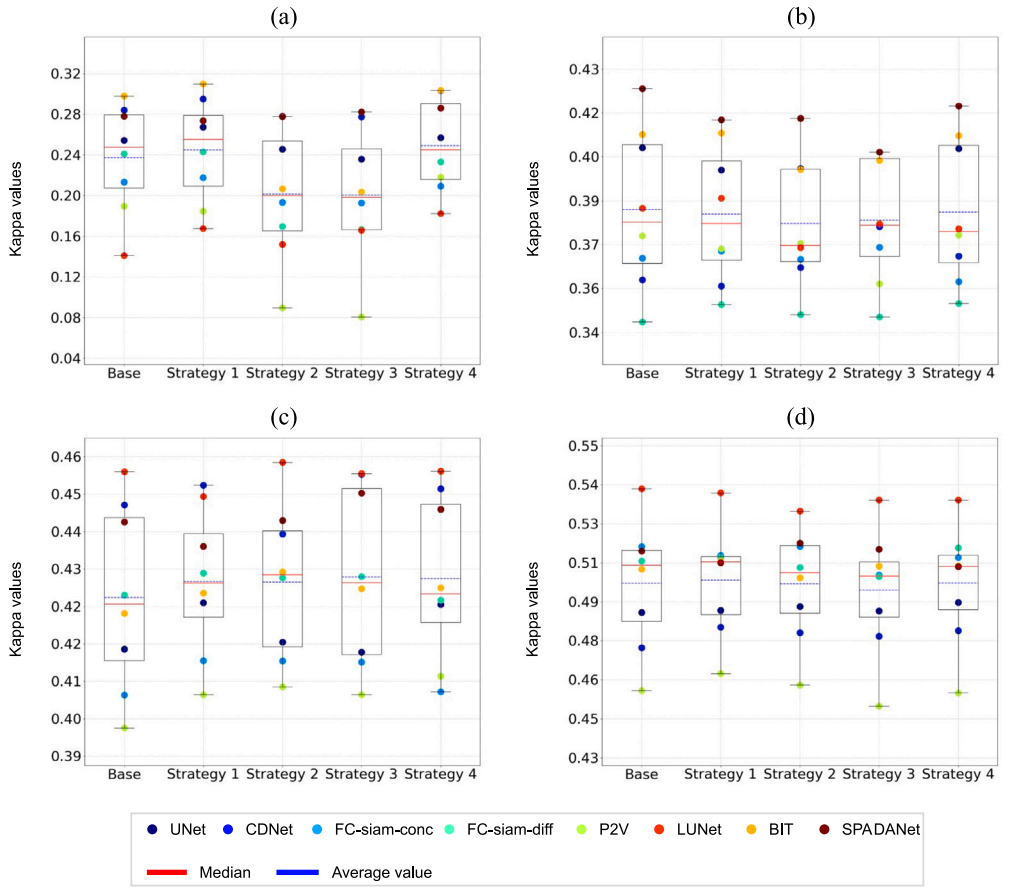


Fig. 7. Boxplot of kappa values for all models with different label ratios and strategies. The red solid line represents the median, the blue dashed line represents the average. Figs. (a), (b), (c), and (d) depict the results for the 5%, 10%, 20%, and 50% label ratio datasets, respectively.

Table 2
Average performance of different strategies across different proportions of labeled datasets.

Method	Label Ratio							
	5%		10%		20%		50%	
	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)
Base	37.36	23.74	43.46	38.71	46.53	42.59	51.78	50.20
Strategy 1	37.76	24.49	43.25	38.55	46.79	42.93	51.90	50.33
Strategy 2	35.77	20.17	43.22	38.22	46.91	42.92	51.82	50.19
Strategy 3	35.71	20.07	43.27	38.33	46.91	43.03	51.69	49.95
Strategy 4	37.96	24.91	43.49	38.62	46.76	42.99	51.96	50.22
Supervised	**F1=52.54** and **Kappa=52.52**							

Best Second Third

stands out as the only model that consistently achieves high performance across different label ratios. This phenomenon may reveal an important direction for future model development: seeking architectures that maintain stable performance across different datasets and label ratios.

To further investigate the impact of each SSL strategy, this paper provides more detailed results of the mean and peak values from Fig. 7.

The average performance in Table 2 supports two key conclusions from Fig. 7:

(1) Image-level consistency regularization generally has a positive impact on model performance. Specifically, in terms of kappa values, Strategy 4 outperforms the base model by 1.17% at a 5% label ratio, while Strategy 3 shows a 0.44% improvement at a 20% label ratio, and Strategy 1 achieves a 0.13% gain at a 50% label ratio. These results indicate that consistency regularization becomes more effective as the proportion of unlabeled data increases.

Table 3
Best quantitative results of different methods and labeling ratios.

Method	Label Ratio							
	5%		10%		20%		50%	
	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)	F1 (%)	Kappa (%)
UNet	37.10	25.43	44.98	40.81	44.97	41.49	51.38	49.09
CDNet	38.23	28.40	42.18	36.30	47.98	44.57	50.64	47.73
FC-conc	34.78	21.35	43.59	37.04	46.16	40.51	52.93	51.62
FC-diff	36.87	24.11	40.76	34.86	47.42	42.64	52.86	51.07
SNUNet	17.90	0.00	17.99	0.00	17.90	0.00	17.83	0.00
P2V	36.17	18.96	44.12	37.80	44.71	39.80	49.83	46.08
LUNet	34.36	14.11	42.78	38.75	<u>48.41</u>	<u>45.28</u>	<u>53.21</u>	**53.85**
BIT	<u>42.70</u>	<u>29.80</u>	43.33	41.27	46.11	42.25	50.32	50.75
SPADANet	38.65	27.80	**45.94**	**42.83**	46.47	44.21	53.10	51.45
BIT + Strategy 1	**43.21**	**30.97**						
SPADANet + Strategy 4			<u>45.28</u>	<u>42.24</u>				
LUNet + Strategy 2					**49.16**	**45.48**		
LUNet + Strategy 1							**53.24**	<u>53.69</u>

Best <u>Second</u>

(2) Using pseudo-label predictions from unlabeled data to form the reference distribution yields greater positive impacts (strategies definition refer to Section 4.2). Comparing Strategy 1 and 4 with Strategy 2 and 3, the former two strategies exhibit better average performance at 5%, 10%, and 50% label ratios. Although this trend is less pronounced at a 20% label ratio, the differences in metrics among the strategies are minimal.

