remote sensing

MDPI

Article

Insta-BDA: Instance-Aware Building Damage Assessment and Counting via Foundation Model Fusion

Beyza Gürer [1], Shaaban Sahmoud [2,*], Mohammed Bennamoun [3], Farid Boussaid [4] and Ali Kishk [5]

[1] Independent Researcher, Istanbul 34000, Türkiye; ebeyzagurer@gmail.com
[2] Department of Computer Engineering, Fatih Sultan Mehmet Vakif University, Istanbul 34000, Türkiye
[3] Department of Computer Science and Software Engineering, The University of Western Australia, Perth 6009, Australia; mohammed.bennamoun@uwa.edu.au
[4] Electrical, Electronic and Computer Engineering, The University of Western Australia, Perth 6009, Australia; farid.boussaid@uwa.edu.au
[5] Investigative Department, Aljazeera Media Network, Doha P.O. Box 23123, Qatar; kishka@aljazeera.net
* Correspondence: ssahmoud@fsm.edu.tr

Highlights

What are the main findings?

- A fusion framework (Insta-BDA) integrating ChangeMamba change detection with SAM3 zero-shot instance segmentation enables building-level damage assessment using only pixel-level supervision, reducing aggregate building count deviation from −47.06% to −34.96%.
- A lightweight learnable fusion variant (MLP with ∼2400 parameters) improves damage classification F1 from 0.670 to 0.696, with cross-dataset evaluation on IAN-BD and IDA-BD demonstrating improved generalization.

What are the implications of the main findings?

- Foundation model fusion provides a practical pathway toward operational, instance-level building damage assessment for disaster response without requiring costly instance-level annotations.
- Per-building damage counts and classifications offer directly actionable outputs for emergency resource allocation and recovery planning.

Abstract

Accurate building damage assessment from satellite imagery is essential for post-disaster response and recovery. Most deep learning approaches treat this task as semantic segmentation, producing pixel-level damage maps without separating individual buildings. This formulation inherently limits reliable per-building damage quantification, which is often far more informative for operational decision-making. We present Insta-BDA, an instance-aware framework that integrates change detection with foundation-model-based instance segmentation for building-level damage assessment using only pixel-level supervision. The approach combines ChangeMamba for spatiotemporal change detection with SAM3 for zero-shot building instance extraction, reconciled through a confidence-guided fusion mechanism. The framework does not require instance-level annotations (polygons, bounding boxes, or instance masks) for training; instance-level structure is derived automatically from SAM3's zero-shot detections, and the learnable fusion variant requires only pixel-level damage labels. We investigate two fusion strategies—a rule-based approach (Insta-BDA-RB) and a learnable variant (Insta-BDA-MLP)—using a compact multilayer perceptron on per-instance features. To improve robustness under typical satellite resolutions and annotation variability, we adopt a binary damage formulation. Experiments

Academic Editors: Massimiliano Pepe and Wen Liu

Received: 16 April 2026
Revised: 5 July 2026
Accepted: 8 July 2026
Published: 14 July 2026

on xBD show that Insta-BDA reduces the aggregate building count deviation to −34.96%, compared with −47.06% for ChangeMamba, while maintaining competitive damage classification performance. The learnable fusion further improves damage F1 (0.70 vs. 0.67 for rule-based fusion). Cross-dataset evaluation on IAN-BD and IDA-BD indicates improved generalization. These results suggest that integrating foundation model segmentation with change detection offers a practical pathway toward operational, instance-level building damage assessment.

Keywords: building damage assessment; change detection; foundation models; instance segmentation; disaster response

1. Introduction

Timely and reliable building damage assessment after natural and human-made disasters is critical for emergency response, resource prioritization, and long-term recovery planning [1]. Remote sensing data, including satellite and aerial imagery acquired before and after disaster events, enable large-scale and geographically extensive damage analysis [2]. In recent years, deep learning has substantially advanced automated damage assessment, with approaches spanning convolutional neural networks to more recent Transformer-based architectures.

Despite these advances, most existing methods formulate building damage assessment as a semantic segmentation task [3], assigning a damage label to each pixel. Some two-stage approaches, such as the Siamese attention U-Net proposed by Wu et al. [4], use a building localization branch to constrain damage classification to building regions. While this improves classification quality by masking non-building pixels, the output remains a pixel-level damage map: individual buildings are not assigned unique identifiers, and per-building counts or labels cannot be derived without additional post-processing such as connected component analysis. Thus, even localization-aware segmentation methods do not resolve this limitation: both formulations support end-to-end optimization and achieve strong pixel-level evaluation metrics, but they inherently discard instance-level structure. In particular, semantic segmentation does not directly provide the number of damaged buildings [5], which is often a more operationally meaningful statistic than aggregate pixel-wise damage measurements.

The absence of instance-level reasoning becomes especially problematic in semi-dense urban environments where buildings are spatially adjacent. Although connected component analysis can be applied to semantic masks to approximate building counts [6], this strategy frequently fails when neighboring buildings are merged into a single contiguous region. Such merging is common in semi-dense areas. Additional failure modes include complex building geometries, partial occlusions, mixed damage predictions within a single structure, and noisy outputs that fragment individual buildings into multiple components.

Mixed damage predictions within individual buildings also complicate evaluation. Ground-truth annotations are typically provided at the polygon level, assigning a single damage category per building. In contrast, semantic segmentation models may produce heterogeneous pixel labels within a single structure. These fragmented predictions can inflate pixel-level F1 scores by partially overlapping with the ground truth, even when the overall building-level classification is incorrect. Instance-aware approaches that enforce one label per building yield cleaner and more interpretable outputs, but they incur an "all-or-nothing" penalty in evaluation when a building is misclassified.

One possible solution is to train dedicated instance segmentation models, such as Mask R-CNN [7] or Faster R-CNN [8], to explicitly detect and segment individual buildings. However, instance segmentation is intrinsically more challenging than semantic segmentation because it requires jointly solving object detection and mask prediction to separate individual instances [9]. This paradigm depends on bounding-box and mask annotations for each object, must resolve overlapping instances, and is evaluated under strict instance-level matching criteria rather than per-pixel accuracy. The collection of detailed instance-level annotations is substantially more costly than pixel-wise labeling, constraining dataset scale and diversity.

Recent foundation models, particularly the Segment Anything Model (SAM) [10] family, provide an alternative pathway. These models can generate high-quality instance masks in a zero-shot manner without task-specific training. Nevertheless, SAM alone does not perform damage assessment: while it can delineate building instances, it lacks the capacity to classify their damage state without additional supervision.

In this work, instead of introducing a new backbone architecture, we propose Insta-BDA, a framework centered on principled fusion. The key idea is to integrate pixel-level change detection with instance-level segmentation in a unified pipeline. Conventional post-processing schemes that apply connected component analysis to semantic masks degrade in dense urban scenes where adjacent buildings coalesce. In contrast, Insta-BDA performs fusion directly at the instance level. It executes confidence-guided decisions per building rather than per pixel, reconciles discrepancies in spatial coverage between change detection outputs and instance masks, and incorporates damage-aware thresholds to better capture partial structural damage.

Building on this, we advance building damage assessment beyond pixel-level reasoning by formalizing it as an instance-level decision-making problem. This formulation aligns directly with operational needs in disaster response, where actionable outputs—such as the number, location, and status of individual structures—are critical for resource allocation and recovery planning. Our approach integrates the spatial priors of foundation models for building delineation with temporally grounded change signals derived from bi-temporal imagery. Through this fusion, the framework reconciles structural boundaries with dynamic damage evidence, enabling robust, interpretable, and practically deployable per-building damage assessment.

Our contributions are summarized as follows:

1. Instance-Level Decision Framework: We reformulate building damage assessment as an instance-level decision-making problem, enabling explicit per-building damage classification and counting. This formulation aligns with operational disaster response requirements, where actionable outputs at the building level are critical for resource allocation and recovery planning.
2. Confidence-Guided Fusion: We propose a principled fusion strategy that integrates pixel-level change detection with instance-level segmentation from foundation models. The framework combines spatiotemporal change signals with strong spatial priors, enabling consistent and interpretable per-building predictions.
3. Improved Building Counting and Generalization: By leveraging foundation-model-derived instances, our approach reduces aggregate building count deviation and demonstrates enhanced generalization across datasets, while maintaining competitive damage classification performance.

The remainder of this paper is structured as follows: Section 1.1 reviews related work. Section 2 presents the materials and methods, including the proposed fusion framework, datasets, and implementation details. Section 3 reports experimental results on three

benchmark datasets with detailed quantitative and qualitative analyses. Section 4 discusses the findings and limitations. Section 5 concludes the paper.

1.1. Related Work

This section surveys the primary research areas relevant to the proposed framework: deep learning-based building damage assessment from satellite imagery, state-space models for change detection, and foundation models for building extraction.

1.1.1. Deep Learning for Building Damage Assessment

Automated assessment of building damage from bi-temporal satellite imagery is commonly structured as a two-stage process: building localization from pre-disaster imagery, followed by damage classification using paired pre- and post-event images [3]. The winning solution to the xView2 challenge [11] formalized this paradigm using U-Net ensembles equipped with high-capacity encoders. BDANet [12] incorporated multiscale convolutional neural networks with cross-directional attention mechanisms, although CNN-based architectures remain constrained by limited receptive fields. The Microsoft Siamese CNN [13] established a practical end-to-end baseline through joint detection and classification objectives. ChangeOS [14] extended the scope of damage assessment to include man-made disasters using object-based semantic change detection. DDNet [15] further introduced dual-temporal joint attention modules to model correlations between pre- and post-disaster imagery while reducing sensitivity to variations in acquisition conditions.

Transformer-based approaches have been proposed to overcome the receptive field limitations of convolutional networks. SDAFormer [16] integrates Swin Transformer blocks within a Siamese U-Net framework, HRTBDA [17] combines HRNet backbones with Transformer-based feature fusion, and DAHiTrA [18] employs hierarchical Transformer architectures with difference blocks to capture multiscale damage features. DamageCAT [19] introduces typology-based damage categorization using a hierarchical U-Net Transformer, replacing ordinal severity levels with categorical damage types. Although Transformer-based models achieve strong predictive performance, their quadratic computational complexity can limit scalability to large-area, high-resolution imagery. Across both CNN and Transformer paradigms, severe class imbalance among damage categories remains a persistent challenge [20]. More recently, ECADS-CNN [21] combines efficient channel attention with depthwise separable convolutions for four-class damage classification on xBD (xView2 score = 0.758), and ADSFNet [22] fuses adaptive attention with state-space modeling to achieve an xView2 score of 0.740 on xBD using optical imagery alone.

1.1.2. State Space Models for Change Detection

Structured state spaces efficiently model long-range dependencies in sequential data [23], making them a scalable alternative to Transformers for bi-temporal change detection.

