Article

Not peer-reviewed version

Geographic Bias Analysis and Cross-Domain Generalization in Deep Learning-Based Building Damage Assessment

Shruti Kshirsagar , Bharath Chandra , Unaza Tallal * , Rajiv Bagai , Atri Dutta

Posted Date: 26 March 2026

doi: 10.20944/preprints202603.2114.v1

Keywords: building damage assessment; satellite imagery; artificial intelligence; deep learning; geographic bias; domain adaptation; data augmentation; cross-domain generalization; disaster response; remote sensing; transfer learning; computer vision

 doi:10.20944/preprints202603.2114.v1

Article

Geographic Bias Analysis and Cross-Domain Generalization in Deep Learning-Based Building Damage Assessment

Shruti Kshirsagar [1,*], Bharath Chandra [1], Unaza Tallal [1,*], Rajiv Bagai [1] and Atri Dutta [2]

[1] School of Computing, Wichita State University, Wichita, Kansas, USA
[2] Department of Aerospace Engineering, Wichita State University, Wichita, Kansas, USA
* Corrspondence: uxtallal@shockers.wichita.edu (U.T.); shruti.kshirsagar@wichita.edu (S.K.)

Abstract

Automated building damage assessment from satellite imagery has become increasingly critical for rapid disaster response and humanitarian relief operations. However, current state-of-the-art deep learning models exhibit significant generalization challenges when deployed to geographically and environmentally diverse regions. This study investigates the nature and extent of geographic bias in building damage detection systems, revealing that model performance degradation stems primarily from geographic and structural characteristics rather than insufficient training data representation. Through systematic evaluation of top-performing xView2 competition solutions across 17 disaster locations spanning multiple climate zones, we found that even state-of-the-art models struggle with generalization, particularly for Minor and Major damage classes, and exhibit strong geographic biases toward certain regions. Strikingly, Nepal despite having the largest training dataset (15,234 images) achieves the worst performance, demonstrating that geographic and structural characteristics dominate generalization behavior more than training data quantity. To address these fundamental limitations, we explore Fusion Augmentation, a novel methodology that enhances edge detection and structural feature representation by integrating auxiliary information channels with standard RGB imagery. Experimental results demonstrate substantial improvements of 7.1% overall F1 score, with dramatic gains for intermediate damage categories such as Minor and Major damage. Domain adaptation experiments on three unseen locations show that combining Fusion Augmentation with supervised fine-tuning yields 40.8% and 60.0% improvements over Minor and major classes, while unsupervised CORAL achieves 24.2% and 39.5% improvements over Minor and major damage classes compared to benchmarks. These findings challenge prevailing assumptions about data-driven generalization in remote sensing AI systems and demonstrate that structural feature enhancement combined with domain adaptation is essential for robust detection across geographically diverse deployment scenarios, providing practical strategies for globally deployable damage assessment systems.

Keywords: building damage assessment; satellite imagery; artificial intelligence; deep learning; geographic bias; domain adaptation; data augmentation; cross-domain generalization; disaster response; remote sensing; transfer learning; computer vision

1. Introduction

Climate change is increasing the frequency and intensity of natural disasters such as hurricanes, earthquakes, floods, and wildfires, impacting worldwide communities [1]. Natural disasters globally displaced 26.4 million people in 2023, causing economic damages of roughly $380 billion [2]. Emergency action following such events is crucial, as timely and precise building damage assessment can save trapped survivors. Due to infrastructure damage, safety hazards, and the scale of affected areas, structural engineers' manual field inspections are generally insufficient for damage assessment. These constraints necessitate automated, scalable, and reliable post-disaster building damage assessment.

 doi:10.20944/preprints202603.2114.v1

Satellite and airborne photos provide speedy, wide-area coverage of disaster-affected areas without endangering assessment crews [3]. Within hours of a tragedy, modern Earth observation satellites may acquire high-resolution photos of structural damage across cities or regions. Manual interpretation of this footage is challenging and time-consuming, especially when studying hundreds of thousands of buildings in large catastrophe zones. This bottleneck slows emergency response decision-making, which is crucial for search-and-rescue and resource allocation.

Recent advancements in deep learning (DL), especially CNNs, have enabled automated building damage detection from satellite images [4]. These models can evaluate massive visual data and classify damage severity with accuracy approaching or exceeding human specialists. The xView2 Challenge, organized by the Defense Innovation Unit (DIU), has advanced building damage assessment by providing the largest publicly available dataset of annotated satellite imagery, spanning 19 disasters and 850,000 labeled buildings [5]. Top-performing solutions from this competition achieved F1 scores above 0.85 for building location and damage classification, proving automated damage assessment systems are technically feasible. Despite these impressive benchmark results, artificial intelligence systems for building damage assessment face critical challenges when deployed across geographically and environmentally diverse regions. Models trained on specific disasters often fail to generalize due to domain shift systematic differences in data distributions caused by varying architectural styles, environmental conditions, and disaster types. Recent operational deployments by humanitarian organizations have revealed that models achieving 90% accuracy in North American contexts may perform at only 30-40% accuracy in South Asian or tropical environments, highlighting fundamental AI generalization limitations.

Our work addresses this critical gap by investigating how structural feature enhancement and domain adaptation can enable AI systems to generalize robustly across diverse geographic deployment scenarios. Despite these impressive benchmark results, a critical gap remains between laboratory performance and real-world deployment: model generalizability. Deep learning models trained on specific disaster scenarios often exhibit significant performance degradation when applied to new geographic regions, different disaster types, or unfamiliar architectural contexts [5–7]. Geographic distribution of training data, class imbalance favoring undamaged structures, and climate zone variations limit this. The consequences are severe: models trained predominantly on North American hurricanes and wildfires have shown degraded performance when deployed to earthquake-affected regions in South Asia, as evidenced in recent operational use by the United Nations and World Bank [8]. A model that performs well in temperate regions but fails in tropical environments could lead to underestimation of damage, delayed humanitarian aid, and ultimately, preventable loss of life. The performance of Building damage detection has improved; however, some important gaps limit its practicality. These gaps are addressed by detailed empirical analysis and an integrated methodology integrating fresh data augmentation with domain adaptation in this paper. While significant progress has been made in building damage detection, several critical gaps limit real-world applicability. In this paper, we address these gaps through comprehensive empirical analysis and propose an integrated methodology combining novel data augmentation with domain adaptation. Our main contributions are:

1. Geographic Bias Analysis: We conduct extensive analysis across 17 locations revealing geographic and structural bias beyond representational imbalance. Models exhibit systematic failures in tropical/volcanic/mountainous regions despite adequate training samples Nepal performs poorly despite having the most training data, demonstrating that geographic characteristics drive generalization failures.
2. Integrated Augmentation-Adaptation Framework: We systematically evaluate supervised fine-tuning and unsupervised CORAL domain adaptation combined with Fusion Augmentation across three out-of-domain locations (Texas, Ayiti, Portugal).