Table 3 highlights the best-performing methods enhanced by consistency regularization, corresponding to the peak performance in Fig. 7. For a 5% label ratio, the BIT + Strategy 1 approach outperformed the second-best BIT method, with a 0.51% improvement in F1 score and a 1.17% improvement in kappa value. At the 10% label ratio, SPADANet + Strategy 4 showed a decline of 0.66% in F1 score and 0.59% in kappa value compared to the best-performing SPADANet. At the 20% label ratio, LUNet + Strategy 2 emerged as the top performer, with a 0.75% improvement in F1 score and a 0.2% improvement in kappa value over the second-best LUNet. At the 50% label ratio, LUNet + Strategy 1 exhibited a marginal 0.03% improvement in F1 score but a slight 0.16% decrease in kappa value compared to LUNet. Metric improvements were observed at 5%, 20%, and 50% label ratios, while a negative impact occurred at 10%.

Visual inspection of the SSL results, exemplified in Fig. 8, provides further qualitative insights into the performance and limitations of these methods across different scenarios and label ratios: Fig. 8(a) showcases the BIT backbone combined with different SSL strategies at the 5% label ratio. Comparing the baseline BIT (column 5) with the BIT + Strategy 1 (column 6), there is a clear improvement. This configuration is also the numerically best method, significantly reduces the misclassifications of 'major damage' pixels as 'no damage'. This visually confirms the positive impact of SSL Strategy 1 in low-data regimes for a strong backbone like BIT, outperforming other SSL combinations visually in this example. Fig. 8(b) is a negative example, SPADANet is the backbone at 10% labels. While SPADANet+Strategy 4 yielded the best overall quantitative score, this specific visual example reveals a potential drawback. Strategy 4 appears to misclassify a significant portion of the 'major damage' area (yellow in GT) as 'minor damage' (blue). Fig. 8(c) uses the LUNet backbone at 20% labels, focusing on the rare 'destroyed' class (magenta in GT). The baseline LUNet struggles significantly with this class (low IoU/F1). However, combining LUNet with SSL Strategy 2 (the numerically best combination here) demonstrates a dramatic improvement in identifying the 'destroyed' building. Fig. 8(d) examines performance at 50% labels, focusing on 'minor damage' (blue). It reveals a general weakness across most models and SSL strategies in identifying 'minor damage' within complex scenes containing dense building clusters and vegetation. In this challenging example, only the baseline BIT model manages to correctly identify some of the 'minor damage' buildings, while LUNet and its SSL combinations fail to detect them accurately.

The positive examples (Fig. 8a, c) often occur in scenarios with relatively isolated buildings, potentially surrounded by water, where SSL appears effective at refining classifications. Conversely, the negative or mixed examples (Fig. 8 b, d) highlight challenges in complex scenes where water bodies, vegetation, and buildings are adjacent to each other, and potential degradation in performance for intermediate classes ('minor', 'major').

5.5.2. Supervised learning results

Since this paper focuses on a specific downstream task (Flood-BDA) and DA models are often inherited from CD models, the SSL experiments in this study primarily use CD models to provide supervised signals. This raises the question of whether post-flood DA model design should differ from traditional CD approaches. Clarifying this point can guide future DA model design, helping to decide whether to directly transfer CD models or develop task-specific designs. To address this, this section conducted supervised learning experiments comparing selected CD methods with the proposed SPADANet.

Given that the dataset contains four damage levels, no single model can achieve optimal performance across all categories. Therefore, this section converted the four-class results into binary results (full results can be found in supplementary material Table S1), classifying building damage caused by floods as 'no damage' or 'damaged' (combining 'minor damage', 'major damage', and 'destroyed').

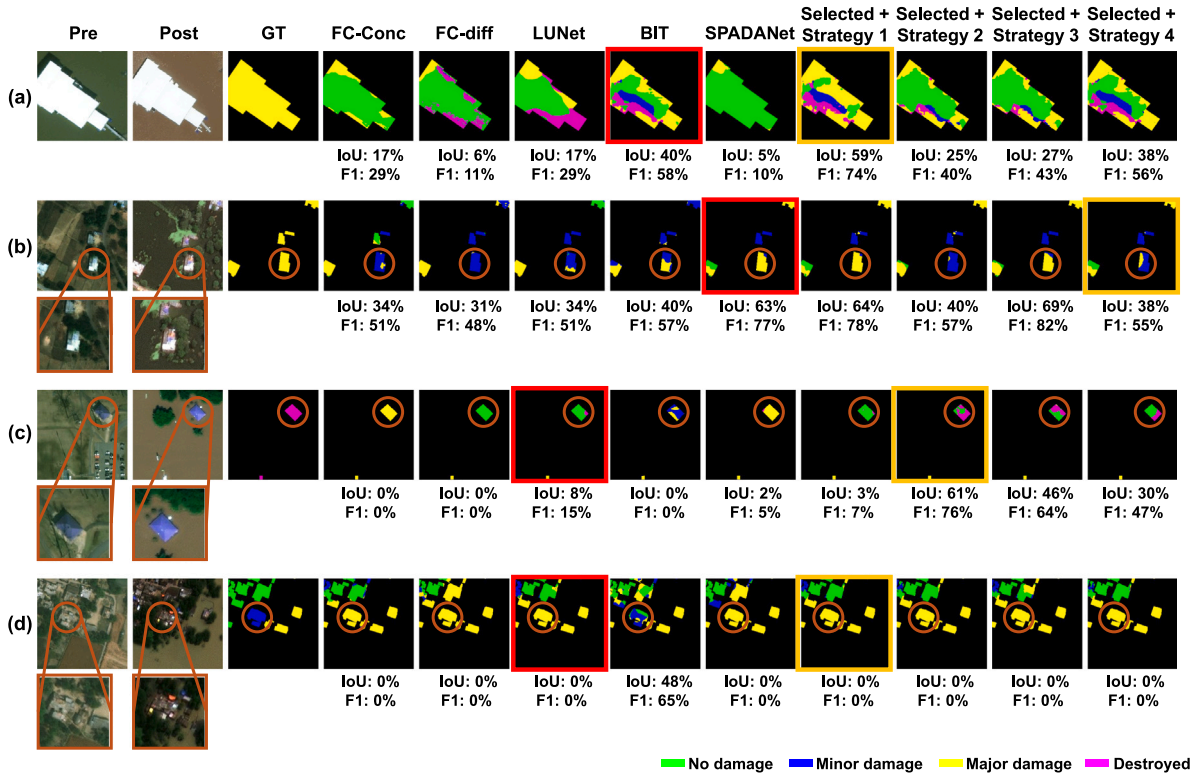


Fig. 8. Visual comparison of four-class semi-supervised learning results using different methods and labeling ratios. Rows correspond to scenarios using 5% (a), 10% (b), 20% (c), and 50% (d) labeled data, respectively. Columns display Pre/Post images, Ground Truth (GT), backbone model results, and results enhanced with various SSL strategies. Highlights: Brown boxes show zoomed image of critical parts. Red boxes show example results involving the backbone combined with SSL. Orange boxes indicate the method achieving the best quantitative performance for that label ratio according. Metrics Shown: Representative IoU and F1 scores are displayed for specific examples: 'major damage' classification in rows (a), (b), "destroyed" in (c), and 'minor damage' in (d). Color legend defines damage classes".