The Mamba architecture [24], derived from structured state-space models, enables sequence modeling with linear computational complexity while retaining modeling capacity comparable to that of Transformers. Extensions such as VMamba [25] and Vision Mamba [26] adapt this formulation to visual representation learning, achieving competitive performance with improved efficiency.

ChangeMamba [27] represents the first application of Mamba-based architectures to remote sensing change detection, introducing MambaBCD, MambaSCD, and MambaBDA for binary change detection, semantic change detection, and building damage assessment, respectively. This framework achieves state-of-the-art results on the xBD benchmark while reducing computational cost. Subsequent variants include CDMamba [28], which enhances local detail modeling; SpectMamba [29], which integrates frequency-domain representations; and DC-Mamba [30], which targets challenging boundary conditions. M-CD [31]

further demonstrates the effectiveness of Mamba-based Siamese networks for change detection, achieving improvements over Transformer baselines with linear-time training. Despite these advances, these approaches operate within a semantic segmentation paradigm and do not explicitly model instance-level building separation, which can limit robustness in densely built or heavily damaged urban environments.

1.1.3. Foundation Models for Building Extraction

The Segment Anything Model (SAM) [10] introduced large-scale, prompt-driven segmentation trained on over one billion masks, enabling high-quality zero-shot mask generation. SAM2 [32] extends this capability to video, while SAM3 [33] incorporates open-vocabulary segmentation through text prompts. In the remote sensing domain, Ren et al. [34] analyzed SAM's performance on overhead imagery, identifying domain shift and scale sensitivity as key limitations. This has motivated adaptations such as RSAM-Seg [35] and LiDAR-assisted approaches [36] tailored to aerial and satellite imagery.

However, SAM-family models are designed for spatial segmentation and do not model temporal changes. A recent exception is the YOLO-E + SAM2 pipeline [37], which uses object detection to prompt SAM2 for pixel-level building damage extraction; however, this approach requires a supervised detector trained on building damage data and does not perform bi-temporal change reasoning. Consequently, SAM-family models have not been directly applied to building damage assessment tasks that require reasoning over pre- and post-disaster imagery.

By integrating SAM3-based instance extraction with MambaBDA's semantic damage prediction, the proposed framework combines zero-shot instance segmentation with efficient spatiotemporal change modeling. This two-stage design facilitates the identification of individual buildings even in scenarios where severe destruction obscures structural boundaries. Comprehensive evaluation on xBD, IDA-BD, and IAN-BD, alongside comparisons with ChangeOS [14] and the Microsoft Siamese CNN [13], enables assessment across diverse disaster types and imaging conditions.

2. Materials and Methods

Figure 1 illustrates the overall architecture of the Insta-BDA framework. The system processes paired pre- and post-disaster imagery using two branches: SAM3 for instance-level building extraction and ChangeMamba for pixel-level change detection. The outputs of these branches are subsequently integrated using a confidence-guided fusion strategy.

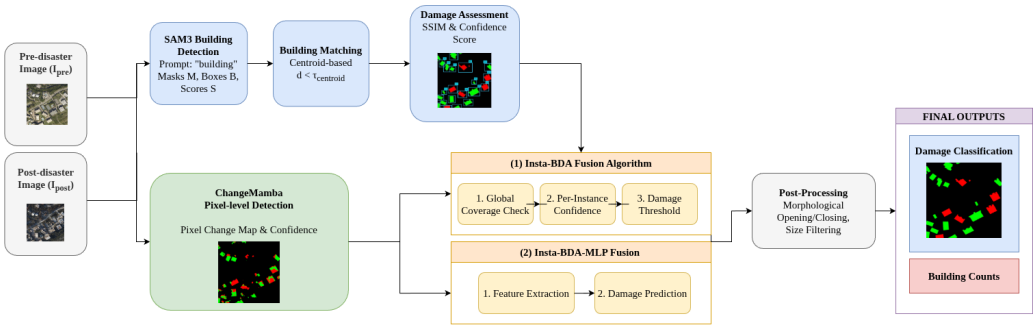

Figure 1. Overview of the Insta-BDA architecture. Pre- and post-disaster imagery is processed in parallel, with SAM3 performing instance-level building detection and ChangeMamba conducting pixel-level change analysis. The resulting predictions are integrated through a fusion strategy that leverages global coverage consistency and instance-wise confidence-guided fusion. A final post-processing stage refines the outputs to produce consolidated damage classifications and accurate building counts.

2.1. Problem Formulation

Given a pre-disaster image I_{pre} and a corresponding post-disaster image I_{post}, the objective is twofold:

1. Assign a damage label to each detected building instance.
2. Accurately estimate the total number of buildings, as well as the counts of non-damaged and damaged structures.

The original four damage categories are consolidated into a binary formulation:

- Non-Damaged: No damage and minor damage.
- Damaged: Major damage and destroyed.

This simplification is motivated by two considerations: First, distinguishing fine-grained damage levels (e.g., minor versus major damage) is challenging due to subtle and context-dependent visual cues in satellite imagery. Empirically, minor damage is the weakest class for all evaluated models: even the best-performing model (ChangeMamba) achieves only 0.576 F1 on the minor-damage class, compared to 0.945 for no-damage and 0.871 for destroyed (detailed in Section 2.7). This performance gap widens substantially on out-of-distribution datasets (e.g., 0.509 Minor F1 on IAN-BD). Additionally, the xBD test set exhibits severe class imbalance among damage subcategories: only 11.0% of buildings are minor damage, 8.8% major damage, and 8.3% destroyed, compounding the classification difficulty. Second, from an operational perspective, the binary distinction between intact and severely affected buildings provides sufficient granularity for emergency response and resource planning.

We emphasize that the binary formulation is a design choice for the damage classification head, not a limitation of the fusion framework itself. The core contribution of Insta-BDA is instance-level reasoning: producing per-building damage labels and counts by fusing change detection with foundation model segmentation. This capability is absent from purely pixel-level methods regardless of the number of damage classes. The fusion mechanism (Section 2) operates on per-instance features and is agnostic to the number of output categories; extending to four-class damage assessment requires recalibration of the fusion thresholds, treatment of class imbalance, and evaluation on datasets with more balanced damage distributions. We consider this to be an important direction for future work, noting that the binary formulation already demonstrates the effectiveness of the instance-level fusion strategy: Insta-BDA reduces the aggregate building count deviation from -47.06% to -34.96% while improving cross-dataset generalization (detailed in Section 3).

2.2. SAM3 Damage Assessment Pipeline

The SAM3 branch performs instance-level damage estimation through three stages: detection, matching, and structural comparison.

2.2.1. Building Detection

SAM3 is applied to both I_{pre} and I_{post} using the text prompt "building". For each detected structure, the model produces an instance mask M, bounding box B, and confidence score S.

2.2.2. Building Matching

Detected instances in pre- and post-disaster images are associated using centroid-based matching:

$$d(m_i^{pre}, m_j^{post}) = \|c(m_i^{pre}) - c(m_j^{post})\|_2, \quad (1)$$

where $c(\cdot)$ denotes the centroid of a mask. A match is accepted if $d < \tau_{centroid}$, with $\tau_{centroid} = 30$ pixels by default.

Pre-disaster buildings that cannot be matched to any post-disaster detection—for example, due to complete structural collapse—are classified as damaged by default, since the absence of a building in the post-disaster image is itself evidence of destruction. Post-disaster detections without a pre-disaster match are classified as non-damaged, as they likely represent structures that were not detected pre-disaster rather than newly constructed buildings.

2.2.3. Damage Assessment

For each matched pair, structural change is quantified using the Structural Similarity Index (SSIM):

$$\text{damaged}_i = \mathbf{1}[\text{SSIM}(R_i^{pre}, R_i^{post}) < \tau_{ssim}], \tag{2}$$

where R_i denotes the masked building region and $\tau_{ssim} = 0.6$.

The final confidence score integrates detection confidence and SSIM-derived damage confidence:

$$\text{conf}_i = \frac{S_i^{pre} + \sigma(\tau_{ssim} - \text{SSIM}_i)}{2}, \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. Direct subtraction $(\tau_{ssim} - \text{SSIM}_i)$ would yield unbounded and asymmetric values. The sigmoid transformation constrains the result to $(0, 1)$, ensuring symmetric behavior around the threshold.

2.3. Insta-BDA Fusion Algorithm

The principal innovation of Insta-BDA lies in a confidence-guided fusion of pixel-level predictions and instance-level segmentation. The fusion proceeds in three stages: (1) a global coverage check determines whether ChangeMamba predictions or an empty mask initializes the fused output, (2) a per-instance confidence decision selects the more reliable source for each building, and (3) a damage ratio threshold assigns the final building-level label when ChangeMamba is selected. A morphological post-processing step follows to remove artifacts.

2.3.1. Global Coverage Comparison

We first compare the spatial coverage of predictions from ChangeMamba and SAM3. Let p_{cm} denote the number of non-zero pixels in the ChangeMamba mask and p_{sam} the total pixels covered by SAM3 instances. The ratio

$$\rho = \frac{p_{cm}}{\max(p_{sam}, p_{cm}, 1)} \tag{4}$$

quantifies relative coverage. If $\rho > 0.5$, the fused mask is initialized using ChangeMamba predictions to avoid discarding relevant detections when change detection is more comprehensive. Otherwise, initialization begins with an empty mask. The threshold of 0.5 represents a natural decision boundary: when ChangeMamba covers more than half the spatial extent of either prediction source, it provides sufficient spatial context to serve as a reliable initialization. Below this threshold, ChangeMamba predictions are sparse relative to SAM3 detections, and initializing from them would introduce noise from isolated pixel-level predictions that lack instance-level coherence. A sensitivity analysis of this threshold is presented in Section 3.4.

2.3.2. Instance-Level Confidence Decision

For each SAM3 instance a_i, we compute the region R_i and the mean ChangeMamba confidence $\bar{c}_{cm}$ over non-background pixels within R_i. This value is compared with the SAM3 detection score, a_i.score. The model exhibiting higher confidence determines the building-level classification.

The ChangeMamba confidence $\bar{c}_{cm}$ is derived from the softmax output of the semantic segmentation head, representing the predicted class probability at each pixel. For a given SAM3 instance region R_i, we average these probabilities over all non-background pixels, yielding an estimate of how confidently ChangeMamba assigns damage labels within that building footprint. The SAM3 detection score a_i.score corresponds to the model's predicted objectness probability for the detected instance, reflecting SAM3's confidence that the region corresponds to a valid building. These two scores originate from different model architectures and training objectives; consequently, they are not inherently calibrated to the same scale. The per-instance confidence comparison is therefore a heuristic that assumes the model with higher confidence on a given building is more likely to be correct. We evaluate this assumption empirically and discuss calibration properties in Section 3.5.