The remainder of this paper is organized as follows: Section 2 reviews related work in building damage classification, data augmentation, domain adaptation, and model bias. Section 3 discusses methods

 doi:10.20944/preprints202603.2114.v1

and material including the dataset, distribution, our approach, Fusion Augmentation, and domain adaptation methodologies, experimental setup, and evaluation protocols. Section 4 presents in-domain and out-of-domain test results, geographic bias patterns. Section 5 concludes with an overview of contributions and their potential impact on operational disaster response systems.

2. Related Work

This section reviews recent advancements in deep learning models for building damage detection, focusing on damage classification approaches, model biases, data augmentation strategies, and domain adaptation techniques.

2.1. State-of-the-Art Models for Building Damage Detection

Deep learning and high-resolution Earth observation data have advanced satellite-based building damage classification. Satellite imagery is now widely used by the International Charter and Copernicus Emergency Management Service for disaster damage evaluation and reference maps [9]. The use of Convolutional Neural Networks (CNNs) for object segmentation [10] led to robust capabilities in automated building damage detection [11]. Two main methods for assessing building damage are bi-temporal (analyzing pre- and post-disaster imagery)[11,12] and single-image (assessing damage using only post-disaster imagery) [13,14]. However, bi-temporal techniques require temporally aligned image pairs to measure changes and give more thorough damage assessment information. Single-image methods are faster and useful for sudden, unexpected disasters without pre-disaster imagery. After the xBD dataset and xView2 Challenge, building damage detection models have quickly evolved to focus on architectural improvements, class imbalance handling, and cross-domain generalization. Kaur et al. [15] developed DAHiTrA, a transformer-based approach that maps pre- and post-disaster images into a common feature space using hierarchical spatial features, capturing long-range dependencies crucial for damage assessment. The Ordinal Class Distance Penalty Loss (OCDPL) by Tsai and Lin [16] addresses class imbalance by penalizing misclassifications based on the difference between predicted and actual damage levels. This approach utilizes damage category structure to enhance minority class performance. According to Wiguna et al. [7], a semi-supervised framework using pseudo-labeling and iterative fine-tuning improved accuracy by 21% on the Noto Peninsula Earthquake dataset, highlighting the potential of using abundant unlabeled disaster data. Jakubik et al. [17] developed foundation models for Earth monitoring, using large-scale pre-training and unsupervised domain adaptation for transferable representations that adapt to new locations and hazards with minimum fine-tuning.

2.2. Bias and Generalization Challenges

Despite impressive performance on benchmark datasets, deep learning models for building damage detection face significant challenges related to bias and limited generalization. Bias refers to systematic errors causing models to perform inconsistently across different datasets or deployment scenarios, particularly when training data fails to represent the full variety of real-world conditions [18]. Many types of bias have been recognized [19]: Class imbalance bias is caused by the xBD dataset's imbalance, where undamaged buildings outnumber damaged ones, leading to model bias towards the majority class [5,7,11,15]. Geographic bias occurs because training data is concentrated in North America, causing models to encode region-specific building materials, architectural styles, and landscape attributes. Melamed et al. [18] found that damaged structures are spatially grouped, with building damage levels substantially associated with surrounding buildings. The top-5 xView2 solutions perform well in highly damaged locations but poorly in isolated damaged structures, which could lead to important oversights and delayed emergency assistance. Real-world applicability is further limited by sensor and image quality bias from satellite sensors, atmospheric conditions, and acquisition parameters. These biases drive our research on data augmentation and domain adaptation to improve generalization.

 doi:10.20944/preprints202603.2114.v1

2.3. Data Augmentation Techniques

Data augmentation is essential for deep learning, especially in labeled limited data areas like building damage detection. Research from various fields shows that effective augmentation must be task-specific, preserve crucial traits, and offer meaningful variability. Augmentation in computer vision relies on traditional geometric transformations such as flips, rotations, scale adjustments, and color jittering [20,21]. Adedeji et al. [20] systematically evaluated these techniques for satellite image classification, noting that while they improve robustness to orientation and perspective changes, their impact is somewhat limited for satellite imagery due to inherent uniformity. DCGAN-based generative techniques for satellite image synthesis are more effective at creating different instances, but require more computing resources [20]. Advanced techniques include fusion-based augmentation where auxiliary information channels combine with RGB data to enrich representations [22], RICAP which patches multiple images together [23], the Albumentations library offering efficient transformations [24], and Cutout which masks regions to force broader context utilization [25]. Critically, research demonstrates that effective augmentation must be quality-aware and task-specific: strategies should adapt to input quality characteristics [26], work synergistically with representation learning [27], and most importantly, preserve task-relevant features while introducing meaningful variability [28]. For building damage assessment, this implies retaining structural borders and boundaries while adding environmental resilience. Advances in building damage identification have not addressed structural edge preservation and enhancement, which are vital for distinguishing damage levels, especially in "minor" and "major" categories.citeparupati2025towards.

2.4. Domain Adaptation Techniques

Domain adaptation transfer knowledge from source domains (training data) to target domains (deployment scenarios) with different data distributions. Supervised approaches utilize labeled target samples for fine-tuning, while unsupervised methods align representations without labels [29,30]. Jakubik et al. [17] used unsupervised domain adaptation, combining pre-training on related tasks and zero-shot adaptation to new hazard types, enhancing generalizability and reducing labeled data needs. UniDA allows adaptation without target label set knowledge [29], while source-free adaptation works when source data is unavailable [30]. Successful adaptation requires both feature distribution alignment and preservation of task-relevant characteristics, with effectiveness enhanced when combined with task-specific augmentation under significant domain shift [26,27] Hu and Tang [30] found that employing CycleGAN and adversarial training for transfer learning and domain adaptation enhances cross-disaster generalization for building damage assessment across 14 events. Recent research by Parupati et al. [31] suggests that combining augmentation and adaptation leads to better generalization than either strategy alone. In line with this, Mouradi et al. [32] investigated cross-disaster building damage detection using a two-stage ensemble framework with supervised domain adaptation. The results showed that supervised domain adaptation was essential for robust four-class damage classification in unseen geographic regions.

3. Materials and Methods

This section describes the dataset, preprocessing procedures, fusion augmentation approach, domain adaptation techniques, baseline models, and evaluation metrics.

Figure 1 presents the overall pipeline of the proposed framework for building damage detection and geographic bias analysis. First, the input dataset is preprocessed. Next, fusion-based augmentation is applied to increase the number of training examples. Finally, either supervised or unsupervised domain adaptation methods are applied, as described in the following sections.