Table 4
Comparative analysis of supervised models for binary Flood-BDA. "Dmg." denotes "damage.".

Network	Precision (%)		Recall (%)		F1 Score (%)		OA	Kappa	Params	GFLOPs
Name	No Dmg.	Dmg.	No Dmg.	Dmg.	No Dmg.	Dmg.	(%)	(%)	(M)	
UNet	93.44	70.35	94.01	68.27	93.73	69.29	89.58	63.02	**1.35**	**3.61**
CDNet	93.30	69.88	93.30	69.88	93.74	68.22	89.54	61.96	1.43	24.09
FC-conc	92.55	76.79	95.87	63.88	94.18	69.74	90.24	63.98	1.55	5.36
FC-diff	93.24	70.83	94.40	66.53	93.81	68.61	89.66	62.43	**1.35**	4.76
SNUNet	92.21	72.71	95.36	60.54	93.76	66.07	89.46	59.89	10.2	44.4
P2V	93.29	71.59	94.66	66.40	93.97	68.90	89.90	62.88	5.42	32.97
LUNet	92.30	**78.66**	**96.57**	61.10	94.39	68.78	90.49	63.27	9.45	17.34
BIT	93.10	75.88	95.76	65.29	**94.41**	70.19	**90.59**	64.64	3.03	8.81
SPADANet	**95.53**	64.94	91.27	**79.10**	93.35	**71.32**	89.20	**64.75**	**1.35**	**3.61**

Best Second

Table 4 presents the binary classification results. SPADANet significantly improves recall to 79.10%, outperforming the second-best model CDNet by 9.22% and the backbone UNet by 10.83%. Despite lower precision compared to CD methods, SPADANet achieves the best overall F1 score, surpassing the second-best model BIT by 1.13%. In terms of kappa values, SPADANet scores 64.75%, 0.11% higher than BIT. Additionally, SPADANet maintains the lowest parameter count (1.35M) and computational cost (3.61 GFLOPs) without increasing the backbone UNet's complexity.

The visual results presented in Fig. 9 offer qualitative insights into the performance of SPADANet compared to other supervised methods, highlighting both strengths and areas where challenges remain.

Fig. 9(a) and (b) demonstrate SPADANet's effectiveness in reducing false negatives (missed damage) compared to the baseline UNet and other methods. Fig. 9(a) shows significant improvement where the baseline struggled, while Fig. 9(b) shows refinement even when the baseline performance was already high. The positive Fig. 9 (a, b) often depict scenarios where buildings are clearly surrounded by water. In these cases, SPADANet's tendency to give more complete predicted area (driven by the prior attention

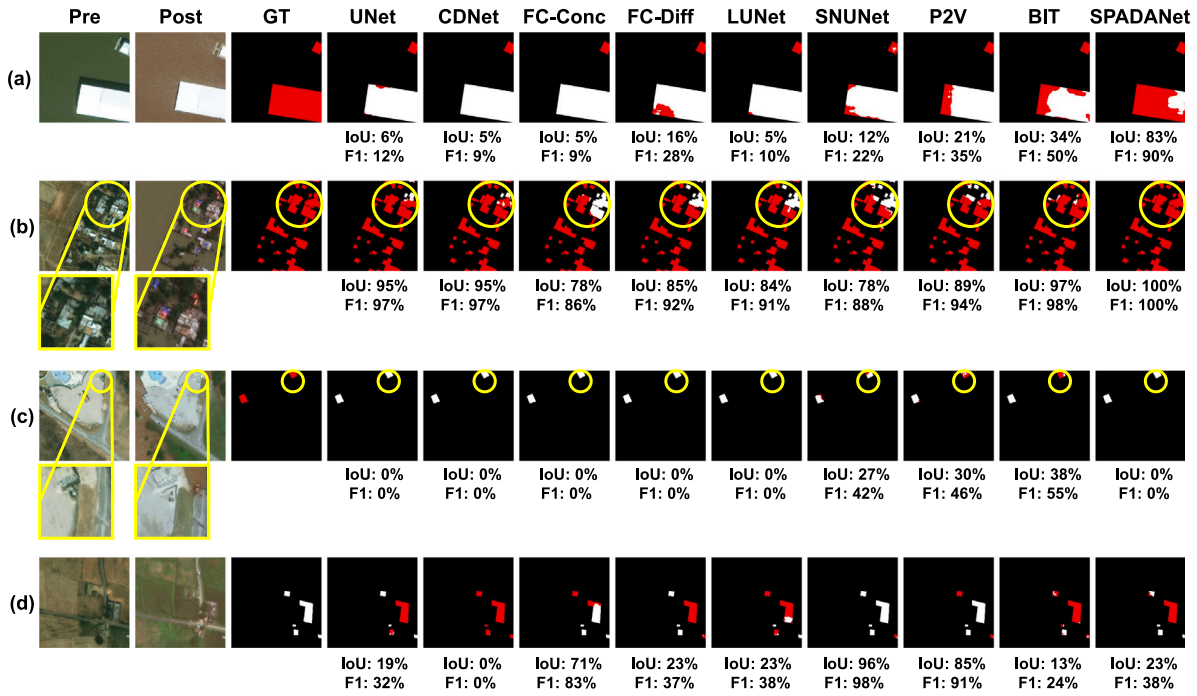


Fig. 9. Visual comparison of binary classification supervised learning results. White denotes undamaged buildings, and red denotes damaged buildings. "Pre" and "Post" refer to satellite images before and after the disaster. "GT" denotes Ground Truth. Yellow boxes show zoomed image of critical parts.

module enhancing subtle changes) leads to superior recall compared to other methods. Fig. 9(c) highlights a failure case where SPADANet, like most other methods shown, fails to identify the damaged building. Observing the original pre- and post-disaster images for this example reveals a potential reason: the building appears visually very similar to surrounding concrete or bare ground in the satellite imagery, making it difficult to identify. This ambiguity likely contributes to the model's failure to detect the damage. Fig. 9(d) SPADANet incorrectly classifies undamaged buildings as damaged. Examination of the source images shows complex environmental factors: while widespread water might not be directly visible on the buildings, the surrounding land cover (vegetation, soil, roads) exhibits significant color and texture changes between the pre- and post-disaster images, characteristic of post-flood conditions. The model may be reacting to these strong contextual changes in the surroundings, leading to false positive predictions on the buildings themselves.