2.3.3. Damage Ratio Threshold

When ChangeMamba is selected as the more reliable source, we compute the proportion of pixels classified as damaged:

$$r_{dmg} = \frac{\#\text{ damaged pixels in } R_i}{|R_i|}. \tag{5}$$

If $r_{dmg} \geq 1/3$, the entire building is labeled as damaged; otherwise, the dominant class within R_i is assigned. This threshold mitigates false negatives arising from partial damage predictions that fall below majority voting.

2.3.4. Post-Processing

After instance-wise decisions, morphological opening removes small artifacts, while closing fills minor gaps. Connected components smaller than 10% of the median building size are discarded to suppress spurious detections.

2.4. Learnable Fusion (Insta-BDA-MLP)

To complement the rule-based fusion, we introduce a learnable alternative in which per-instance decisions are predicted by a lightweight multilayer perceptron.

2.4.1. Feature Representation

For each SAM3 instance, an 8-dimensional feature vector is constructed:

1. $\bar{c}_{cm}$: Mean ChangeMamba confidence within the instance (excluding background);
2. r_{dmg}: Proportion of damaged pixels predicted by ChangeMamba;
3. r_{nd}: Proportion of non-damaged pixels;
4. r_{bg}: Proportion of background pixels;
5. s_{sam}: SAM3 detection confidence;
6. d_{sam}: SAM3 SSIM-based binary damage label;
7. ρ: Global coverage ratio;
8. $\Delta c = \bar{c}_{cm} - s_{sam}$: Confidence difference.

2.4.2. Network Architecture

The fusion network consists of two hidden layers:

$$f(\mathbf{x}) = W_3\sigma(W_2\sigma(W_1\mathbf{x} + b_1) + b_2) + b_3, \tag{6}$$

where $\mathbf{x} \in \mathbb{R}^8$, the hidden layer sizes are 64 and 32, and each hidden layer employs batch normalization, ReLU activation, and dropout ($p = 0.3$). The output is a two-class logit (non-damaged vs. damaged). The total parameter count is approximately 2400. The overall network architecture is illustrated in Figure 2.

An MLP is selected over more expressive architectures (e.g., attention-based models or graph neural networks) due to the low dimensionality of the feature space. The design achieves negligible computational overhead (less than 0.01 s loading time and 0.41 s per image). More complex models would increase the training and inference cost without proportional benefit.

2.4.3. Training Procedure

Training uses per-instance features from the xBD holdout set (31,679 building instances). Focal loss [38] with $\gamma = 2.0$ addresses class imbalance (75% non-damaged, 25% damaged). Optimization is performed using AdamW with cosine annealing. Stratified 5-fold cross-validation is applied, and features are standardized using training statistics. The classification threshold is selected on the validation fold to maximize Macro F1, resulting in an optimal value of 0.63. Cross-validation yields a Macro F1 of 0.860 ± 0.005.

During inference, the MLP replaces the rule-based confidence comparison. Rather than directly comparing $\bar{c}_{cm}$ and a_i.score or applying a fixed damage ratio heuristic, the model predicts the building-level damage label from the feature vector. The global coverage initialization and post-processing steps remain unchanged.

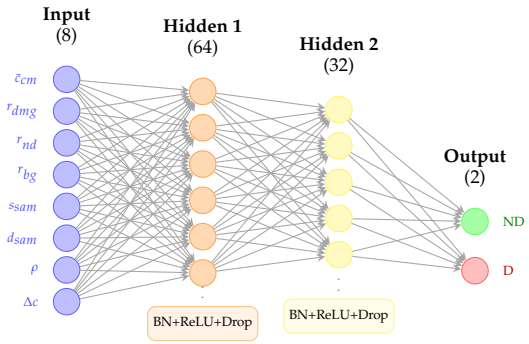

Figure 2. Architecture of the learnable fusion MLP. Eight-dimensional per-instance features extracted from ChangeMamba and SAM3 are fed into a two-layer fully connected network (64 and 32 neurons) to predict the building damage category. The model contains ∼2400 trainable parameters. ND: non-damaged; D: damaged.

2.5. Datasets

We conducted evaluation on three benchmark datasets. The xBD dataset [39] is the largest publicly available benchmark for building damage assessment, comprising more than 850,000 annotated building instances across 19 disaster events, with four damage severity levels. We used the official xBD test split (933 images) for final evaluation. Representative xBD imagery with annotated building footprints is shown in Figure 3. The xBD dataset provides three official partitions (train, holdout, and test), with no overlap among them. We used the holdout partition (933 images) with polygon-level JSON annotations for hyperparameter tuning. The holdout set was used exclusively for hyperparameter tuning

and MLP training; it played no role in any of the final reported evaluation results. Table 1 provides an explicit mapping of each experimental table to its data partition.

To evaluate cross-dataset generalization, we further tested on IDA-BD [40], which contains satellite imagery from Hurricane Ida (2021), and IAN-BD [41], which consists of aerial imagery from Hurricane Ian (2022).

Table 1. Experimental table-to-partition mapping.

Tables	Partition	Description
2, 4–8, 10, 11 and 16	xBD test (933 img)	Primary evaluation results
3	Cross-paper	Results cited from publications
9	xBD holdout (933 img)	Ablation study
12	xBD holdout (933 img)	Threshold sensitivity analysis
13	xBD holdout (5-fold CV)	Feature ablation (MLP)
14 and 17	xBD test (20 img)	Runtime and compute
15	xBD holdout (933 img)	SSIM robustness

Figure 3. Representative satellite imagery from the xBD dataset with annotated building footprints.

2.6. Implementation Details

2.6.1. Insta-BDA-MLP Training

The learnable fusion MLP was trained on per-instance features extracted from the xBD holdout set. The eight-dimensional feature vector for each SAM3 building instance was computed from ChangeMamba and SAM3 outputs as described in Section 2. We used AdamW [42] with an initial learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and cosine annealing (minimum LR 1×10^{-6}) over a maximum of 200 epochs. Features were standardized to zero mean and unit variance using training-set statistics. Training used focal loss [38] ($\gamma = 2.0$) with inverse class frequency weighting to address the class imbalance (approximately 75% non-damaged, 25% damaged instances). Stratified 5-fold cross-validation grouped by disaster type was used for model selection, with early stopping (patience = 20 epochs) based on validation Macro F1. The classification threshold was optimized per fold, yielding a final threshold of 0.63. The batch size was 512. All training was performed on a single NVIDIA GeForce RTX 3060 GPU (NVIDIA Corporation, Santa Clara, CA, USA).

2.6.2. SAM3 Configuration

We used the SAM3 model [33] loaded from the `facebook/sam3` checkpoint. Building instances were detected using the text prompt "building" applied to both pre-disaster and post-disaster images independently. SAM3 returned per-instance binary masks, bounding

boxes, and confidence scores. A minimum detection confidence threshold of 0.3 was applied to filter low-confidence detections. No additional fine-tuning or domain adaptation was performed; SAM3 operates entirely in a zero-shot setting.

2.6.3. Pre- and Post-Processing

Input images were resized to 1024 × 1024 pixels for both ChangeMamba and SAM3. For ChangeMamba, the output was a per-pixel class probability map obtained via softmax. For the fusion stage, SAM3 instance masks were rasterized to the same spatial resolution. Post-processing consisted of morphological opening followed by closing (both with a 3 × 3 kernel) to remove small artifacts and fill minor gaps. Connected components smaller than 10% of the median building size in the image were removed. The final output assigns a single damage label per building instance and aggregates instance counts for building-level statistics.

2.7. Baseline Model Selection

To determine the base change detection model, we compared three building damage assessment approaches across multiple datasets (Table 2). We report localization F1 (building detection accuracy), classification F1 (damage severity classification), overall F1, and per-class F1 scores for the four damage categories. Evaluation was performed on xBD (in-distribution), IAN-BD (aerial imagery), and IDA-BD (satellite imagery) to assess both in-domain performance and generalization. ChangeMamba achieved the strongest overall performance across most metrics and datasets—with the exception of IDA-BD, where it had the lowest classification F1 (0.026) and overall F1 (0.236) of the three baselines (Table 2)—and was therefore selected as the base change detection backbone for Insta-BDA.

Table 2. Baseline Model Comparison: Localization, classification, and per-class F1 Scores.

Model	Dataset	Loc F1	Clf F1	OA F1	No-Dmg	Minor	Major	Destroyed
ChangeOS	xBD	0.848	0.747	0.777	0.924	0.571	0.747	0.844
	IAN-BD	0.761	0.007	0.233	0.278	0.002	0.630	0.144
	IDA-BD	0.738	0.042	0.250	0.733	0.220	0.021	0.023
Microsoft	xBD	0.735	0.361	0.474	0.832	0.253	0.228	0.655
	IAN-BD	0.670	0.205	0.344	0.298	0.109	0.384	0.226
	IDA-BD	0.602	0.086	0.241	0.701	0.214	0.171	0.029
ChangeMamba	xBD	**0.857**	**0.758**	**0.787**	**0.945**	**0.576**	**0.748**	**0.871**
	IAN-BD	**0.762**	**0.377**	**0.492**	0.299	**0.509**	0.587	**0.277**
	IDA-BD	**0.726**	0.026	0.236	**0.836**	**0.361**	0.007	**0.077**

Loc F1: Localization F1. Clf F1: Classification F1. OA F1: Overall F1. Best results per dataset in **bold**.

To contextualize the performance of the selected backbone, Table 3 compares recent methods on the xBD dataset using the xView2 combined score (0.3 × Loc F1 + 0.7 × Clf F1) and per-class damage classification F1.

Table 3. Comparison with recent methods on xBD (four-class damage classification).

Method	Year	xView2 Score	No-Dmg	Minor	Major	Destroyed
BDANet [12] †	2022	0.806	0.925	0.616	0.788	0.876
DAHiTrA [18] †	2023	**0.819**	**0.978**	**0.711**	0.765	0.772
ADSFNet [22] ‡	2026	0.740	0.850	0.522	0.689	0.753
ChangeOS [14] *	2021	0.777	0.924	0.571	0.747	0.844
ChangeMamba [27] *	2024	0.787	0.945	0.576	0.748	**0.871**

xView2 Score = 0.3 × Loc F1 + 0.7 × Clf F1. † Results as reported in [18]. ‡ Results as reported in [22]. * Our evaluation. Best score in each column is shown in **bold**. Note: Cross-paper scores may differ due to evaluation protocols.

DAHiTrA achieves the highest xView2 combined score (0.819), with notably strong Minor damage classification (0.711 F1). ChangeMamba achieves the second-highest destroyed-class F1 (0.871), behind BDANet's 0.876, and competitive overall performance (0.787). We note that cross-paper comparisons should be interpreted with caution, as differences in evaluation protocols, pre-processing, and test set handling may affect the reported scores. The bottom two rows use our evaluation pipeline for consistency. ChangeMamba was selected as the backbone due to its strong performance and public availability of pretrained weights. The Insta-BDA framework is backbone-agnostic, and stronger backbones can be substituted without modifying the fusion logic.