Figure 1. End-to-end pipeline of the proposed framework building damage detection.

 doi:10.20944/preprints202603.2114.v1

3.1. Dataset

For building damage assessment the xBD (xView2 Building Damage) dataset [5] is the largest publicly available dataset for from satellite imagery. The dataset consisted of paired pre- and post-disaster images collected by Maxar/DigitalGlobe satellites at a spatial resolution of 0.3-0.8 meters. The dataset covered around 45,362 square kilometers, providing more than 850,000 annotated building polygons, collected from 19 disaster events across five continents. A wide range of hazard types are included in xBD dataset such as severe wildfires (Santa Rosa, Woolsey, Pinery, SoCal, Portugal), major hurricanes (Michael, Harvey, Matthew, Florence), destructive tornadoes (Moore, Joplin), volcanic eruptions (Guatemala, Lower Puna), earthquakes (Nepal, Mexico), floods (Midwest), and the Palu tsunami. These events takes place across diverse climate zones, including tropical, volcanic, arid, temperate and mountainous regions, making xBD one of the most diverse datasets available. Each building instance is labeled with a polygon footprint and a damage level is assigned according to four-tier Joint Damage Scale, which includes four categories: No Damage, Minor Damage, Major Damage, and Destroyed. The dataset also offers global disaster maps of affected locations and illustrative pre- and post-disaster image pairs.

Table 1 outlines four damage classes presented in xBD dataset, ranging from undamaged building to completely destroyed structures. These classes describe the level of damage that has occurred on each building from no structural issues to full collapse or destroyed. The xBD dataset has a severe damage class distribution data imbalance, a typical pattern commonly seen in damage-related datasets. Table 2 presents the frequency of each category across the 19 disaster events. Out of a total of 411,357 labeled buildings, 76% falls in No Damage class, whereas, Minor Damage, Major Damage and Destroyed makes up only 7-9 % of the whole dataset. This skewed distribution underscores the class imbalance challenge when training models on the xBD dataset.

Table 1. Scale of Damage Descriptions [5].

Damage Class	Description
No Damage	Undisturbed. No sign of water, structural or single damage, or burn marks.
Minor Damage	Building partially burnt, water surrounding structure, roof elements missing or visible cracks.
Major Damage	Partial wall or roof collapse, encroaching, surrounded by mud/water.
Destroyed	Completely collapsed, partially or completely covered with water/mud or no longer present.

Table 2. Class Distribution in xBD Dataset.

Damage Class	Count	Percentage
No Damage	313,033	76.2%
Minor Damage	36,860	9.0%
Major Damage	29,904	7.3%
Destroyed	31,560	7.5%
Total	411,357	100%

Furthermore, the distribution of samples varies significantly across disaster types and geographic regions, as illustrated in Figure 2. About 45% of the training set is dominated by North American events, including Hurricanes Michael, Harvey, Matthew, and Florence, as well as the Santa Rosa and Woolsey wildfires, which highlights the dataset's geographic bias. In contrast, the events from South Asian, African, and South American regions are noticeably underrepresented. This geographic skew introduces inherent bias, which is the central focus of our investigation in this study.

 doi:10.20944/preprints202603.2114.v1

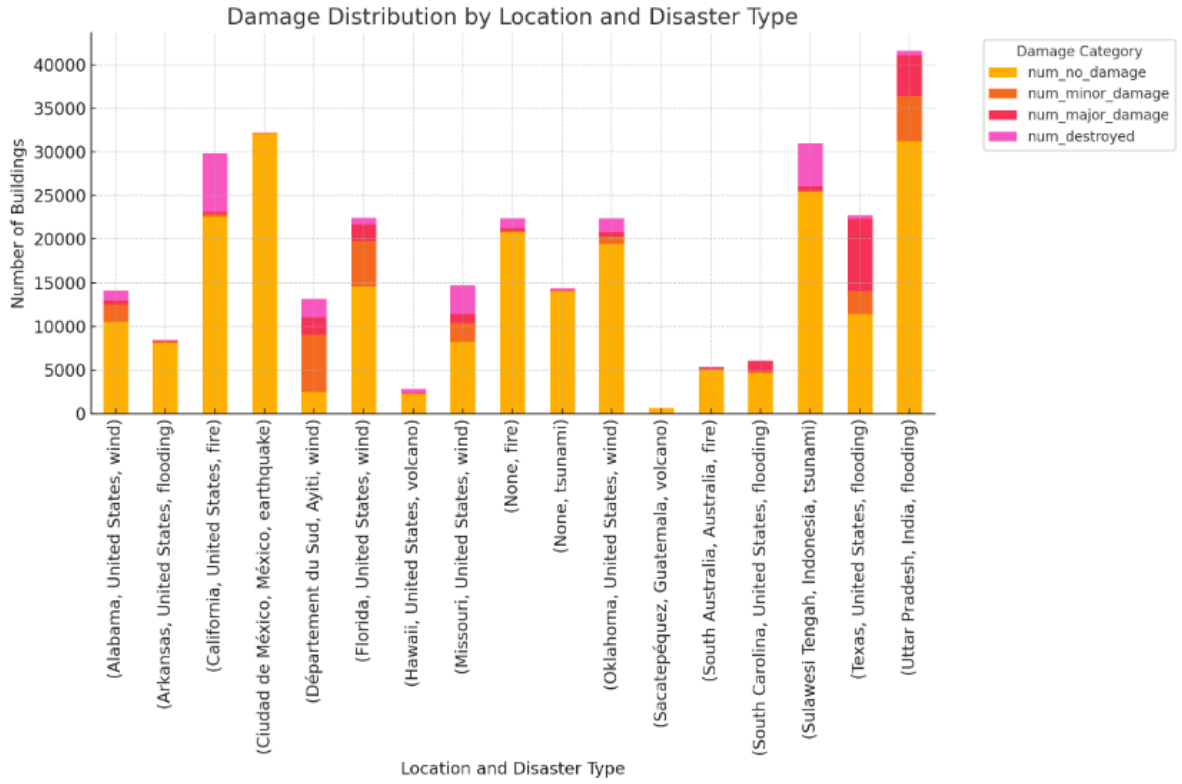

Figure 2. Number of training samples for each location [5].

3.2. Image Preprocessing

The xBD dataset includes building damage annotations in GeoJSON format, that contain polygon coordinates for each structure, with metadata including damage level, disaster type, and image identifiers. We converted these annotations to pixel-wise masks for model training using the following pipeline:

3.2.1. Polygon Extraction and Mask Generation

First, we parsed the GeoJSON files to extract the polygon coordinates (x,y) where each point represents a vertex of a building. We then converted these geographic coordinates into pixel coordinates using the geotransformation information of each image. Next, we generated masks for each building polygon using OpenCV's fillPoly function. For building localization, pixels belonging to buildings were labeled as 1 and background pixels as 0. For damage classification, labels 0,1,2,3,4 represent background, no damage, minor, major, and destroyed, respectively. When multiple buildings overlapped at the same pixel, we assigned the label with the highest damage level. This ensured that regions with severe damage were properly captured. Figure 3 shows the full preprocessing pipeline from raw GeoJSON data to final masks.