Although binary results are used in this section, it is important to note that DA tasks are inherently multi-class. To better explore multi-class performance, Fig. 10 presents normalized confusion matrices for four-class results. Fig. 10(a) to (g) show traditional CD methods, while Fig. 10(h) shows SPADANet. SPADANet outperforms in the 'minor damage' and 'major damage' categories. However, the limitations of directly transferring CD methods to DA tasks are evident in other two categories. For example, SPADANet's performance in the 'no damage' category is lower, but since undamaged buildings are not the primary focus in post-disaster scenarios, high accuracy in this category has limited practical impact. In the 'destroyed' category, both CD methods and SPADANet perform poorly (below 20%). CD models tend to misclassify damaged buildings as 'no damage', as seen in the first column of Fig. 10(a) to (g). In contrast, SPADANet's misclassifications are more concentrated in adjacent categories, as shown in Fig. 10(h). In real-world applications, using CD models could lead to severe misallocation of rescue resources, as many severely damaged buildings might be classified as undamaged. While SPADANet also struggles to accurately predict the 'destroyed' category, it tends to classify them as 'major damage', which may not cause this problem. These confusion matrices results clarify the adaptability of SPADANet in post flood DA scenario.

6. Discussion

6.1. Defining task-specific design principles for post-flood damage assessment

A core motivation for this study stems from the hypothesis that DLCD and DLDA, particularly for post-flood scenarios, require distinct design considerations despite their shared heritage. The comprehensive benchmarking experiments, evaluating both transferred DLCD methods and novel semi-supervised and supervised approaches tailored for Flood-BDA, provide substantial evidence supporting this distinction:

Task Objective: The supervised learning results clearly illustrate a divergence in optimization focus. Early DL methods applied to CD, such as UNet [33] and CDNet [29], exhibit relatively balanced precision and recall when benchmarked on flood dataset

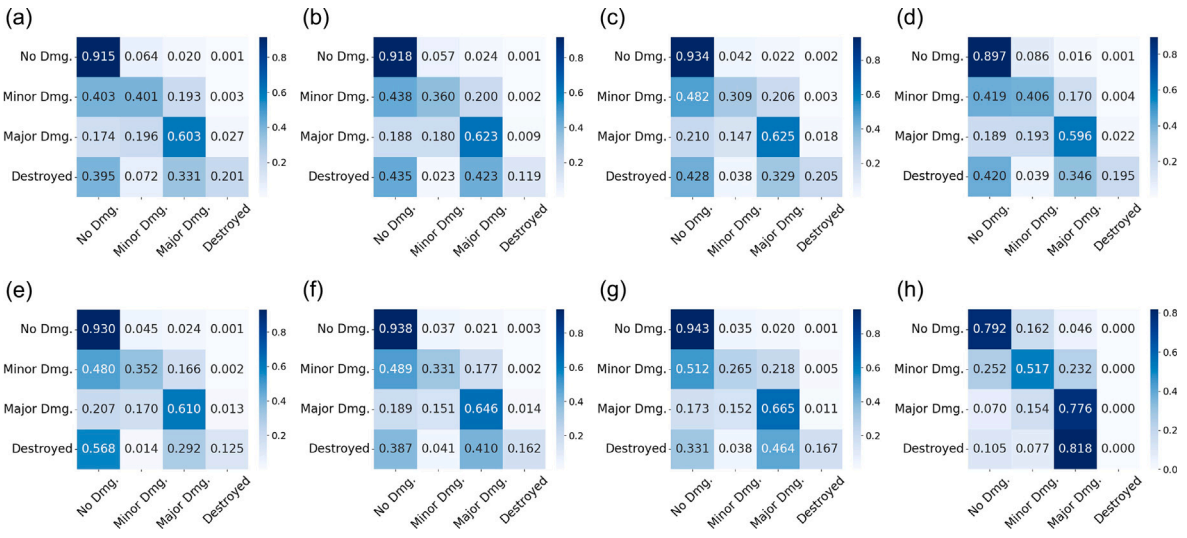


Fig. 10. Normalized confusion matrices for four-class results using different methods. (a) UNet. (b) CDNet. (c) FC-siam-conc. (d) FC-siam-diff. (e) LUNet. (f) P2V. (g) BIT. (h) SPADANet. The *y*-axis represents the ground-truth labels, the *x*-axis represents the predicted labels, and the values in the cells indicate the proportion of predictions for each category in the overall dataset. "Dmg". denotes "damage".

(Table 4). However, as DLCD research progressed, incorporating more sophisticated techniques like siamese architectures [30], temporal modeling [26,31], and various self-attention mechanisms [24,32], a trend towards prioritizing precision often emerged, likely driven by the equally important goals of the changed and unchanged categories in the CD scenario. When these advanced CD models are directly applied to the Flood-BDA task, this precision preference persists, often at the expense of recall. However, in post-disaster DA, this balance is inappropriate. The failure to identify a damaged building (a false negative) can prevent the allocation of life-saving resources, making it an unacceptable error [2]. Consequently, high recall is the paramount objective. By incorporating a parameter-free prior attention module aimed at amplifying subtle, outlier-like features, SPADANet demonstrates a successful shift towards a high-recall operational profile. As shown in the confusion matrices (Fig. 10), its error pattern is also more operationally benign; unlike many CD models that misclassify destroyed buildings as no damage, SPADANet's errors are concentrated between adjacent damage classes (e.g., destroyed as major). This high-recall performance and more favorable error distribution define its high adaptability to the DA application context.

Data Characteristics: This paper creates benchmarks based on two characteristics of Flood-BDA. First, Label scarcity and dataset imbalance are common problems in CD and DA. Among SSL paradigms that can address this, consistency regularization has been notably under-explored, especially within the DA domain. This work provides, to our knowledge, the first benchmark evaluation of consistency regularization, specifically using the image-level strategy, for Flood-BDA. The positive results demonstrate that this approach effectively leverages unlabeled data, highlighting SSL's significant potential as a crucial strategy for overcoming data limitations in DA. Second, Flood damage frequently manifests as subtle visual alterations, unlike the more distinct changes often targeted by CD. The supervised benchmark experiments reveal that directly transferred CD methods struggle to effectively capture these subtle cues for DA.