3. Results

Table 4 reports segmentation performance on xBD (in-distribution) and cross-dataset generalization results on IAN-BD and IDA-BD (out-of-distribution). We compare four configurations: SAM3 only, ChangeMamba, Insta-BDA-RB (rule-based fusion), and Insta-BDA-MLP (learnable fusion).

Table 4. Segmentation and cross-dataset generalization results. xBD: In-distribution test set (933 images). IAN-BD: Out-of-distribution aerial imagery (101 images). IDA-BD: Out-of-distribution satellite imagery (87 images).

	xBD Test Set				IAN-BD (OOD)			IDA-BD (OOD)		
Method	**Macro F1**	**mIoU**	**ND F1**	**D F1**	**Loc F1**	**Dmg F1**	**mIoU**	**Loc F1**	**Dmg F1**	**mIoU**
SAM3 Only	0.393	0.253	0.525	0.260	0.796	0.410	0.258	0.642	0.301	0.188
ChangeMamba	**0.778**	**0.640**	**0.833**	**0.722**	0.741	0.520	0.352	0.726	0.355	0.269
Insta-BDA-RB	0.731	0.580	0.792	0.670	**0.786**	0.545	0.375	**0.746**	0.366	0.279
Insta-BDA-MLP	0.741	0.591	0.787	0.696	**0.786**	**0.548**	**0.378**	**0.746**	**0.366**	**0.279**

ND: Non-damaged. D: Damaged. Loc: Localization. Dmg: Damage classification. OOD: Out-of-distribution. Best per column in **bold**.

On its training domain (xBD), ChangeMamba achieves the highest segmentation performance (Macro F1 = 0.778). However, under cross-dataset evaluation, Insta-BDA consistently outperforms ChangeMamba. On IAN-BD, Insta-BDA-MLP improves damage F1 from 0.520 to 0.548 (+5.4%), and on IDA-BD from 0.355 to 0.366 (+3.1%). This improved generalization can be attributed to SAM3's zero-shot instance detection capability, which remains robust across imaging modalities.

The learnable fusion model (Insta-BDA-MLP) further improves over the rule-based variant (Insta-BDA-RB). On xBD, damage F1 increases from 0.670 to 0.696 (+3.9%). On IAN-BD, damage F1 improves from 0.545 to 0.548. Although cross-dataset gains are modest, they indicate that a learned decision boundary transfers more effectively than manually designed thresholds. On IDA-BD, both fusion variants achieve identical damage F1 (0.366).

Table 5 compares building counting accuracy on the xBD test set.

Table 5. Building counting accuracy on xBD test set (933 images).

Method	**Detected**	**Deviation**	**Dev. (%)**	**MAE**	**MRPE (%)**
Ground Truth	54,862	—	—	—	—
SAM3 Only	35,683	−19,179	−34.96%	—	—
ChangeMamba	29,045	−25,817	−47.06%	35.0	−25.9
Insta-BDA-RB/MLP	35,683	−19,179	**−34.96%**	41.5	+27.6

Note: Insta-BDA-RB/MLP and SAM3 Only yield identical counts, as Insta-BDA relies on SAM3 for instance detection. MAE and MRPE are per-image averages over 752 images containing at least one building. The proposed method (Insta-BDA) and the best counting deviation are shown in **bold**.

Both SAM3 and Insta-BDA detect 35,683 buildings, corresponding to a −34.96% counting error, since Insta-BDA adopts SAM3's instance detections for counting. The key distinction is that Insta-BDA additionally provides per-building damage classification, which SAM3 alone does not. In contrast, ChangeMamba's semantic segmentation significantly undercounts buildings (−47.06%) due to fragmented predictions. SAM3-based undercounting results from overlapping instances merging when rasterized to spatial masks, as well as missed small or occluded structures.

It is important to note that semantic segmentation models such as ChangeMamba are not designed for building counting; their building counts are derived post hoc through connected component analysis of pixel-level predictions. The counting results presented here should therefore be understood as demonstrating a capability gap rather than a performance benchmark: instance-aware methods enable per-building counting as an additional output that semantic models cannot natively provide. Within the instance-level evaluation, SAM3 Only (zero-shot, instance-aware) serves as the most appropriate direct baseline, as it shares the same detection mechanism as Insta-BDA but lacks the damage classification fusion.

Qualitative results on the xBD test set (Figure 4) show that both Insta-BDA variants generate cleaner, instance-level predictions compared to ChangeMamba's pixel-wise outputs. The Insta-BDA-MLP variant further improves classification consistency by learning per-instance decision boundaries.

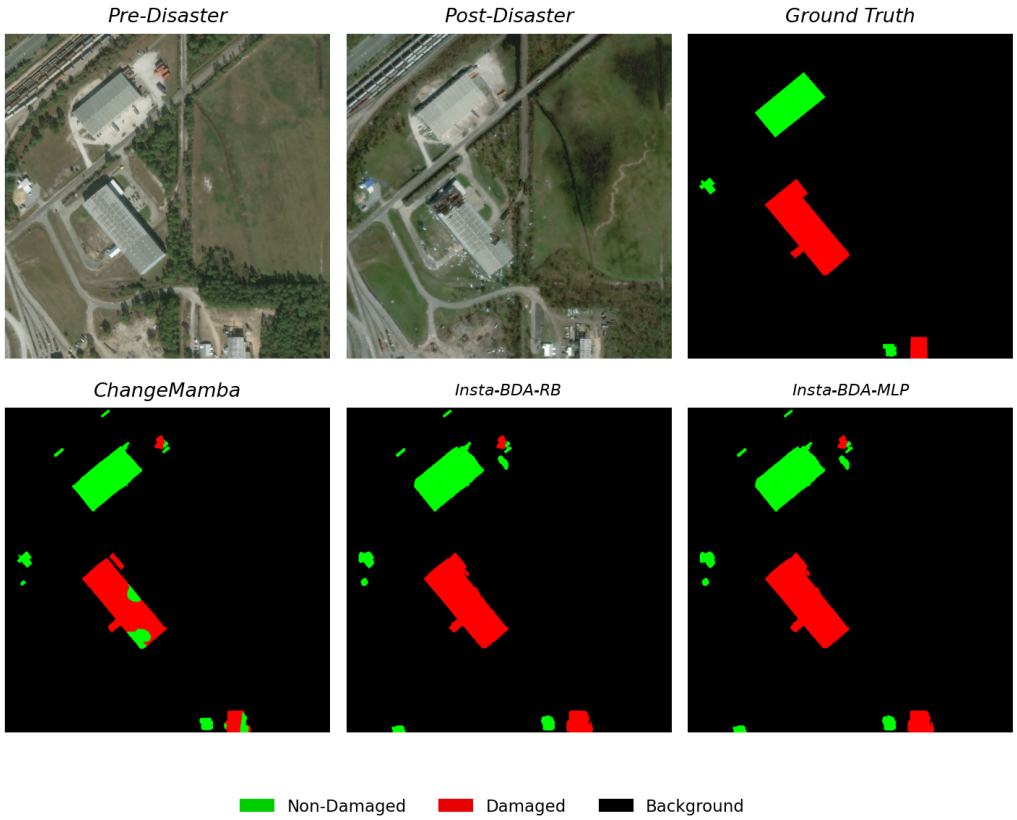

Figure 4. Qualitative comparison on the xBD test set (Hurricane Michael). Top row: Input imagery and ground truth. Bottom row: ChangeMamba produces fragmented predictions, while Insta-BDA-RB and Insta-BDA-MLP yield cleaner instance-level masks with correct damage classification. Green = non-damaged, red = damaged.

On the out-of-distribution IAN-BD dataset (Figure 5), ChangeMamba produces fragmented predictions with mixed damage labels within individual buildings. In contrast, Insta-BDA-RB and Insta-BDA-MLP assign a single damage label per instance, better align-

ing with the ground-truth annotation protocol. Insta-BDA produces cleaner building boundaries with consistent labels, whereas ChangeMamba scatters damage pixels within the same structure.

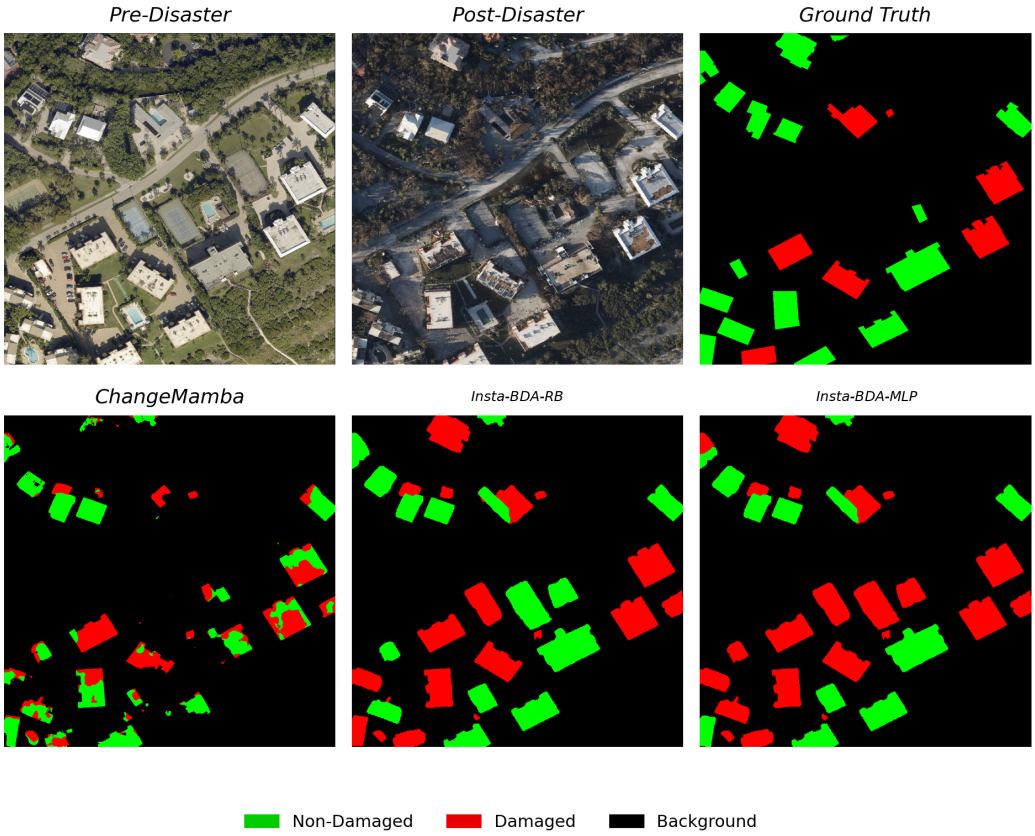

Figure 5. Qualitative comparison on IAN-BD (out-of-distribution Hurricane Ian imagery). ChangeMamba produces fragmented predictions with mixed classes within buildings, while Insta-BDA-RB and Insta-BDA-MLP assign one damage class per instance. Green = non-damaged, red = damaged.