(a) Pre-disaster image (b) Pre-disaster mask (c) Post-disaster image (d) Post-disaster mask

Figure 3. Pre-disaster and post-disaster images of the Santa Rosa wildfire with their corresponding masks [5].

 doi:10.20944/preprints202603.2114.v1

3.2.2. Image Normalization

All imagery from satellites are normalized using ImageNet statistics [33] to utilize transfer learning from pre-trained encoders.

$$I_{norm} = \frac{I - \mu}{\sigma} \tag{1}$$

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ represent the mean and standard deviation for RGB channels. All images are downsized to 1024x1024 pixels, with zero-padding for smaller ones. To process images larger than 1024×1024, we adopt a sliding window technique with 50% overlap and aggregate predictions using majority voting during inference.

3.3. Fusion Augmentation

Post-disaster satellite imagery presents significant challenges for AI-based damage detection: structural edges are obscured by debris, contrast is reduced by atmospheric conditions, and fine-grained damage features are masked by environmental factors. Our Fusion Augmentation approach addresses these challenges by explicitly encoding structural information through auxiliary feature channels that enhance the model's ability to detect damage indicators across diverse environmental conditions. Fusion-based data augmentation is designed to enhance model performance by integrating auxiliary information channels with standard RGB inputs [22]. Unlike instance-based augmentation–which modifies the original image through transformation techniques such as rotation, flipping, or noise addition–fusion augmentation enriches the input by adding additional spectral or spatial information. Instead of increasing the number of training samples, this technique increases the information dimensionality available to the model, thereby improving its ability to capture complex patterns. In addition, buildings in vegetation-dense or visually complex environments may be partially occluded or surrounded by debris, making structural edges difficult to detect from RGB imagery alone. Conventional augmentation may even degrade these crucial features by introducing noise or reducing edge sharpness through excessive color jittering. Fusion Augmentation, addresses these challenges by enriching the input representation with auxiliary channels that explicitly encode structural information relevant to damage detection. To achieve this, we incorporate several feature enhancing transformations. Canny edge detection is applied to extract building boundaries and highlight structural discontinuities commonly associated with damage. Histogram equalization is used to improve contrast, making intensity variations that signal structural deterioration more distinct. Furthermore, unsharp masking sharpens local details, revealing fine-grained structural features such as cracks, missing roof components, and other damage indicators.

3.3.1. Canny Edge Detection

Structural edges are critical for AI-based damage detection, as damage manifests through geometric discontinuities. Edges play a crucial role in characterizing building layout, especially in post-disaster images where environmental factors may obscure these features [34]. In this study, we employ canny detector [35] due to its robustness and proven effectiveness in extracting well-defined structural boundaries. The algorithm consists of the following stages:

1. Noise Reduction: A Gaussian filter is applied to supress high-frequency noise whiles preserving key structural edges intact. The Gaussian kernel is defined as:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{2}$$

where we use a 5×5 Gaussian kernel with $\sigma = 1.4$.

2. Gradient Calculation: Image gradients G_x and G_y are computed using sobel operators. The gradient magnitude and orientation are then derived as:

$$\text{Edge\_Gradient}\,(G) = \sqrt{G_x^2 + G_y^2} \tag{3}$$

 doi:10.20944/preprints202603.2114.v1

$$\text{Angle}\,(\theta) = \tan^{-1}\left(\frac{G_y}{G_x}\right) \tag{4}$$

3. Non-Maximum Suppression: To produce thin, well-localized edges, non-maximum suppression retains only pixels whose gradient magnitude is a local maximum along the gradient direction.

4. Double Thresholding: Edge pixels are classified using high (T_{high}) and low (T_{low}) thresholds. We set $T_{high} = 100$ and $T_{low} = 50$ based on validation experiments.

5. Edge Tracking by Hysteresis: Weak edges that are connected to the strong edges are retained; whereas isolated weak edges are supressed. This steps remove noise while guarantees edge continuity.

The resulting edge map $E(x, y) \in \{0, 1\}$ highlights building boundries, structural discontinuities, and damage-related features. By integrating these edge maps into the network enables early emphasis on geometric cues that are essential for identifying building damage.

3.3.2. Contrast Enhancement

Contrast enhancement is applied to improves the visibility of subtle damage features by redistributing pixel intensities across the full dynamic range. In this study, we apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [36], which performs contrast enhancement locally with small image tiles set 8×8 pixels– while limiting the amplification of histogram bins through a clip limit of 2.0. This prevents the excessive noise enhancement that often occurs using global histogram equalization. CLAHE is applied independently to each RGB channel, the redistribution of histogram values is controlled by the following formulation:

$$H'(i) = \min\left(H(i), \frac{N_{pixels}}{N_{bins}} \times \text{clip\_limit}\right) \tag{5}$$

where $H(i)$ is the histogram count for bin (i). The resultant enhanced image $I_{enhanced}$ reveals finer texture variations, weathering patterns, and subtle structural differences that help distinguish visually altered but structurally intact surfaces such as soot or ash deposits from genuine damage.

3.3.3. Unsharp Masking

Unsharp masking [37] is applied to emphasize fine structural details by sharpening the image through the combination of original and a blurred version. The sharpened output is computed as:

$$I_{sharp} = I + \alpha \cdot (I - G_\sigma * I) \tag{6}$$

where $G_\sigma * I$ denotes the Gaussian-blurred image using 5×5 kernel with $\sigma = 1.0$, and the sharpening strength is set to $\alpha = 1.5$. This enhancement improves the visibility of subtle structural cues, such as small cracks, and missing roof elements, while avoiding the artifacts that conventional sharpening filters may introduce.

3.3.4. Channel Fusion Strategy

After generating the auxiliary representations of the edge map, contrast-enhanced image, and unsharp-masked image. We convert each auxiliary channel to a single-channel grayscale format and concentrate on the original RGB image. The fused input is constructed as

$$I_{fused} = [I_{RGB}, E, C, S] \tag{7}$$

resulting in a six-channel representation consisted of three RGB bands and three structural-enhancement channels. This fused repesentation provides richer information, allows the model to detect subtle structural and damage related features that may not be clearly visible in RGB imagery alone. We construct a six channel input consisted of three RGB bands and three greyscale auxiliary channels (contrast, edge and sharpness). Using single-channel auxiliary features keeps the representation compact while retaining the structural information required for reliable damage detection. The

 doi:10.20944/preprints202603.2114.v1

proposed design is computationally more efficient than a full 10-channel RGB-based fusion and as, we observed empirically, attained comparable performance. Therefore, the six channel configuration offers an effective balance between representational richness and computational cost.

3.4. Domain Adaptation Techniques

Domain adaptation addresses the challenge of transferring knowledge learned from a source domain to a target domain with different underlying data distribution [38]. This capability is especially important in building damage assessment, where models trained on specific disasters and geographic regions must generalize to new and unseen scenarios. In this study, we investigate two complementary strategies: supervised fine-tuning, which depends on small amount of labeled target data and CORAL [39], an unsupervised feature alignment approach.