This study confirms that simply transferring DLCD methods to Flood-BDA is suboptimal due to fundamental differences in task objectives, and data characteristics. CD models often discard subtle information critical for DA and may exhibit undesirable error modes. The benchmark results highlight the potential of alternative approaches tailored to DA: prior attention mechanisms show promise for handling subtle changes and boosting recall, while image-level consistency regularization offers an effective pathway to address data scarcity by leveraging unlabeled data. These findings provide clear directions for future DA research, emphasizing the need for task-specific model design and evaluation rather than relying solely on advancements from the general CD field. SPADANet and the image-level SSL framework presented here serve as initial, validated steps along these DA-specific research avenues.

6.2. Effect of image level consistency regularization

This paper develops an SSL approach leveraging image-level consistency regularization to address the challenges of data scarcity and long-tail distribution in post-flood DA tasks. While existing SSL research often focuses on consistency derived from intra-image augmentations [23,40,41], this work explores the potential of using statistical distributions derived from groups of images as the reference for consistency, given the increasing availability of unlabeled remote sensing data from disaster events.

The results presented quantitatively (Tables 2, 3, S3–S22) and qualitatively (Fig. 7, 8) demonstrate the general positive impact of this image-level consistency regularization approach. Model performance typically improves with SSL, particularly as the proportion of labeled data decreases (Table 2), indicating effective utilization of unlabeled data in low-resource scenarios. Even well-performing baseline models show potential for enhancement through these techniques (Table 3).

Table 5
Ablation study on the placement of the prior attention module.

Network Name	Precision (%)	Recall (%)	F1-score (%)
UNet	**70.35**	68.27	69.29
SPADANet-Conv	62.26	77.68	69.12
SPADANet-Up	62.62	77.01	69.07
SPADANet	64.94	**79.10**	**71.32**

A key finding emerges when comparing the four reference distribution strategies: Strategies 1 and 4, which construct the reference distribution Q using pseudo-label predictions generated by the model itself on unlabeled data, generally yield greater positive impacts than Strategies 2 and 3, which rely on distributions derived from the ground-truth labels or model predictions on the limited labeled set. This suggests that using a reference distribution that is directly adapted to the characteristics of the unlabeled data pool provides a more effective learning signal, especially in low-label regimes (e.g., 5%, see Table 2, Fig. 7).

Despite the acknowledged limitations inherent in using pseudo-labels – namely the potential for error propagation if the base model is weak and the lack of exhaustive tuning for SSL-specific hyperparameters like confidence thresholds – the empirical results presented offer valuable insights. This work's primary goal in exploring SSL was to investigate whether the consistency regularization paradigm, largely unexplored in DA, could help bridge the methodological gap between these fields by effectively leveraging abundant unlabeled data. The findings suggest that image-level consistency regularization, particularly when guided by adaptive pseudo-label distributions (Strategies 1 & 4), does indeed provide a positive impact on model performance under the data-scarce conditions typical of Flood-BDA.

This research mitigated the risk of error propagation by initializing the SSL process with strong, pre-trained CD networks identified through the initial benchmarking (Subsection 5.5.2), ensuring a reasonably reliable starting point for pseudo-label generation. While this research did not implement complex mechanisms like confidence thresholding, the consistent improvements observed across various strong backbones (Fig. 7, Tables S3–S22) provide preliminary validation for the effectiveness of the image-level consistency regularization approach itself.

Therefore, this study successfully demonstrates the potential of applying image-level consistency regularization to Flood-BDA. It serves as an important first step in characterizing the performance of this SSL paradigm in this context, showing its ability to extract useful information from unlabeled data.

6.3. Effect of prior attention

The investigation into suitable supervised signal in SSL, motivated by the need for task-specific designs rather than direct transfer from CD methods [17,21,32,34], led to the development of SPADANet. As discussed (Section 2.1, 2.3), while DA research often inherits CD architectures like U-Net [24,32,34] and attention modules [17,24], the specific priors embedded may not align with DA's requirement to capture subtle damage cues and prioritize recall. SPADANet combines the robust U-Net structure with a parameter-free prior attention module [25] specifically chosen for its theoretical potential to amplify subtle, outlier-like pixel changes characteristic of flood damage—a less explored approach compared to standard self-attention.

The supervised learning results demonstrate the effectiveness of this approach. In binary classification (Table 4), SPADANet achieves state-of-the-art recall (79.10%), significantly outperforming strong CD baselines like CDNet (+9.22%). This high recall, achieved while maintaining the lowest parameter count (1.35M) and computational cost (3.61 GFLOPs), directly addresses this research claimed requirement of DA, enabling models to be rapidly deployed at low cost and places greater emphasis on recall. SPADANet also achieves the highest overall F1 score and Kappa value, indicating strong overall performance.

To further explore the role of the prior attention module, this study conducted ablation studies by integrating it into different parts of the U-Net architecture:

(1) UNet: The baseline U-Net without any prior attention module.
(2) SPADANet: The proposed SPADANet, with the prior attention module added to skip connections.
(3) SPADANet-Up: A variant of SPADANet, with the prior attention module added to upsampling.
(4) SPADANet-Conv: A variant of SPADANet, with the prior attention module added to Conv blocks.

Table 5 confirmed the optimal placement within skip connections, suggesting that retaining subtle information lost during down sampling is critical. The module works by amplifying outlier neurons; while this can introduce noise, placing it in skip connections appears to balance the retention of useful subtle information against noise introduction effectively.

Qualitative analysis provides further nuance. The visual examples in Fig. 9 confirm SPADANet's ability to drastically reduce false negatives compared to other methods, especially in scenarios with clearer damage indicators (Fig. 9a, b). However, they also reveal limitations: SPADANet can still miss damage in low-contrast environments (Fig. 9 c) and may produce false positives when strong contextual changes occur near undamaged buildings (Fig. 9 d).

Further insights come from the multi-class confusion matrices (Fig. 10). Compared to traditional CD methods (Fig. 10a–g) which exhibit a strong tendency to misclassify damaged buildings (minor, major, destroyed) as 'no damage' (high values in the first column of rows 2–4), SPADANet (Fig. 10h) shows a markedly different error pattern. While still struggling with the 'destroyed'