3.1. Detailed Analysis

We analyzed performance across the six disaster categories represented in the xBD test set (Table 6), comparing Insta-BDA-RB and Insta-BDA-MLP.

Table 6. Performance by disaster type (RB and MLP).

Disaster	#	Macro F1		Cnt Err	
		RB	MLP	RB	MLP
Hurricane	387	0.363	**0.365**	−28.5%	−28.4%
Wildfire	381	0.272	0.271	−27.5%	−27.5%
Flood	80	0.313	**0.314**	−25.2%	−25.2%
Tsunami	42	0.488	**0.491**	−78.4%	−78.4%
Earthquake	38	0.403	0.403	−78.9%	−78.9%
Volcanic	5	0.329	0.329	−15.6%	−15.6%
Overall	933	0.329	0.329	−50.1%	−50.1%

Per-image Macro F1 and per-image counting error (%), averaged across images. Aggregate total counting error is −34.96% (Table 5). For each disaster type, the higher Macro F1 between RB and MLP is shown in **bold**.

Both fusion strategies exhibit consistent behavior across disaster types. The highest F1 scores occur for tsunami (0.491) and earthquake (0.403), which contain densely clustered buildings. Wildfire scenes are the most challenging (0.271), likely due to ambiguous building

footprints caused by fire damage. Counting errors are negative across all disaster types, with the largest undercounting observed in earthquake (−78.9%) and tsunami (−78.4%) scenes, where collapsed structures yield indistinct boundaries. The MLP variant provides slight F1 improvements over RB for the hurricane and tsunami categories.

We further evaluated performance with respect to building density (Table 7), grouping images by ground-truth building count.

Table 7. Performance by building density.

Density	#	Macro F1		Cnt Err	MAE	
		RB	MLP	RB/MLP	ChangeMamba	Insta-BDA
1–10 (sparse)	241	0.319	0.318	+13.6%	1.5	3.7
11–50 (moderate)	253	0.433	**0.439**	−4.0%	6.7	13.3
51–100 (dense)	105	**0.467**	0.462	−23.5%	17.0	17.9
101–200 (v. dense)	89	0.466	**0.467**	−45.8%	47.0	**37.0**
200+ (ultra dense)	64	0.461	0.461	−75.1%	286.6	339.6
Overall	933	0.329	0.329	−50.1%	35.0	41.5

Per-image averages. Images with no buildings (181) excluded. Cnt Err: RB values shown; MLP differs by ≤0.2%. MAE: mean absolute error (buildings per image). Best value between the compared columns in each row is shown in **bold**.

Insta-BDA slightly overcounts in sparse scenes (+13.6%), where SAM3 occasionally detects non-building structures. Undercounting increases with density, ranging from −4.0% (moderate density) to −75.1% (ultra-dense scenes), where building boundaries become difficult to distinguish. Per-image Macro F1 increases with density, from 0.319 (sparse) to 0.461 (ultra-dense). The MLP variant shows slight improvements in moderate-density (0.439 vs. 0.433) and very dense scenes (0.467 vs. 0.466). Notably, in very dense scenes (101–200 buildings), Insta-BDA outperforms ChangeMamba in per-image MAE (37.0 vs. 47.0), as SAM3's instance-level detections handle densely packed buildings better than connected-component counting, which tends to merge adjacent structures.

We also examined performance as a function of average building size (Table 8), where size is defined as the mean ground-truth building area (in pixels) per image.

Table 8. Performance by average building size (RB and MLP).

Avg Size	#	Macro F1		Cnt Err	
		RB	MLP	RB	MLP
<500 px	194	0.307	0.307	−55.7%	−55.6%
500–2k px	489	0.443	**0.444**	−50.0%	−50.0%
2k–5k px	54	**0.450**	0.448	−35.0%	−35.0%
5k+ px	13	0.447	**0.448**	−21.4%	−21.0%
Overall	933	0.329	0.329	−50.1%	−50.1%

Per-image averages. Images with no buildings (183) excluded. Cnt: counting error (%). The higher Macro F1 between RB and MLP in each row is shown in **bold**.

Macro F1 increases with building size, from 0.307 for small buildings (<500 px) to 0.448 for very large buildings (>5k px), reflecting clearer segmentation boundaries in larger structures. Undercounting correspondingly decreases, from −55.7% (small buildings) to −21.4% (very large buildings).

3.2. Ablation Studies

To isolate the contribution of each design component in Insta-BDA, we conducted controlled ablation experiments (Table 9).

- SAM3 + SSIM: Uses SAM3 detections with SSIM-based damage classification, without incorporating ChangeMamba's pixel-level change predictions.
- ChangeMamba + Connected Components: Uses ChangeMamba predictions with standard connected component counting instead of SAM3 instances.
- Damage Threshold Variants: Tests thresholds of 0, 0.25, and 0.50 (default = 0.33) for labeling a building as damaged.
- No Morphological Filtering: Removes the post-processing cleanup stage.

Table 9. Ablation study on xBD holdout set (933 images).

Configuration	Macro F1	ND F1	D F1
SAM3 + SSIM	0.393	0.525	0.261
ChangeMamba + conn. comp.	**0.780**	**0.832**	**0.728**
Damage thresh. = 0	0.736	0.791	0.682
Damage thresh. = 0.25	0.744	0.793	0.696
Damage thresh. = 0.50	0.742	0.792	0.691
No morph. filtering	0.744	0.792	0.695
Full Insta-BDA-RB	0.744	0.793	0.695

ND: Non-damaged. D: Damaged. Conn. comp.: Connected components. Best value in each column is shown in **bold**.

SAM3 combined with SSIM achieves only 0.393 Macro F1, confirming that ChangeMamba's pixel-level change detection is critical. Using ChangeMamba with connected components achieves the highest Macro F1 (0.780), demonstrating strong segmentation quality. However, connected component counting underestimates buildings by −47.06%, which motivates the use of SAM3 instance detections in Insta-BDA.

Damage threshold variations have minimal effect (Macro F1 between 0.736 and 0.744 across thresholds 0–0.50), and removing morphological filtering produces negligible change. The full Insta-BDA-RB pipeline achieves 0.744 Macro F1, representing a modest reduction relative to ChangeMamba + connected components (0.780), in exchange for substantially improved counting accuracy.

Note that Insta-BDA-MLP is excluded from the ablation table because it was trained on the same holdout split used for ablation evaluation. Its performance is reported separately in Table 4 (xBD Macro F1 = 0.741, D F1 = 0.696), where it demonstrates consistent gains over the rule-based fusion, particularly under cross-dataset evaluation.

3.3. Instance-Level Evaluation

To complement the pixel-level metrics reported above, we evaluated Insta-BDA at the instance level using the standard COCO evaluation protocol [43]. Predicted building instances from SAM3 were matched to ground-truth polygons from the xBD test set using mask IoU, and we report Average Precision (AP) and Average Recall (AR) following the COCO convention. We evaluated both fusion variants: Insta-BDA-RB (rule-based) and Insta-BDA-MLP (learnable). Table 10 summarizes the results over 752 test images containing at least one building (54,862 GT annotations, 35,601 predicted instances).

Table 10. Instance-level evaluation on xBD test set (COCO protocol).

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Insta-BDA-RB	0.131	0.291	0.098	0.080	0.224	0.210
Insta-BDA-MLP	**0.134**	**0.298**	**0.101**	**0.081**	**0.230**	**0.214**

AP: Average Precision at IoU 0.50:0.05:0.95. AP_{50}/AP_{75}: AP at IoU 0.50/0.75. AP_S/AP_M/AP_L: AP for small/medium/large buildings (COCO area thresholds). Best per column in **bold**.

Table 11 reports recall and per-category AP to characterize detection completeness and damage classification quality at the instance level.

Table 11. Instance-level recall and per-category AP on xBD test set.

Method	AR_{100}	AR_S	AR_L	AP_{ND}	AP_D
Insta-BDA-RB	0.213	0.129	0.548	0.124	0.138
Insta-BDA-MLP	0.213	0.128	**0.552**	0.124	**0.145**

AR_{100}: Average Recall with 100 detections per image. AR_S/AR_L: AR for small/large buildings. AP_{ND}/AP_D: per-category AP for non-damaged/damaged buildings. Best value in each column is shown in **bold**.

Insta-BDA-MLP consistently outperforms Insta-BDA-RB across all AP metrics, with the largest improvement on damaged buildings (AP_D: 0.145 vs. 0.138, $AP_{50,D}$: 0.328 vs. 0.311). This confirms that the learnable fusion transfers pixel-level damage cues to the instance level more effectively than the rule-based heuristic. Both variants achieve substantially higher AP on medium and large buildings ($AP_M \approx 0.23$) than on small buildings ($AP_S \approx 0.08$), reflecting SAM3's difficulty in detecting small structures under the text-prompted "building" setting. Overall recall ($AR_{100} = 0.213$) indicates that approximately one in five GT buildings is detected with sufficient mask overlap, with recall increasing to 0.55 for large buildings.

For reference, supervised Mask R-CNN [7] models achieve AP 35–40 on their training domain; a zero-shot pipeline achieving $AP_{50} \approx 0.30$ on out-of-domain remote sensing imagery underscores the practical viability of this approach. We note that SAM3 operates entirely in a zero-shot setting and is not expected to match supervised building extraction models trained on domain-specific data. The Insta-BDA framework is modular: any improved building instance provider can replace SAM3 without modifying the fusion logic, and the localization performance would improve accordingly.

We note that instance-level building damage assessment on xBD represents a gap in the existing literature: nearly all published work on this dataset uses semantic segmentation, and to the best of our knowledge no instance segmentation results on xBD have been reported. Training a supervised instance segmentation model (e.g., Mask R-CNN) on xBD would require polygon-level instance annotations, which contradicts the zero-shot premise of our framework. The natural instance-level baseline is SAM3 Only (SSIM-based damage classification without ChangeMamba fusion), which shares the same underlying detections as Insta-BDA but achieves substantially lower damage classification quality (Macro F1 = 0.393, Table 9). The COCO AP results in Tables 10 and 11 quantify instance-level performance using the standard evaluation protocol and establish an initial benchmark for future instance-aware methods on this dataset.

3.4. Threshold Sensitivity Analysis

To demonstrate that the fusion strategy is robust to specific parameter choices, and to address the design of the centroid matching threshold, we systematically vary four key thresholds and report their effect on performance. Table 12 presents results on the xBD holdout set, which is used here to avoid contaminating the test set with parameter selection decisions.