3.4.1. Supervised Fine-Tuning

Supervised fine-tuning adapts pre-trained models to target domains using a limited number of labeled samples [40]. In this approach, the early encode layers remain frozen to preserve low-level, transferable features such as edges and textures, while the deepr encode layers and decoders are updated to learn domain-specific building characteristics and damage patterns. This selective updating allows the model to retain the generalizable representations while effectively adapting to the structural and environmental variations present in the target domain. The fine-tuning process follows a standard training loop. Let M_θ be the pre-trained model and $D_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ the labeled target dataset. The early encoder layers are kept fixed, while the final encoder blocks and the decoder are unfrozen, and the model is optimized using AdamW. For each training batch (X_b, Y_b), predictions $\hat{Y}_b = M_\theta(X_b)$ are computed, and the parameters of the unfrozen layers are updated to minimize the composite loss

$$L_{\text{total}} = \lambda_1 L_{\text{Dice}} + \lambda_2 L_{\text{Focal}} + \lambda_3 L_{\text{CE}}, \tag{8}$$

with weights $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.2$, following the top-performing xView2 formulation [41]. Through this procedure, the model gradually aligns its higher-level representations with the target distribution while preserving useful pre-trained structure.

3.4.2. Unsupervised CORAL Domain Adaptation

Unsupervised domain adaptation is performed using CORrelation ALignment (CORAL) [42], which aligns second-order statistics between the source and target domains by matching their feature covariance matrices. CORAL minimizes the distributional discrepancy through the covariance-matching objective

$$L_{\text{CORAL}} = \frac{1}{4d^2} \|C_S - C_T\|_F^2, \tag{9}$$

where C_S and C_T denote the covariance matrices of source and target features, respectively, and $\|\cdot\|_F$ is the Frobenius norm. The covariance matrices are computed as

$$C_S = \frac{1}{n_s - 1} \left(D_S^\top D_S - \frac{1}{n_s} (1^\top D_S)^\top (1^\top D_S) \right), \tag{10}$$

$$C_T = \frac{1}{n_t - 1} \left(D_T^\top D_T - \frac{1}{n_t} (1^\top D_T)^\top (1^\top D_T) \right), \tag{11}$$

where $D_S = \{f_i^s\}_{i=1}^{n_s}$ and $D_T = \{f_j^t\}_{j=1}^{n_t}$ are source and target feature sets, respectively. To improve adaptation, CORAL loss is applied at multiple network layers, and the total multi-layer objective is

$$L_{\text{CORAL}}^{\text{total}} = \sum_{i=1}^{t} \lambda_i L_{\text{CORAL}}^{(i)}. \tag{12}$$

 doi:10.20944/preprints202603.2114.v1

3.5. Benchmark Systems

In this paper, we used three xView2 competition winners as benchmark systems. The first-place xView2 solution [41] employs a combination of UNet-like architectures for building localization and Siamese neural networks for damage classification. Localization models are trained on pre-disaster images using several encoder backbones (e.g., ResNet34, SE-ResNeXt50, SENet154), which are then converted into Siamese networks for analyzing paired pre- and post-disaster inputs. The approach incorporates extensive data augmentation and a composite loss consisting of Dice, Focal, and Cross-Entropy terms.

The second-place solution [43] adopts UNet-like networks with pretrained DPN92 and DenseNet161 encoders for binary localization. Mixed-precision training (Apex) is used to improve computational efficiency, and localization and classification stages are optimized separately. Multiclass damage classification is performed using a Focal-with-Dice loss.

The third-place system [44] ensembles multiple semantic segmentation models trained with weighted cross-entropy to mitigate class imbalance. A shared encoder processes pre- and post-disaster imagery, whose extracted features are concatenated before entering the decoder. The ensemble spans several encoder families, including ResNets, DenseNets, and EfficientNets paired with UNet or FPN decoders, and integrates pseudo-labeling and weighted model averaging to enhance robustness. All three benchmark systems report F1 scores above 0.85 on the official xView2 test set, reflecting the effectiveness of ensemble modeling and advanced training strategies. However, as demonstrated in our experiments, these approaches degrade substantially when evaluated on geographically out-of-distribution regions, motivating the development of our Fusion Augmentation and domain adaptation methods.

3.6. Evaluation Metrics

Model performance is assessed using metrics that capture both localization accuracy and damage classification quality, consistent with the official xView2 evaluation protocol.

F1 Score. The primary metric is the F1 score [45], defined as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{13}$$

where

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}. \tag{14}$$

We compute F1 scores separately for localization and for each damage category (no damage, minor, major, and destroyed). Following the xView2 protocol, overall damage-classification performance is reported as the harmonic mean of the class-wise F1 scores:

$$F1_{\text{damage}} = \frac{4}{\sum_{c \in C} \frac{1}{F1_c}}, \tag{15}$$

where $C = \{\text{no-damage, minor, major, destroyed}\}$ ensuring that all categories contribute equally despite dataset imbalance.

4. Results and Discussion

This section presents the experimental results and discusses the findings in the context of existing literature.

4.1. Baseline Performance

The top-3 winning solutions from the xView2 challenge [41,43,44] represent the current state-of-the-art in automated building damage assessment from satellite imagery. Table 3 presents the performance of these models on the official competition test set, demonstrating an overall capabilities with localization F1 scores exceeding 0.84 and particularly high accuracy on the "No Damage" class

 doi:10.20944/preprints202603.2114.v1

(F1 > 0.90). However, these models exhibit substantial performance degradation on two damage categories, such as minor damage and major damage. This pattern aligns with findings from the original xBD dataset paper [5].

Table 3. Performance of Top-3 Winning Solutions on the xView2 Challenge Test Set [46].

Class	First-place	Second-place	Third-place
Localization	0.8621	0.8532	0.8465
No Damage	0.9147	0.9018	0.9071
Minor Damage	0.6385	0.6181	0.6173
Major Damage	0.7819	0.7702	0.7651
Destroyed	0.8542	0.8486	0.8462

4.2. Geographic Bias

To investigate potential geographic biases, we constructed 17 location-specific test sets by withholding 10% of images from each disaster location in the xBD dataset. This approach allow systematic evaluation of model generalization across diverse geographic contexts, disaster types, and environmental conditions. Tables 4–6 present comprehensive performance results for the top-3 winning solutions across all 17 xBD disaster locations. Performance stratification shows three levels. In temperate US regions like Arkansas, Alabama, Florida, Missouri, Oklahoma, South Carolina, and Texas, models achieve localization F1 values above 0.87 and perform well across all damage categories. Several things contributed to this achievement. Most of the training data in the xBD dataset comes from these locations, accounting for about 60% [5]. Second, they have relatively flat terrain, sparse-to-moderate vegetation, consistent building materials (mostly wood-frame construction with asphalt roofing), and similar architectural styles, creating a homogeneous feature space that models can reliably learn. Third, hurricanes, floods, and tornadoes in these regions provide damage patterns well-represented in the training corpus. This confirms Weber and Kané's findings [47] that training data similarity greatly impacts model performance in remote sensing applications.

Table 4. Location-wise performance of the top-1 winning solution across 17 xBD disaster locations. Per-class F1 scores (Localization, No-Damage, Minor, Major, Destroyed) reveal significant geographic variability, forming the basis for subsequent geographic bias analysis [41] on Each Location from the xBD Dataset.