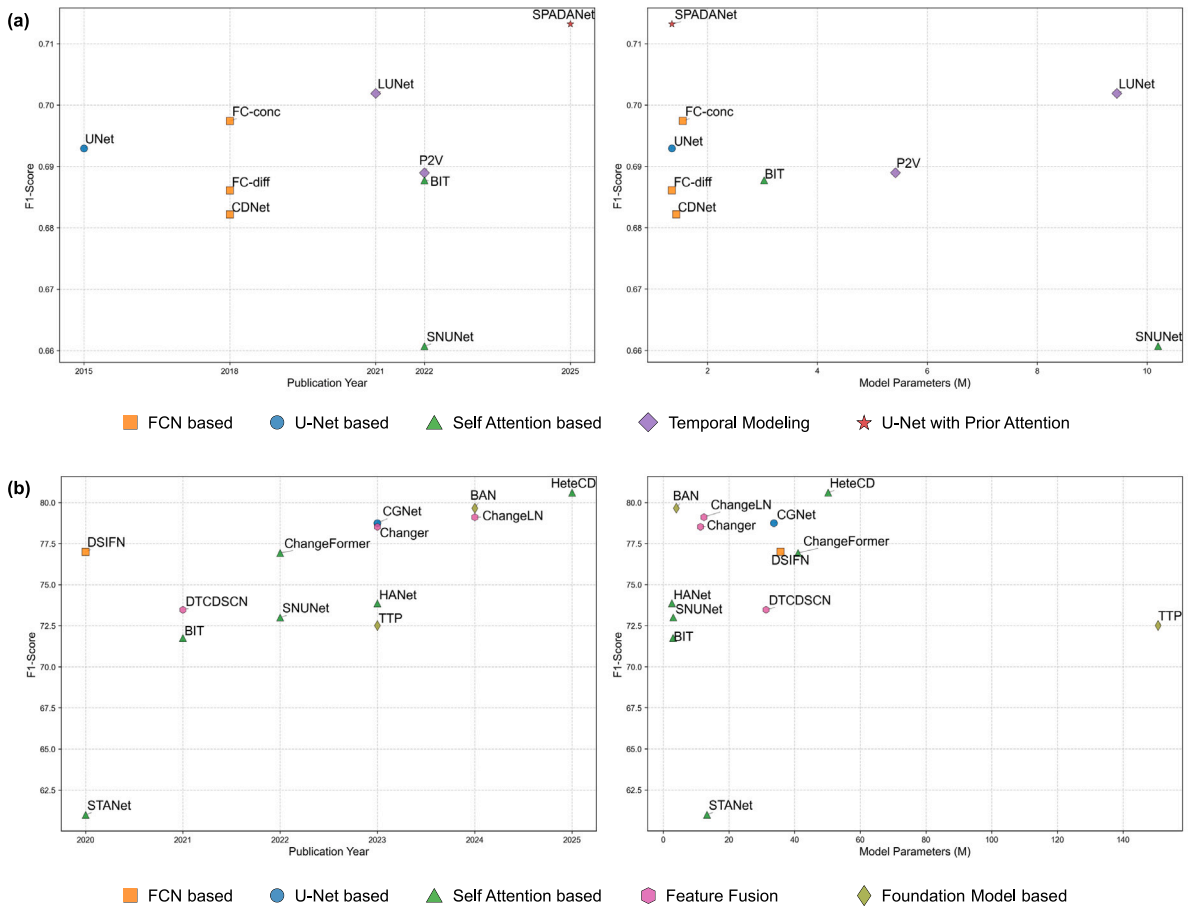


Fig. 11. A comparative analysis of performance, efficiency, and architectural trends. (a) The benchmark results: performance vs. publication year (left) and vs. model parameters (right). (b) Recent DLCD landscape.
Source: Adapted from [48].

class (low diagonal value), its misclassifications are more concentrated in adjacent categories (e.g., 'destroyed' often predicted as 'major damage'). From a disaster response perspective, misclassifying 'destroyed' as 'major' is significantly less detrimental than misclassifying it as 'no damage', as the building will still likely be flagged for inspection. This shift in error patterns underscores SPADANet's alignment with the recall-oriented needs of DA, prioritizing the identification of any damage over precise categorization, especially compared to precision-focused CD methods.

This study primarily serves as a benchmark demonstration that explicitly designing for Flood-BDA priors – in this case, using spatial inhibition prior attention to capture subtle changes – offers a distinct and promising direction compared to directly transferring self-attention-based CD models. It strongly suggests that attention mechanisms, specifically those incorporating task-relevant priors, warrant significant further exploration and development within the DA domain.

6.4. Limitation and future work

While this study provides valuable benchmarks and insights into applying prior attention mechanism and SSL for Flood-BDA, several limitations should be acknowledged, as they point towards important directions for future research.

SSL and Model Generalizability: The exploration of image-level consistency regularization demonstrated its potential but was not exhaustive. The SSL hyperparameters (e.g., buffer size, loss weights) were set based on preliminary analysis, and a more comprehensive optimization could yield further improvements. Similarly, while SPADANet's performance on the 'destroyed' class indicates current limitations, these also highlight the need for future work on more robust feature extraction and generalization across diverse disaster scenarios and types.

The Precision–Recall Trade-off and the Future of SPADANet: A key limitation of SPADANet in its current form is its lower precision compared to CD models. This trade-off, while justified by the operational needs of Flood-BDA, prevents it from being presented as the definitive SOTA solution in terms of balanced metrics. This research addresses this from three perspectives:

(1) The primary objective was not to compete with the absolute SOTA, but to establish a benchmark of foundational and classic neural network architectures to understand their inherent suitability for the Flood-BDA task. As shown in the survey of recent literature [48] (Fig. 11(b)), research post-2022 has increasingly shifted towards complex models leveraging Feature Fusion and large Foundation Models, often driven by scaling laws. These approaches, while powerful, rely heavily on extensive pre-training and make it difficult to isolate the contribution of core architectural principles. By intentionally excluding these models from our direct comparison, this paper was able to focus the analysis on the fundamental adaptability of modules like prior attention, which was the central question of this research.
(2) The performance trajectory of the DLCD field (Fig. 11) shows that recent gains are incremental. For instance, over a five-year span from 2020 to 2025, the performance improvement from a strong baseline to the SOTA HeteCD model is approximately 3% in F1-score. SPADANet's performance gain of nearly 2% over the next-best model within its architectural class is therefore highly competitive and consistent with this broader trend of diminishing returns. The massive increase in parameters seen in recent models does not yield a proportional leap in performance.
(3) The promising aspect of SPADANet's current trade-off is its potential for future development. As shown in Fig. 11(a) (right), it achieves its high performance with one of the lowest parameter counts (1.35M). This exceptional efficiency means that SPADANet is far from its performance ceiling. Future work can strategically enhance its architecture.

Broader Research Directions: Addressing these limitations will be crucial for developing more robust and operationally effective deep learning solutions. Future work should not only focus on improving precision in models like SPADANet but also explore novel metrics that better capture the unique balance of recall and error severity required for disaster response.

7. Conclusion

This study addressed the critical need for tailored DL approaches for Flood-BDA, distinct from methods directly transferred from CD. Through comprehensive benchmarking experiments, this research evaluated the performance of representative CD techniques and modules when applied to Flood-BDA, revealing limitations related to recall prioritization, error patterns, and handling subtle visual changes characteristic of flood damage.