The results indicate that Insta-BDA-RB is robust across a wide range of threshold values. The centroid matching threshold $\tau_{centroid}$ has moderate impact: Macro F1 varies by only 0.010 across the full 10–50 pixel range, with a gradual improvement at larger values. The selected value of 30 pixels provides a balance between matching precision and recall. Making this threshold adaptive (e.g., proportional to individual building size) is

a promising direction; however, the fixed threshold already achieves stable performance across the diverse building scales present in xBD.

Table 12. Sensitivity to key thresholds on xBD holdout set.

Parameter	Value	Macro F1	D F1	Cnt Err (%)
$\tau_{centroid}$ (px)	10	0.755	0.720	6.3
	20	0.761	0.725	7.8
	30 (default)	0.763	0.727	8.3
	40	0.764	0.728	8.4
	50	0.765	0.730	8.3
τ_{ssim}	0.4	0.763	0.728	8.3
	0.5	0.763	0.728	8.3
	0.6 (default)	0.763	0.727	8.3
	0.7	0.762	0.727	8.3
r_{dmg} threshold	0.20	0.763	0.728	8.3
	0.33 (default)	0.763	0.727	8.3
	0.50	0.760	0.723	8.3
	0.67	0.760	0.723	8.3
ρ threshold	0.3	0.762	0.727	9.2
	0.5 (default)	0.763	0.727	8.3
	0.7	0.761	0.725	5.9
	0.9	0.750	0.709	2.0

Each parameter is varied independently while others remain at default values. Bold indicates the default configuration. Note: morphological post-processing is excluded from this analysis to isolate the direct effect of each threshold on the fusion decision.

The SSIM threshold τ_{ssim} and damage ratio r_{dmg} threshold both exhibit minimal performance variation (Macro F1 ± 0.001 and ± 0.003, respectively), with no abrupt degradation at any tested value. The global coverage ratio ρ threshold shows stable behavior for values $\leq$ 0.7, with Macro F1 varying by less than 0.002; only the extreme value $\rho = 0.9$ causes a notable drop (-0.013), as it restricts the fusion to very few images. These results confirm that Insta-BDA's performance is not critically dependent on any single parameter choice.

While the fixed threshold values are calibrated for satellite imagery at approximately 0.5 m/pixel ground sample distance (GSD), as in xBD, they may require adjustment for imagery at different resolutions. For deployment on higher- or lower-resolution imagery, the centroid matching threshold can be scaled proportionally: $\tau_{centroid} = \tau_{base} \times (\mathrm{GSD}_{xBD}/\mathrm{GSD})$, where $\tau_{base} = 30$ pixels and $\mathrm{GSD}_{xBD} \approx 0.5$ m/pixel. Similarly, SSIM-based thresholds may require recalibration when image quality characteristics differ substantially from the training domain. We note that this adaptive formula has not been validated on non-xBD resolutions, and empirical evaluation across multiple GSD values remains an important direction for establishing the generalizability of the fixed-threshold approach.

3.5. Feature Ablation and Confidence Calibration

To understand which input features contribute most to the MLP fusion decision, we conducted leave-one-out ablation experiments. For each of the eight input features, we zeroed out their values across all instances and retrained the MLP, evaluating on the xBD holdout set using stratified 5-fold cross-validation, ensuring that the model was never evaluated on its own training data within each fold. We adopted zero-out ablation rather than feature removal to preserve the network architecture across all experiments, ensuring that differences in performance would be attributable to the information content of each feature rather than changes in model capacity. Table 13 reports the resulting changes in Macro F1.

Table 13. Feature ablation for Insta-BDA-MLP (leave-one-out).

Removed Feature	Macro F1	Δ F1
None (full model)	0.912	—
$\bar{c}_{cm}$ (CM confidence)	0.907	−0.005
r_{dmg} (damage ratio)	0.907	−0.005
r_{nd} (non-damaged ratio)	0.907	−0.005
r_{bg} (background ratio)	0.906	−0.006
s_{sam} (SAM3 score)	0.908	−0.004
d_{sam} (SAM3 damage)	0.911	−0.001
ρ (coverage ratio)	0.915	+0.003
Δc (conf. difference)	0.911	−0.002

Five-fold cross-validation on xBD holdout set. Δ F1: Change relative to full model. Note: The full-model baseline (0.912) differs from the CV Macro F1 reported in Section 2.6 (0.860) because the ablation experiment involves an independent training run with different random weight initialization; the relative Δ F1 values, which are the focus of this analysis, are unaffected.

The most influential feature is r_{bg} (background ratio), whose removal causes the largest F1 drop (-0.006). The pixel-level features $\bar{c}_{cm}$, r_{dmg}, and r_{nd} each contribute approximately -0.005, indicating that ChangeMamba's per-pixel damage predictions are the primary signal driving the fusion decision. The SAM3 detection score s_{sam} contributes moderately (-0.004), while the SAM3 damage label d_{sam} and confidence difference Δc have smaller individual effects (-0.001 and -0.002), as they capture information that is partially redundant with other features. Notably, removing the coverage ratio ρ slightly improves F1 (+0.003), suggesting that this global feature may introduce noise in per-instance decisions; the MLP can compensate for its absence through the remaining features.

To assess whether Insta-BDA-MLP produces calibrated predictions, we computed the Expected Calibration Error (ECE) using out-of-fold predictions from the 5-fold cross-validation on the xBD holdout set. For each fold, the model trained on the remaining four folds produced softmax probabilities for the held-out instances; these out-of-fold predictions were then binned into 10 equally spaced confidence intervals to compare predicted probability against observed accuracy. This ensured that ECE was computed on data not seen during training, providing an unbiased estimate of calibration quality. Insta-BDA-MLP achieved an ECE of 0.208, indicating moderate miscalibration; that is, the model's predicted confidence did not closely match its actual accuracy. This is distinct from overfitting: the model may generalize well in terms of classification accuracy while still producing poorly calibrated probability estimates. ECE could be further reduced through temperature scaling or Platt calibration in future work. The calibration concern is most acute for the rule-based variant (Insta-BDA-RB), which directly compares raw confidence scores from ChangeMamba and SAM3 that are not inherently on the same scale. The MLP variant (Insta-BDA-MLP) mitigates this by learning the fusion decision from the full eight-dimensional feature vector rather than relying solely on raw confidence comparison, reducing its sensitivity to miscalibrated individual scores. However, as the ECE of 0.208 indicates, the MLP's own predicted probabilities remain moderately miscalibrated; applying post hoc calibration techniques such as temperature scaling or Platt calibration could further improve reliability.

3.6. Runtime Across Resolutions

To assess scalability for operational deployment, we measured the inference time for each pipeline component at three input resolutions (Table 14). All measurements were averaged over 20 images on a single NVIDIA GeForce RTX 3060 GPU.

Table 14. Inference time (seconds) by image resolution.

Component	256 × 256	512 × 512	1024 × 1024
ChangeMamba inference	0.11	0.22	0.83
SAM3 inference	2.46	2.64	3.27
RB fusion	0.03	0.16	0.81
MLP fusion	0.02	0.14	0.65
Total Insta-BDA-RB	2.60	3.02	4.91
Total Insta-BDA-MLP	2.59	3.00	4.75
GPU Memory (GB)	4.4	4.6	6.9

Averaged over 20 images. GPU: NVIDIA GeForce RTX 3060. **Bold** rows denote the total end-to-end pipeline times.

SAM3 constitutes the primary computational bottleneck across all resolutions, accounting for over 65% of the total inference time. At 512 × 512, end-to-end processing completes in approximately 3 s, making near-real-time assessment feasible for tiled large-area imagery. At 1024 × 1024, the total time increases to under 5 s, with ChangeMamba inference and fusion each scaling roughly quadratically while SAM3 scales sub-linearly due to its internal tiling strategy. GPU memory usage remains under 7 GB across all tested resolutions, confirming that the pipeline is deployable on consumer-grade hardware.

3.7. SSIM Robustness Under Perturbation

To quantify the sensitivity of the SSIM-based damage cue to common acquisition artifacts, we introduced controlled synthetic perturbations to the post-disaster imagery in the xBD holdout set and measured the resulting change in pipeline performance. This analysis was conducted on the holdout set using the rule-based variant (Insta-BDA-RB), which does not involve any learned parameters, so there was no risk of data leakage. Perturbations were applied exclusively to the post-disaster image provided to SAM3, which uses SSIM for building-level damage assessment; ChangeMamba predictions were computed from the original unperturbed imagery. This isolated the effect of each perturbation on the SSIM-based damage component of the pipeline. Table 15 reports results for Insta-BDA-RB under each perturbation type.

Table 15. Effects of synthetic perturbations on Insta-BDA-RB performance (xBD holdout).

Perturbation	Macro F1	D F1	Δ F1
None (baseline)	0.707	0.636	—
Shift ±2 px	0.708	0.635	+0.001
Shift ±5 px	0.701	0.631	−0.006
Shift ±10 px	0.691	0.619	−0.017
Rotation ±1°	0.695	0.619	−0.012
Rotation ±3°	0.667	0.597	−0.040
Rotation ±5°	0.616	0.575	−0.092
Brightness ±10%	0.704	0.630	−0.003
Brightness ±20%	0.709	0.642	+0.002
Brightness ±30%	0.706	0.633	−0.002

Perturbations applied to post-disaster images only. Δ F1: Change relative to unperturbed baseline.

Small spatial shifts (±2 px) and mild brightness variations (±10–30%) have negligible impact on performance ($|\Delta|$ F1 ≤ 0.003), indicating that the pipeline tolerates typical co-registration errors and illumination differences present in satellite imagery. Larger shifts (±10 px) reduce Macro F1 by 0.017, primarily due to centroid mismatches that prevent correct building pairing. Rotations are the most disruptive perturbation type: ±3° degrades

Macro F1 by 0.040, and $\pm 5^\circ$ causes a substantial drop of 0.092, as the SSIM computation becomes unreliable when building regions are spatially misaligned. Brightness changes up to $\pm 30\%$ cause negligible impact ($|\Delta|$ F1 ≤ 0.003), demonstrating SSIM's robustness to radiometric variation within typical acquisition conditions. These results suggest that, for datasets with significant geometric misalignment, additional pre-processing (e.g., local image registration) would be beneficial, while radiometric normalization is less critical.

4. Discussion

4.1. Interpretation of Results

The lower pixel-level F1 observed for Insta-BDA can be attributed to its instance-aware smoothing behavior. In the ground truth, each building is assigned a single damage label based on polygon-level annotation. However, pixel-wise semantic segmentation models such as ChangeMamba frequently produce mixed predictions within a single building, labeling some pixels as non-damaged and others as damaged within the same structure.