Location	Disaster Type	Localization	No-Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8615	0.8962	0.7669	0.8882	0.0370
Alabama	Wind	0.8801	0.9448	0.7485	0.8796	0.9125
California	Fire	0.8689	0.8548	0.5572	0.2965	0.8934
Mexico	Earthquake	0.8769	0.9469	0.0000	0.127	0.0000
Florida	Wind	0.9112	0.9303	0.7444	0.8796	0.7668
Hawaii	Volcano	0.8790	0.8694	0.1202	0.0000	0.8161
Missouri	Wind	0.9405	0.9694	0.7170	0.8571	0.9358
Nepal	Flood	0.8375	0.8856	0.3036	0.8692	0.2060
Oklahoma	Wind	0.9416	0.8772	0.8054	0.8435	0.8983
Portugal	Fire	0.8391	0.8827	0.4691	0.3509	0.6328
South Australia	Fire	0.8357	0.8486	0.4257	0.0678	0.7188
South Carolina	Flood	0.8757	0.8667	0.6696	0.8270	0.1086
Indonesia	Tsunami	0.8741	0.9408	0.0000	0.6483	0.8649
Texas	Flood	0.9004	0.9246	0.7087	0.7518	0.297
Guatemala	Volcano	0.9089	0.8800	0.0000	0.0000	0.1264
Sunda	Tsunami	0.8653	0.8334	0.0000	0.0000	0.0000
Ayiti	Wind	0.8628	0.9380	0.5150	0.3694	0.5378

In contrast to temperate regions, Mediterranean and tropical regions (Hawaii, Portugal, South Australia, Indonesia, Ayiti) show modest performance degradation, with F1 scores typically 10-15% lower.

 doi:10.20944/preprints202603.2114.v1

The training distribution does not account for environmental variability in these places. For instance, Hawaii's volcanic landscapes feature dense tropical vegetation and unique geological formations that obscure building boundaries. The Mediterranean climate of Portugal affects sunlight, vegetation (eucalyptus forests), and construction materials (stone and concrete) differently than North American training data. High humidity, dense vegetation, and informal settlement patterns combine with non-Western architecture styles in Indonesia's tropical climate. Environmental variability leads to domain shift [48], where visual characteristics from temperate training data may not transfer well to climatically diverse test conditions. Sublime and Kalinicheva [49] found that vegetation density greatly affects building detection accuracy in satellite imagery.

Table 5. Performance of 2nd Winning Solution [43] on Each Location from the xBD Dataset.

Location	Disaster Type	Localization	No-Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8285	0.8607	0.749	0.8665	0.0362
Alabama	Wind	0.8621	0.9221	0.7255	0.8581	0.8782
California	Fire	0.8453	0.8231	0.5454	0.2877	0.8651
Mexico	Earthquake	0.8465	0.9237	0.0	0.1234	0.0
Florida	Wind	0.8878	0.9073	0.7188	0.8567	0.7365
Hawaii	Volcano	0.8555	0.8426	0.1165	0.0	0.7877
Missouri	Wind	0.9152	0.9446	0.6923	0.8342	0.9015
Nepal	Flood	0.8089	0.8564	0.2943	0.8446	0.1994
Oklahoma	Wind	0.9169	0.8521	0.7788	0.8212	0.8663
Portugal	Fire	0.8143	0.8547	0.4549	0.3406	0.6121
South Australia	Fire	0.8113	0.8225	0.4126	0.0658	0.6954
South Carolina	Flood	0.8501	0.8594	0.6487	0.8036	0.1052
Indonesia	Tsunami	0.8489	0.9163	0.0	0.6292	0.8357
Texas	Flood	0.8756	0.8988	0.6866	0.7307	0.2877
Guatemala	Volcano	0.8833	0.8533	0.0	0.0	0.1223
Sunda	Tsunami	0.8401	0.8076	0.0	0.0	0.0
Ayiti	Wind	0.8383	0.9122	0.4989	0.3585	0.5204

Table 6. Performance of 3rd Winning Solution [44] on Each Location from the xBD Dataset.

Location	Disaster Type	Localization	No-Damage	Minor	Major	Destroyed
Arkansas	Flood	0.8124	0.8412	0.7312	0.8451	0.0355
Alabama	Wind	0.8489	0.9087	0.7089	0.8396	0.8564
California	Fire	0.8312	0.8089	0.5334	0.2811	0.8445
Mexico	Earthquake	0.8321	0.9102	0.0	0.1205	0.0
Florida	Wind	0.8743	0.8934	0.7021	0.8378	0.7201
Hawaii	Volcano	0.8412	0.8289	0.1138	0.0	0.7689
Missouri	Wind	0.9023	0.9312	0.6767	0.8156	0.8801
Nepal	Flood	0.7956	0.8421	0.2878	0.8267	0.1945
Oklahoma	Wind	0.9034	0.8389	0.7612	0.8045	0.8478
Portugal	Fire	0.8012	0.8401	0.4445	0.3329	0.5978
South Australia	Fire	0.7978	0.8089	0.4034	0.0643	0.6789
South Carolina	Flood	0.8367	0.8445	0.6334	0.7856	0.1028
Indonesia	Tsunami	0.8334	0.9012	0.0	0.6156	0.8167
Texas	Flood	0.8623	0.8834	0.6712	0.7145	0.2811
Guatemala	Volcano	0.8644	0.8540	0.0	0.0	0.1222
Sunda	Tsunami	0.8213	0.7978	0.0	0.0	0.0
Ayiti	Wind	0.8161	0.8989	0.4876	0.3543	0.5108

Low performing regions, such as Nepal, had the worst performance (Minor Damage F1 = 0.304, Major Damage F1 = 0.870) despite having the most training samples (Table 4). This interesting finding challenges the idea that model generalization failures are primarily caused by inadequate training data representation in disaster assessment literature [5,50]. Gupta et al. [5] proposed that

 doi:10.20944/preprints202603.2114.v1

diversified catastrophe data would enhance model robustness, assuming representational bias as the main limitation. Our Nepal data contradict this claim by showing geographic and structural bias. Figure 6 visualizes average damage F1 across 17 locations, revealing substantial geographic variability and motivating the need for domain adaptation. Nepal's poor performance has many causes beyond sample size. First, Nepal's mountainous terrain's steep slopes, varied elevations, and dramatic shadows challenge flat land computer vision algorithms. Ghorbanzadeh et al. [51] found that topographic complexity decreases building detection accuracy by 20-30% compared to flat terrain baselines. Second, Nepal has traditional Himalayan architecture brick and stone construction with terracotta roof tiles, irregular building shapes, and compact urban layouts unlike North American training data's wood frame rectangular structures.

According to Ji et al. [52], the visual feature mismatch caused by architectural variety cannot be compensated by training data from other regions in cross-domain building segmentation. Third, Nepal's dense urban areas' high building density, narrow streets, frequent occlusions, and complicated building adjacencies hamper instance segmentation algorithms. Finally, Nepal's 2015 earthquake caused foundation failure, wall buckling, and partial collapse, unlike temperate disasters' wind or water damage.