Based on these benchmark findings, this research investigated and validated two key strategies specifically relevant to DA's challenges:

Image-Level Consistency Regularization: To combat data scarcity and label imbalance, this paper benchmarked an image-level consistency regularization approach, demonstrating its effectiveness in leveraging unlabeled data. Crucially, this paper found that reference distributions derived from pseudo-labels generally yielded better performance than those based on limited ground-truth information, providing a practical pathway for SSL in DA.

Prior Attention (SPADANet): To address the challenge of subtle change detection, this research proposed SPADANet, integrating a parameter-free prior attention module into a U-Net architecture. SPADANet significantly improved recall and exhibited more operationally relevant error patterns compared to benchmarked CD models, validating the potential of incorporating DA-specific priors into attention mechanisms.

In conclusion, this work establishes important benchmarks for both supervised and semi-supervised learning in the Flood-BDA context. It provides quantitative and qualitative evidence that DA necessitates distinct design considerations compared to general CD, particularly regarding recall emphasis and the handling of subtle features and limited labeled data. The success of SPADANet and the image-level SSL framework demonstrates the promise of prior attention and image-level consistency regularization as valuable directions for future DA research. While limitations remain, such as the need for further SSL optimization and the fact that SPADANet currently demonstrates primarily its operational adaptability rather than definitive SOTA performance, this study provides a foundational benchmark and validated methodologies for developing more effective DL solutions tailored to the unique demands of Flood-BDA.

CRediT authorship contribution statement

Jiaxi Yu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tomohiro Fukuda:** Writing – review & editing, Supervision, Resources. **Nobuyoshi Yabuki:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgment

During the preparation of this work the authors used Deepseek to improve readability and language of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Code will be availability at https://github.com/JX-OctoNeko/Flood_BDA_benchmark.git.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijdrr.2025.105664.

Data availability

Data will be made available on request.

References

[1] IPCC, Cities, settlements and key infrastructure, in: Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution To the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, 2023, pp. 907–1040.

[2] P. Ge, H. Gokon, K. Meguro, A review on synthetic aperture radar-based building damage assessment in disasters, Remote Sens. Environ. 240 (2020) 111693, http://dx.doi.org/10.1016/j.rse.2020.111693.

[3] M. Leidig, R. Teeuw, Free software: A review, in the context of disaster management, Int. J. Appl. Earth Obs. Geoinf. 42 (2015) 49–56, http://dx.doi.org/10.1016/j.jag.2015.05.012, URL: https://www.sciencedirect.com/science/article/pii/S030324341500121X.

[4] D. Peng, X. Liu, Y. Zhang, H. Guan, Y. Li, L. Bruzzone, Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges, Int. J. Appl. Earth Obs. Geoinf. 136 (2025) 104282, http://dx.doi.org/10.1016/j.jag.2024.104282, URL: https://www.sciencedirect.com/science/article/pii/S1569843224006381.

[5] Z. Zheng, Y. Zhong, J. Wang, A. Ma, L. Zhang, Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters, Remote Sens. Environ. 265 (2021) 112636, http://dx.doi.org/10.1016/j.rse.2021.112636.

[6] C. Wu, F. Zhang, J. Xia, Y. Xu, G. Li, J. Xie, Z. Du, R. Liu, Building damage detection using U-net with attention mechanism from pre- and post-disaster remote sensing datasets, Remote. Sens. 13 (5) (2021) URL: https://www.mdpi.com/2072-4292/13/5/905.

[7] D. Kim, J. Won, E. Lee, K.R. Park, J. Kim, S. Park, H. Yang, M. Cha, Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages, Front. Env. Sci. 10 (2022) http://dx.doi.org/10.3389/fenvs.2022.969758.

[8] L.P. Cui, X.P. Wang, A.X. Dou, X. Ding, High resolution SAR imaging employing geometric features for extracting seismic damage of buildings, Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-3 (2018) 239–244, http://dx.doi.org/10.5194/isprs-archives-XLII-3-239-2018.

[9] Y. Ito, M. Hosokawa, Damage estimation model using temporal coherence ratio, IGARSS, in: Proc. IEEE Int. Geosci. Remote Sens. Symp., vol. 5, 2002, pp. 2859–2861 vol.5, http://dx.doi.org/10.1109/IGARSS.2002.1026802.

[10] Y. Yamaguchi, Disaster monitoring by fully polarimetric SAR data acquired with ALOS-PALSAR, Proc. IEEE 100 (10) (2012) 2851–2860, http://dx.doi.org/10.1109/JPROC.2012.2195469.

[11] Z.L. Qiqi Zhu, D. Li, A review of multi-class change detection for satellite remote sensing imagery, Geo- Spat. Inf. Sci. 27 (1) (2024) 1–15, http://dx.doi.org/10.1080/10095020.2022.2128902.

[12] H. Jiang, M. Peng, Y. Zhong, H. Xie, Z. Hao, J. Lin, X. Ma, X. Hu, A survey on deep learning-based change detection from high-resolution remote sensing images, Remote. Sens. 14 (7) (2022) URL: https://www.mdpi.com/2072-4292/14/7/1552.

[13] X.X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, F. Fraundorfer, Deep learning in remote sensing: A comprehensive review and list of resources, 5 (4) (2017) 8–36, http://dx.doi.org/10.1109/MGRS.2017.2762307.

[14] L. Dong, J. Shan, A comprehensive review of earthquake-induced building damage detection with remote sensing techniques, ISPRS J. Photogramm. Remote Sens. 84 (2013) 85–99, http://dx.doi.org/10.1016/j.isprsjprs.2013.06.011.

[15] M.T. Marvi, A review of flood damage analysis for a building structure and contents, Nat. Hazards 102 (3) (2020) 967–995.

[16] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, L. Zhang, A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16, http://dx.doi.org/10.1109/TGRS.2021.3085870.

[17] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote. Sens. 12 (10) (2020) URL: https://www.mdpi.com/2072-4292/12/10/1662.

[18] M.A. Lebedev, Y.V. Vizilter, O.V. Vygolov, V.A. Knyaz, A.Y. Rubis, Change detection in remote sensing images using conditional adversarial networks, Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLII-2 (2018) 565–571, http://dx.doi.org/10.5194/isprs-archives-XLII-2-565-2018.

[19] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284, http://dx.doi.org/10.1109/TKDE.2008.239.

[20] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, M. Gaston, Creating xBD: A dataset for assessing building damage from satellite imagery, in: Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshop, 2019.