Since F1 is computed at the pixel level, ChangeMamba's mixed predictions may partially overlap with ground-truth pixels even when the overall building-level classification is incorrect. In contrast, Insta-BDA enforces exactly one damage label per detected building instance, resulting in smoother and more uniform outputs that better reflect the per-building annotation protocol.

The apparent F1 trade-off arises because of the following:

- ChangeMamba's noisy, mixed-class predictions can coincidentally align with subsets of ground-truth pixels;
- Insta-BDA produces clean, single-class predictions for each building instance;
- If Insta-BDA misclassifies a building, all pixels belonging to that instance are counted as incorrect.

Furthermore, ChangeMamba is trained and evaluated on the same dataset (xBD), providing a domain-specific advantage in pixel-level metrics. By contrast, SAM3 operates in a zero-shot setting without any xBD-specific training. This domain discrepancy contributes to the lower F1 observed for Insta-BDA. Despite this, Insta-BDA substantially reduces the aggregate building count deviation (-34.96% vs. -47.06%), highlighting the benefit of foundation-model generalization. The instance-aware formulation aligns more closely with the ground-truth annotation scheme and produces outputs that are more actionable for disaster response scenarios.

Specifically, Insta-BDA enables instance-level building counting for resource allocation planning and provides per-building damage categorization suitable for insurance assessment and reconstruction prioritization.

We further compared the learnable fusion variant, Insta-BDA-MLP, with the rule-based fusion, Insta-BDA-RB, across all three datasets. The MLP was trained on per-instance features extracted from the xBD holdout split (31,679 building instances from 933 images) using stratified 5-fold cross-validation, achieving a Macro F1 of 0.860 ± 0.005 during cross-validation.

On the xBD test set, Insta-BDA-MLP improves Macro F1 from 0.731 (RB) to 0.741 (+1.4%), driven primarily by improved damage classification (D F1: 0.696 vs. 0.670, corresponding to a +3.9% relative improvement). The non-damaged F1 remains comparable (0.787 vs. 0.792), indicating that the MLP slightly reduces non-damaged precision in exchange for substantially improved damage recall.

Under cross-dataset evaluation, Insta-BDA-MLP matches or slightly exceeds the rule-based approach. On IAN-BD, damage F1 increases from 0.545 to 0.548; on IDA-BD, both variants achieve 0.366. These results indicate that the learned decision boundary generalizes to unseen datasets despite being trained exclusively on xBD holdout data.

The learnable fusion introduces negligible computational overhead (Section 4.3). The MLP contains ∼2400 parameters, loads in under 0.01 s, and its per-instance forward pass is faster than the rule-based comparison (0.41 s vs. 0.57 s per image). While Insta-BDA-RB remains attractive for its interpretability and lack of additional training requirements, the MLP provides modest yet consistent performance gains when holdout annotations are available.

4.2. Statistical Robustness

As both models are deterministic (i.e., no stochastic training or inference at test time), we evaluated statistical robustness using bootstrap confidence intervals and hypothesis testing. Specifically, we performed bootstrap resampling with 10,000 iterations over the per-image F1 scores computed across all 933 test images. The resulting confidence intervals and statistical comparisons are summarized in Table 16.

Table 16. Statistical robustness analysis (10,000 bootstrap resamples). F1 metrics computed over 752 images containing buildings, counting MAE over all 933 images.

Method	Metric	Mean ± Std	95% CI
ChangeMamba	Macro F1	0.279 ± 0.184	$[0.266, 0.292]$
	ND F1	0.533 ± 0.369	$[0.506, 0.559]$
	D F1	0.026 ± 0.098	$[0.019, 0.033]$
Insta-BDA-RB	Macro F1	0.256 ± 0.172	$[0.243, 0.268]$
	ND F1	0.481 ± 0.346	$[0.456, 0.506]$
	D F1	0.031 ± 0.104	$[0.023, 0.038]$
ChangeMamba	Count MAE	28.3 ± 89.3	$[22.7, 34.4]$
Insta-BDA-RB	Count MAE	45.1 ± 96.5	$[39.1, 51.8]$

Wilcoxon signed-rank: F1 $p < 10^{-56}$, count MAE $p < 10^{-64}$. **Bold** denotes the proposed method (Insta-BDA-RB).

The narrow confidence intervals indicate stable performance across different image subsets. A paired Wilcoxon signed-rank test yields $p < 0.001$ for both F1 and counting MAE, confirming that the observed differences are statistically significant. We note that three different Macro F1 values are reported for Insta-BDA-RB across different tables, reflecting different aggregation methods and image subsets: Table 4 reports aggregate F1 (0.731), computed over all pixels pooled across the test set; Table 6 reports per-image F1 averaged over all 933 images (0.329), where the 181 images with no buildings contribute F1 = 0; and Table 16 reports per-image F1 averaged over the 752 images containing at least one building (0.256). These are not contradictory; aggregate F1 is dominated by large images with many pixels, while per-image averaging weights each image equally regardless of size.

Instance Detection Errors: SAM3 may fail to detect buildings in densely built environments or when structures exhibit atypical appearances (e.g., industrial facilities, temporary shelters), leading to undercounting. Conversely, large buildings or complexes may be oversegmented into multiple instances, artificially inflating counts. Furthermore, false positive detections can arise when SAM3 misidentifies non-building objects such as vegetation or debris as structures, as illustrated in Figure 6. These errors propagate directly through the pipeline, affecting both building counts and damage estimates, and are most pronounced in rural or heavily vegetated scenes.

Scene Complexity and Occlusion: Buildings obscured by vegetation, debris, smoke, or neighboring structures pose challenges for both SAM3 and ChangeMamba. In large-scale collapse scenarios, individual building footprints may merge into rubble fields, making instance delineation difficult. Similarly, boundary inaccuracies in SAM3 masks can cause partial overlaps between adjacent buildings or inclusion of surrounding non-building regions.

Temporal and Geometric Misalignment: The approach assumes reasonable spatial alignment between pre- and post-disaster imagery. Significant viewpoint shifts, seasonal variations, or illumination changes between acquisition dates may degrade the matching performance between pre-disaster instance masks and post-disaster damage predictions. The SSIM-based damage cue (Equation (2)) is particularly sensitive to such perturbations: illumination differences between acquisition dates can reduce SSIM independently of structural damage, producing false positive damage labels, while minor misregistration between co-registered image pairs shifts building boundaries and introduces spurious structural differences. The xBD dataset provides pre-registered image pairs, partially mitigating these effects, but cross-dataset application to imagery with larger geometric or radiometric discrepancies may require additional pre-processing such as histogram matching or local image registration. We quantify the sensitivity of the SSIM signal to synthetic perturbations in Section 3.7.

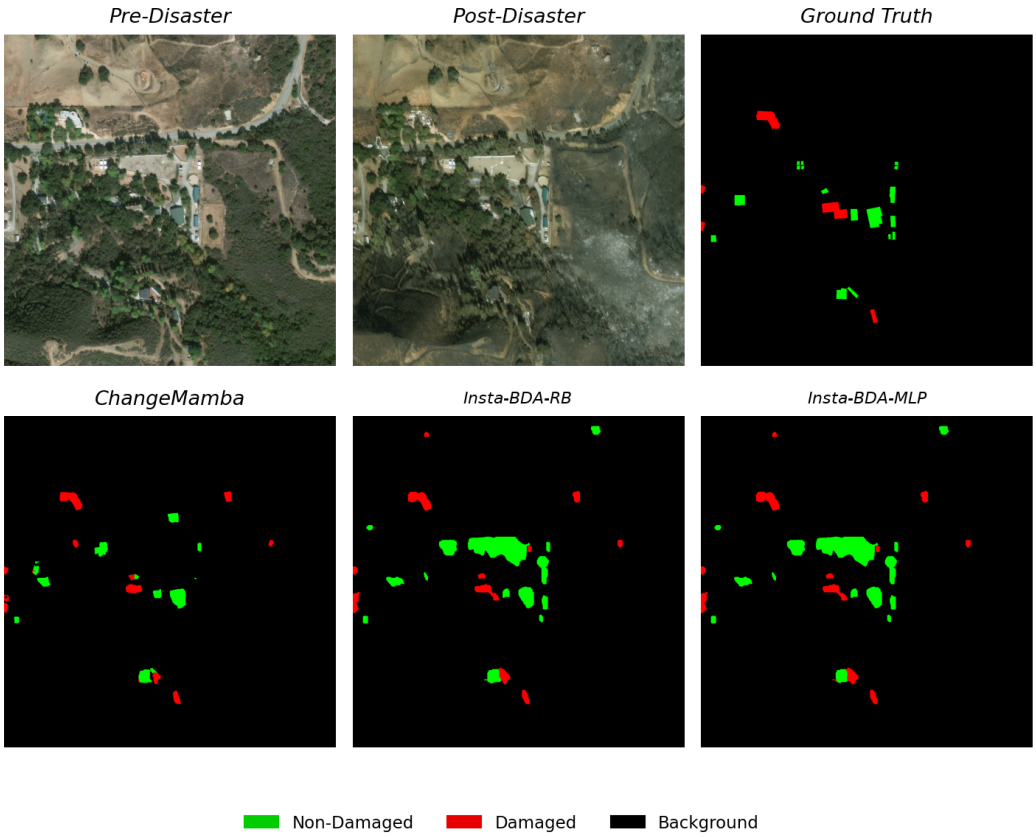

Figure 6. Failure case illustrating SAM3 false positives (Southern California wildfire). Vegetation and debris are mistakenly detected as buildings by SAM3, leading Insta-BDA to generate spurious predictions not present in the ground truth. Green denotes non-damaged; red denotes damaged.

4.3. Computational Requirements

Inference efficiency is essential for disaster response scenarios where rapid assessment is required. Table 17 summarizes the computational performance of each component on a single NVIDIA GeForce RTX 3060 GPU using 1024 × 1024 input images, averaged over 20 test samples.

ChangeMamba and SAM3 inference constitute the primary computational bottlenecks in the pipeline. The MLP-based fusion is marginally faster than the rule-based approach (0.41 s vs. 0.57 s per image), as vectorized forward passes replace iterative comparison logic. For large-scale deployments, ChangeMamba and SAM3 can be executed in parallel on separate GPUs to further reduce end-to-end latency.

Table 17. Computational requirements per image.

Component	Time (s)	GPU Mem (GB)
ChangeMamba inference	0.86	6.5
SAM3 inference	3.06	6.5
RB fusion (rule-based)	0.57	–
MLP fusion (learnable)	0.41	–
Total Insta-BDA-RB	4.50	6.5
Total Insta-BDA-MLP	4.34	6.5

Bold rows denote the total end-to-end pipeline times.

Building damage assessment is inherently an offline analysis task; therefore, strict real-time constraints are not required. The additional inference time is justified by the enhanced outputs, namely, per-building damage classification and more reliable building count estimates.