Tables 4–6 demonstrate that Mexico and Guatemala, with mountainous terrain, unique architectural styles, and earthquake/volcanic damage patterns, also underperform (Mexico: Minor Damage F1 = 0.000; Guatemala: Minor/Major Damage F1 = 0.000). The persistent pattern that geographic and environmental features dominate performance more than training data quantity changes how disaster damage assessment generalization issues are regarded. According to Tuia et al. [53], the domain shift problem in Earth observation is a geographic shift problem driven by spatial autocorrelation and environmental gradients, resulting in systematic feature distribution differences that cannot be easily addressed through data augmentation.Our findings significantly support this view, showing that regional environmental conditions influence model bias.

Geographic bias has major effects on practical disaster response. Despite good aggregate performance measures, state-of-the-art models may function poorly in places with environmental or structural differences from training data. A model trained on North American disasters may have 90% accuracy in Florida but only 30% accuracy in Nepal, making it unreliable for worldwide humanitarian applications. The variety in dependability defies the universality assumption implicit in many catastrophe informatics systems and highlights the necessity for domain adaptation strategies that explicitly address geographic bias.

4.3. Domain Adaptation to Unseen Geographic Regions

To rigorously quantify geographic bias and evaluate mitigation strategies, we designed controlled out-of-domain experiments using three representative locations—Texas (temperate), Ayiti (tropical), and Portugal (Mediterranean)—that were completely excluded from model training. This experimental design simulates realistic deployment scenarios where emergency responders must assess damage in regions for which pre-trained models have no direct experience. For each target location, we trained the first-place winning solution [41] from scratch on the remaining 16 xBD locations, creating truly out-of-domain test conditions. We then evaluated six configurations: (i) pretrained baseline model, (ii) pretrained model with Fusion Augmentation applied during training, (iii) pretrained model with supervised fine-tuning on 30 labeled target samples, (iv) pretrained model with both fine-tuning and Fusion Augmentation, (v) pretrained model with unsupervised CORAL domain adaptation [39], and (vi) pretrained model with CORAL and Fusion Augmentation. This comprehensive evaluation enables isolation of individual technique contributions and assessment of synergistic effects. The results, presented in Tables 7–9, reveal nuanced patterns of adaptation effectiveness across geographic contexts.

Table 7 reveals moderate domain shift in climatically similar but geographically diverse locations. Under these moderate domain shift conditions due to geographic specificity, pretrained baseline localization performance (F1 = 0.824) is acceptable, but intermediate damage categories (Minor: 0.411, Major: 0.553) show significant degradation (33% and 19% respectively) compared to in-domain per-

 doi:10.20944/preprints202603.2114.v1

formance on similar climatic disasters. This degradation, despite Texas being a temperate flood scenario similar to training data from Arkansas and South Carolina, reveals that geographic specificity—urban development patterns, building codes, construction practices—introduces sufficient feature distribution shift to impair model performance. Fusion Augmentation alone shows slight improvements (Minor: +5.8%, Major: +3.5%), indicating that explicit structural feature enhancement can withstand domain transfer without target-specific adaptation. Supervised fine-tuning on has significantly greater impacts (Minor: +15.0%, Major: +9.2%), highlighting the significance of target domain supervision. Combining fine-tuning with Fusion Augmentation yields the highest results (Minor: 0.505, Major: 0.632), with 23% and 14% improvements above baseline. For Texas, unsupervised CORAL domain adaptation [39] yields minimal effectiveness (Minor: +10.9%, Major: -19.6%), with a surprising performance degradation on Major Damage classification This shows that supervised adaptation with less labeled data outperformed unsupervised feature alignment for modest domain shifts with relatively aligned label distributions.

Table 7. Domain Adaptation Results for Texas (Temperate Region) SDA:Supervised Domain Adaptation, FA: Fusion Augmentation, and UDA: Unsupervised Domain Adaptation.

Method	Localization	No-Damage	Minor	Major	Destroyed
Pretrained model [41]	0.8236	0.6673	0.4105	0.5528	0.2141
Pretrained model+ FA	0.8285	0.6631	0.4344	0.5721	0.2154
Pretrained model+SDA fine-tuning	0.8429	0.6742	0.4723	0.6037	0.2931
Pretrained model+SDA+ FA	**0.8485**	**0.6961**	**0.5047**	**0.6320**	**0.3117**
Pretrained model+UDA-CORAL	0.8246	0.6581	0.4553	0.4446	0.2259
Pretrained model+ UDA-CORAL + FA	0.8315	0.6641	0.4616	0.4494	0.2631

Table 8. Domain Adaptation Results for Ayiti (Tropical Region) SDA:Supervised Domain Adaptation, FA: Fusion Augmentation, and UDA: Unsupervised Domain Adaptation

Method	Localization	No-Damage	Minor	Major	Destroyed
Pretrained model [41]	0.6852	0.6214	0.3108	0.1892	0.2046
Pretrained model+FA	0.7043	0.6638	0.3916	0.2631	0.2354
Pretrained model+SDA fine-tuning	0.8124	0.6946	0.4315	0.2981	0.2852
Pretrained model+SDA+Augmentation	**0.8413**	**0.6961**	**0.4958**	**0.3419**	**0.3140**
Pretrained model+UDA-CORAL	0.7246	0.6417	0.4054	0.2531	0.2419
Pretrained model+UDA-CORAL+FA	0.7515	0.6511	0.4212	0.3058	0.2913

Table 9. Domain Adaptation Results for Portugal (Mediterranean Region). SDA:Supervised Domain Adaptation, FA: Fusion Augmentation, and UDA: Unsupervised Domain Adaptation

Method	Localization	No-Damage	Minor	Major	Destroyed
Pretrained model [41]	0.7049	0.6415	0.3049	0.1992	0.3985
Pretrained model+FA	0.7251	0.6832	0.3257	0.2118	0.4157
Pretrained model+SDA fine-tuning	0.8028	0.7212	0.4082	0.3046	0.4453
Pretrained model+SDA + FA	**0.8285**	**0.7516**	**0.4267**	**0.3685**	**0.4875**
Pretrained model+UDA-CORAL	0.7564	0.6762	0.3481	0.2478	0.4295
Pretrained model+UDA-CORAL+ FA	0.7917	0.6839	0.3795	0.3495	0.4495

Table 8 shows more significant domain shift effects in tropical conditions. The pretrained baseline severely fails under severe domain shift in tropical conditions characterized by dense vegetation and environmental occlusion (Localization: 0.685, Minor: 0.311, Major: 0.189), showing that temperate disaster models are unsuitable for tropical environments. Tropical deployments demonstrate a significant geographic bias, with a 60% decrease in Minor Damage performance and a 76% decrease in Major Damage performance compared to temperate regions. Ji et al. [54] found that tropical vegetation presents unique challenges for building detection, including dense canopy cover, confusing shadows, and moisture-dependent reflectance changes hindering damage assessment. Fusion Augmentation

 doi:10.20944/preprints202603.2114.v1

significantly improves performance (Minor: +26.0%, Major: +39.0%), indicating that edge detection and contrast enhancement partially compensate for environmental occlusion by highlighting structural boundaries despite vegetation covering RGB features Combined fine-tuning with Fusion Augmentation leads to the best tropical performance (Localization: 0.841, Minor: 0.496, Major: 0.342), with 59% and 81% gains over baseline. In significant domain shift scenarios, unsupervised CORAL adaptation shows noteworthy outcomes (Minor: +30.4%, Major: +33.8%), indicating that statistical feature distribution alignment can be beneficial even without target labels. CORAL + Fusion Augmentation enhances unsupervised adaptation (Minor: +35.5%, Major: +61.7%), comparable to supervised fine-tuning without labeled target data. This has crucial implications for quick natural disaster response, where labeled data may not be available during first evaluation.