[21] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, G. Liu, A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images, ISPRS J. Photogramm. Remote Sens. 166 (2020) 183–200, http://dx.doi.org/10.1016/j.isprsjprs.2020.06.003.

[22] Y. Ouali, C. Hudelot, M. Tami, An overview of deep semi-supervised learning, 2020, arXiv:2006.05278 URL: https://arxiv.org/abs/2006.05278.

[23] L. Yang, L. Qi, L. Feng, W. Zhang, Y. Shi, Revisiting weak-to-strong consistency in semi-supervised semantic segmentation, in: Proc. Conf. Comput. Vis. Pattern Recognit., CVPR, 2023, pp. 7236–7246, http://dx.doi.org/10.1109/CVPR52729.2023.00699.

[24] H. Chen, Z. Qi, Z. Shi, Remote sensing image change detection with transformers, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14, http://dx.doi.org/10.1109/TGRS.2021.3095166.

[25] L. Yang, R.-Y. Zhang, L. Li, X. Xie, SimAM: A simple, parameter-free attention module for convolutional neural networks, in: M. Meila, T. Zhang (Eds.), Proc. 38th Int. Conf. Mach. Learn., ICML, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 11863–11874, URL: https://proceedings.mlr.press/v139/yang21o.html.

[26] M. Lin, G. Yang, H. Zhang, Transition is a process: Pair-to-video change detection networks for very high resolution remote sensing images, IEEE Trans. Image Process. 32 (2023) 57–71, http://dx.doi.org/10.1109/TIP.2022.3226418.

[27] S. Fang, K. Li, Z. Li, Changer: Feature interaction is what you need for change detection, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11, http://dx.doi.org/10.1109/TGRS.2023.3277496.

[28] S. Xie, J. Duan, S. Liu, Q. Dai, W. Liu, Y. Ma, R. Guo, C. Ma, Crowdsourcing rapid assessment of collapsed buildings early after the earthquake based on aerial remote sensing image: A case study of yushu earthquake, Remote. Sens. 8 (9) (2016) http://dx.doi.org/10.3390/rs8090759.

[29] P.F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, R. Gherardi, Street-view change detection with deconvolutional networks, Auton. Robot. 42 (2018) 1301–1322.

[30] R. Caye Daudt, B. Le Saux, A. Boulch, Fully convolutional siamese networks for change detection, in: Proc. 25th IEEE Int. Conf. Image Process., ICIP, 2018, pp. 4063–4067, http://dx.doi.org/10.1109/ICIP.2018.8451652.

[31] M. Papadomanolaki, M. Vakalopoulou, K. Karantzalos, A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection, IEEE Trans. Geosci. Remote Sens. 59 (9) (2021) 7651–7668, http://dx.doi.org/10.1109/TGRS.2021.3055584.

[32] S. Fang, K. Li, J. Shao, Z. Li, SNUNet-CD: A densely connected siamese network for change detection of VHR images, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5, http://dx.doi.org/10.1109/LGRS.2021.3056416.

[33] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Interv, Springer International Publishing, Cham, 2015, pp. 234–241.
[34] Z. Xing, S. Yang, X. Zan, X. Dong, Y. Yao, Z. Liu, X. Zhang, Flood vulnerability assessment of urban buildings based on integrating high-resolution remote sensing and street view images, Sust. Cities Soc. 92 (2023) 104467, http://dx.doi.org/10.1016/j.scs.2023.104467.
[35] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.
[36] D. Roy, R. Pramanik, R. Sarkar, Margin-aware adaptive-weighted-loss for deep learning based imbalanced data classification, IEEE Trans. Artif. Intell. 5 (2) (2024) 776–785, http://dx.doi.org/10.1109/TAI.2023.3275133.
[37] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, X. Huang, SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images, IEEE Trans. Geosci. Remote Sens. 59 (7) (2021) 5891–5906, http://dx.doi.org/10.1109/TGRS.2020.3011913.
[38] H. Chen, J. Song, C. Wu, B. Du, N. Yokoya, Exchange means change: An unsupervised single-temporal change detection framework based on intra- and inter-image patch exchange, ISPRS J. Photogramm. Remote Sens. 206 (2023) 87–105, http://dx.doi.org/10.1016/j.isprsjprs.2023.11.004.
[39] O. Chapelle, B. Scholkopf, A. Zien (Eds.), Semi-supervised learning (Chapelle, O. et al. eds.; 2006) [Book reviews], IEEE Trans. Neural Netw. 20 (3) (2009) 542, http://dx.doi.org/10.1109/TNN.2009.2015974.
[40] D. Berthelot, N. Carlini, E.D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, C. Raffel, Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring, 2019, arXiv preprint arXiv:1911.09785.
[41] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C.A. Raffel, E.D. Cubuk, A. Kurakin, C.-L. Li, FixMatch: Simplifying semi-supervised learning with consistency and confidence, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), in: Adv. Neural Inf. Proces. Syst., vol. 33, Curran Associates, Inc., 2020, pp. 596–608, URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/06964dce9addb1c5cb5d6e3d9838f733-Paper.pdf.
[42] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, IAUnet: Global context-aware feature learning for person reidentification, IEEE Trans. Neural Netw. Learn. Syst. 32 (10) (2021) 4460–4474, http://dx.doi.org/10.1109/TNNLS.2020.3017939.
[43] Z. Zhang, C. Lan, W. Zeng, X. Jin, Z. Chen, Relation-aware global attention for person re-identification, in: Proc. Conf. Comput. Vis. Pattern Recognit., CVPR, 2020, pp. 3183–3192, http://dx.doi.org/10.1109/CVPR42600.2020.00325.
[44] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Proces. Syst. 33 (2020) 6840–6851.
[45] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proc. Conf. Comput. Vis. Pattern Recognit., CVPR, 2017, pp. 1125–1134.
[46] Y. Grandvalet, Y. Bengio, Semi-supervised learning by entropy minimization, in: L. Saul, Y. Weiss, L. Bottou (Eds.), in: Adv. Neural Inf. Proces. Syst., vol. 17, MIT Press, 2004, URL: https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf.
[47] A. Achille, M. Rovere, S. Soatto, Critical learning periods in deep networks, in: Int. Conf. Learn. Represent., ICLR, 2019, URL: https://openreview.net/forum?id=BkeStsCcKQ.
[48] W. Jing, H. Bai, B. Song, W. Ni, J. Wu, Q. Wang, HeteCD: Feature consistency alignment and difference mining for heterogeneous remote sensing image change detection, ISPRS J. Photogramm. Remote Sens. 223 (2025) 317–327, http://dx.doi.org/10.1016/j.isprsjprs.2025.03.008, URL: https://www.sciencedirect.com/science/article/pii/S0924271625001066.