5. Conclusions

We introduced Insta-BDA, a unified framework that integrates change detection with foundation-model-based instance segmentation for building damage assessment. Through confidence-guided fusion at the building-instance level, Insta-BDA preserves structural boundaries while incorporating pixel-level damage cues. We investigated two fusion strategies: a transparent rule-based formulation (Insta-BDA-RB) and a lightweight learnable variant based on a small MLP (Insta-BDA-MLP).

This work introduces an instance-level fusion mechanism that integrates ChangeMamba change detection with SAM3 instance segmentation, offered in both rule-based and learnable forms. By leveraging instance-aware predictions, the proposed approach reduces the aggregate building count deviation to -34.96%, compared to -47.06% for ChangeMamba, while producing per-building damage classifications aligned with polygon-level ground-truth annotations suitable for structured reporting. Moreover, we have demonstrated that a lightweight learnable fusion module (MLP with $\sim$2400 parameters) improves damage F1 from 0.670 to 0.696 over the rule-based variant, and that both fusion strategies exhibit stronger cross-dataset generalization than ChangeMamba alone.

Importantly, the proposed fusion framework is model-agnostic and can be extended to other change detection backbones beyond ChangeMamba. The learnable fusion results further show that minimal supervision at the instance-feature level can enhance damage classification performance without degrading counting accuracy or introducing meaningful computational overhead.

A key direction for future work is reducing inference latency through model distillation, lighter segmentation architectures, and parallel GPU execution. More broadly, Insta-BDA moves toward operationally deployable disaster assessment systems that translate remote sensing advances into actionable intelligence for emergency response and recovery planning.

Author Contributions: Conceptualization, B.G., M.B., and S.S.; methodology, B.G., M.B., and S.S.; software, B.G.; validation, B.G., M.B., F.B., and S.S.; formal analysis, B.G., M.B., and F.B.; investigation, B.G. and A.K.; resources, B.G. and A.K.; data curation, B.G.; writing—original draft preparation, B.G.; writing—review and editing, S.S., M.B., F.B., and A.K.; visualization, B.G.; supervision, S.S., M.B., and F.B.; project administration, M.B. and S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The xBD dataset is publicly available at https://xview2.org/. The IDA-BD dataset is available at https://doi.org/10.17603/ds2-pd9h-6k05. The IAN-BD dataset is available at https://doi.org/10.17603/ds2-2n0d-4s77.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BDA	Building Damage Assessment
CNN	Convolutional Neural Network
MLP	Multilayer Perceptron
SAM	Segment Anything Model
SSIM	Structural Similarity Index
xBD	xView2 Building Damage
OOD	Out-of-Distribution
ND	Non-Damaged
D	Damaged

References

1. Braik, A.M.; Koliou, M. Automated Building Damage Assessment and Large-Scale Mapping by Integrating Satellite Imagery, GIS, and Deep Learning. *Comput.-Aided Civ. Infrastruct. Eng.* **2024**, *39*, 3438–3460. [CrossRef]
2. Al Shafian, S.; Hu, D. Integrating Machine Learning and Remote Sensing in Disaster Management: A Decadal Review of Post-Disaster Building Damage Assessment. *Buildings* **2024**, *14*, 2344. [CrossRef]
3. Alisjahbana, I.; Li, J.; Strong, B.; Zhang, Y. DeepDamageNet: A Two-Step Deep-Learning Model for Multi-Disaster Building Damage Segmentation and Classification Using Satellite Imagery. *arXiv* **2024**, arXiv:2405.04800.
4. Wu, C.; Zhang, F.; Xia, J.; Xu, Y.; Li, G.; Xie, J.; Du, Z.; Liu, R. Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets. *Remote Sens.* **2021**, *13*, 905. [CrossRef]
5. Melamed, D.; Johnson, C.; Zhao, C.; Blue, R.; Morrone, P.; Hoogs, A.; Clipp, B. xFBD: Focused Building Damage Dataset and Analysis. *arXiv* **2022**, arXiv:2212.13876.
6. Zou, R.; Liu, J.; Pan, H.; Tang, D.; Zhou, R. An Improved Instance Segmentation Method for Fast Assessment of Damaged Buildings Based on Post-Earthquake UAV Images. *Sensors* **2024**, *24*, 4371. [CrossRef] [PubMed]
7. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
9. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9404–9413.
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023.
11. Durnov, V. xView2 1st Place Solution. Available online: https://github.com/vdurnov/xview2_1st_place_solution (accessed on 1 March 2026).
12. Shen, Y.; Zhu, S.; Yang, T.; Chen, C.; Pan, D.; Chen, J.; Xiao, L.; Du, Q. BDANet: Multiscale Convolutional Neural Network with Cross-Directional Attention for Building Damage Assessment from Satellite Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
13. Microsoft. Building Damage Assessment CNN Siamese. Available online: https://github.com/microsoft/building-damage-assessment-cnn-siamese (accessed on 1 March 2026).
14. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A.; Zhang, L. Building Damage Assessment for Rapid Disaster Response with a Deep Object-Based Semantic Change Detection Framework: From Natural Disasters to Man-Made Disasters. *Remote Sens. Environ.* **2021**, *265*, 112636. [CrossRef]
15. Ge, X.; Zhou, L.; Meng, D. DDNet: Disaster Damage Detection for Buildings Based on Dual-Temporal Joint Attention Network. *Sci. Rep.* **2025**, *15*, 42513. [CrossRef] [PubMed]

16. Da, Y.; Ji, Z.; Zhou, Y. Building Damage Assessment Based on Siamese Hierarchical Transformer Framework. *Mathematics* **2022**, *10*, 1898. [CrossRef]
17. Chen, F.; Sun, Y.; Wang, L.; Wang, N.; Zhao, H.; Yu, B. HRTBDA: A Network for Post-Disaster Building Damage Assessment Based on Remote Sensing Images. *Int. J. Digit. Earth* **2024**, *17*, 1. [CrossRef]
18. Kaur, N.; Lee, C.-C.; Mostafavi, A.; Mahdavi-Amiri, A. Large-Scale Building Damage Assessment Using a Novel Hierarchical Transformer Architecture on Satellite Images. *Comput.-Aided Civ. Infrastruct. Eng.* **2023**, *38*, 2072–2091. [CrossRef]
19. Xiao, Y.; Mostafavi, A. DamageCAT: A Deep Learning Transformer Framework for Typology-Based Post-Disaster Building Damage Categorization. *Int. J. Disaster Risk Reduct.* **2025**, *120*, 105704. [CrossRef]
20. Su, J.; Bai, Y.; Wang, X.; Lu, D.; Zhao, B.; Yang, H.; Mas, E.; Koshimura, S. Technical Solution Discussion for Key Challenges of Operational Convolutional Neural Network-Based Building-Damage Assessment from Satellite Imagery: Perspective from Benchmark xBD Dataset. *Remote Sens.* **2020**, *12*, 3808. [CrossRef]
21. Zhao, Z.; Wang, F.; Chen, S.; Wang, H.; Cheng, G. Deep Object Segmentation and Classification Networks for Building Damage Detection Using the xBD Dataset. *Int. J. Digit. Earth* **2024**, *17*, 2302577. [CrossRef]
22. Zhu, R.; Lan, X. Multimodal Building Damage Assessment Method Fusing Adaptive Attention Mechanism and State-Space Modeling. *Sensors* **2026**, *26*, 638. [CrossRef] [PubMed]
23. Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
24. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752.
25. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J. VMamba: Visual State Space Model. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024.
26. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In Proceedings of the International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024.
27. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. ChangeMamba: Remote Sensing Change Detection with Spatiotemporal State Space Model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–20. [CrossRef]
28. Zhang, H.; Chen, K.; Liu, C.; Chen, H.; Zou, Z.; Shi, Z. CDMamba: Incorporating Local Clues into Mamba for Remote Sensing Image Binary Change Detection. *arXiv* **2024**, arXiv:2406.04207.
29. Dong, Z.; Cheng, D.; Li, J. SpectMamba: Remote Sensing Change Detection Network Integrating Frequency and Visual State Space Model. *Expert Syst. Appl.* **2025**, *287*, 127902. [CrossRef]
30. Zhang, J.; Chen, R.; Liu, F.; Liu, H.; Zheng, B.; Hu, C. DC-Mamba: A Novel Network for Enhanced Remote Sensing Change Detection in Difficult Cases. *Remote Sens.* **2024**, *16*, 4186. [CrossRef]
31. Paranjape, J.N.; de Melo, C.; Patel, V.M. A Mamba-Based Siamese Network for Remote Sensing Change Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Tucson, AZ, USA, 28 February–4 March 2025; pp. 1186–1196.
32. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv* **2024**, arXiv:2408.00714.
33. Meta AI. Introducing Meta Segment Anything Model 3. Available online: https://ai.meta.com/blog/segment-anything-model-3/ (accessed on 1 March 2026).
34. Ren, S.; Luzi, F.; Lahrichi, S.; Kassaw, K.; Collins, L.M.; Bradbury, K.; Malof, J.M. Segment Anything, from Space? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 8355–8365.
35. Zhang, J.; Yang, X.; Jiang, R.; Shao, W.; Zhang, L. RSAM-Seg: A SAM-Based Model with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation. *Remote Sens.* **2025**, *17*, 590. [CrossRef]
36. Ji, Y.; Wu, W.; Wan, G.; Zhao, Y.; Wang, W.; Yin, H.; Tian, Z.; Liu, S. Segment Anything Model-Based Building Footprint Extraction for Residential Complex Spatial Assessment Using LiDAR Data and Very High-Resolution Imagery. *Remote Sens.* **2024**, *16*, 2661. [CrossRef]
37. Wu, Z.; Qu, C.; Wang, W.; Miao, Z.; Feng, H. Remote Sensing Extraction of Damaged Buildings in the Shigatse Earthquake, 2025: A Hybrid YOLO-E and SAM2 Approach. *Sensors* **2025**, *25*, 4375. [CrossRef] [PubMed]
38. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
39. Gupta, R.; Hosfelt, R.; Sajeev, S.; Patel, N.; Goodman, B.; Heim, E.; Doshi, J.; Lucas, K.; Choset, H.; Gaston, M. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 10–17.
40. Lee, C.; Kaur, N.; Mahdavi-Amiri, A.; Mostafavi, A. *IDA-BD: Pre- and Post-Disaster High-Resolution Satellite Imagery for Building Damage Assessment from Hurricane Ida*; DesignSafe-CI: Austin, TX, USA, 2022. [CrossRef]

41. Gyegyiri, J.; Su, H.; Thapa, A. *IAN-BD: Pre- and Post-Disaster High-Resolution Aerial Imagery for Building Damage Assessment from Hurricane Ian*; DesignSafe-CI: Austin, TX, USA, 2024. [CrossRef]
42. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
43. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 740–755.