Table 9 shows that Portugal presents a distinct domain shift scenario combining Mediterranean climate characteristics with wildfire-specific damage patterns. Even though Portugal's Mediterranean environment is closer to tropical training regions, intermediate damage detection dropped to tropical Ayiti levels (Localization: 0.705, Minor: 0.305, Major: 0.199). A number of factors cause this poor generalization. Portugal's eucalyptus forest have dense, uniform vegetation canopies with distinct spectral signatures from temperate deciduous forests and tropical rainforests in training data. Ash covering from the 2017 Portuguese wildfires can mask building boundaries and create false damage signatures, as ash-covered roofs may resemble collapsed structures [55]. Additionally, Portuguese construction predominantly features stone and concrete materials with terracotta roofing architectural characteristics diverging from wood-frame North American structurescreating fundamental feature mismatches. Fusion Augmentation yields slightly lower increases (Minor: +6.8%, Major: +6.3%) than Ayiti, possibly due to ash coverage affecting edge recognition techniques. Supervised fine-tuning results in significant improvements (Minor: +33.9%, Major: +52.9%), whereas the combined technique achieves best performance (Minor: 0.427, Major: 0.369), representing 40% and 85% over baseline. CORAL + Fusion Augmentation enhances unsupervised performance (Minor: +24.5%, Major: +75.4%), with CORAL adaptation showing significant benefits (Minor: +14.2%, Major: +24.4%).

A number of patterns show up across all three out-of-domain locations, indicating geographic bias and adaptation techniques. As shown in Table 10, the proposed method achieves consistent improvements across all three locations. First, baseline performance degradation is proportional to environmental dissimilarity, not geographic distance. Ayiti and Portugal have severe degradation despite being geographically and climatically different, while Texas has moderate degradation despite being far from training regions. This shows that domain shift in disaster damage assessment is environmental and structural, not geographical. Fusion Augmentation has constant gains (5-40% depending on shift severity) across every location, demonstrating the fundamental robustness of specific structural feature encoding to environmental variations. Next, fine-tuning and Fusion Augmentation work together to provide complimentary advantages, resulting in 23-85% enhancements based on damage category and location. Fusion Augmentation yields superior structural representations that fine-tuning can better adjust to target domain characteristics, possibly due to complimentary mechanisms. Finally, unsupervised CORAL adaptation shows variable effectiveness limited benefit for moderate domain shifts (Texas) but substantial gains for extreme shifts (Ayiti, Portugal), suggesting statistical feature alignment becomes more valuable as shift severity increases and label distributions diverge.

Table 10. Performance improvement for each damage class across out-of-domain locations using the Top-1 solution. We report the pretrained baseline, the proposed method (pretrained model + supervised domain adaptation + fusion augmentation), absolute improvement (Δ), and relative improvement (Δ%).

Class	Texas				Ayiti				Portugal			
	Baseline	Proposed	Δ	Δ%	Baseline	Proposed	Δ	Δ%	Baseline	Proposed	Δ	Δ%
No-Damage	0.6673	0.6961	0.0288	4.3	0.6214	0.6961	0.0747	12.0	0.6415	0.7516	0.1101	17.2
Minor	0.4105	0.5047	0.0942	22.9	0.3108	0.4958	0.1850	59.5	0.3049	0.4267	0.1218	39.9
Major	0.5528	0.6320	0.0792	14.3	0.1892	0.3419	0.1527	80.7	0.1992	0.3685	0.1693	85.0
Destroyed	0.2141	0.3117	0.0976	45.6	0.2046	0.3140	0.1094	53.5	0.3985	0.4875	0.0890	22.3

 doi:10.20944/preprints202603.2114.v1

5. Conclusions

In this study, we investigate generalization challenges and geographic biases in AI-based building damage assessment from satellite imagery. Through extensive evaluation of top performing xView2 solutions across 17 disaster locations, we found that geographic and structural characteristics, as well as insufficient training data, degrade model performance. Our results demonstrate that structural feature enhancement through auxiliary channels is not merely beneficial but essential for robust AI-based damage detection across geographically diverse disaster scenarios. The Nepal paradox worst performance despitethe largest training dataset (15,234 images) shows that environmental context, architectural diversity, and terrain complexity impact generalization behavior more than data quantity. Experimental results show a 7.1% increase in overall F1 score, with significant gains in intermediate damage categories like Minor and Major damage. Domain adaptation experiments on three unseen locations demonstrate that Fusion Augmentation with supervised fine-tuning improves minor and major classes by 40.8% and 60.0%, respectively, while unsupervised CORAL improves minor and major damage classes by 24.2% and 39.5% over benchmarks.

Future research can explore several directions to further improve cross-domain generalization in building damage assessment. First, incorporating foundation models, such as large-scale vision transformers pre-trained on diverse remote sensing datasets, could help learn more transferable features and reduce the need for domain specific fine tuning. In addition, generative approaches like GAN based domain adaptation (e.g., CycleGAN) could be used to synthesize realistic target domain images from source data, improving performance under strong domain shifts. Finally, exploring more diverse data augmentation strategies and domain adaptation techniques on larger and more varied datasets could further enhance model robustness and generalization.

Author Contributions: Conceptualization, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Data curation, Bharath Chandra and Unaza Tallal; Formal analysis, Bharath Chandra and Unaza Tallal; Funding acquisition, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Investigation, Bharath Chandra; Methodology, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Project administration, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Resources, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Software, Bharath Chandra and Unaza Tallal; Supervision, Shruti Kshirsagar, Rajiv Bagai and Atri Dutta; Validation, Bharath Chandra and Unaza Tallal; Visualization, Bharath Chandra and Unaza Tallal; Writing—original draft, Bharath Chandra and Unaza Tallal; Writing—review & editing, Shruti Kshirsagar, Unaza Tallal and Rajiv Bagai.

Funding: This research was supported by the Kansas NSF EPSCoR grant (R55156).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset is publically available.

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

References

1. NASA. What are the effects of climate change on extreme weather? https://science.nasa.gov/climate-change/extreme-weather/, 2024. Accessed on 26 November 2024.
2. United Nations Office for Disaster Risk Reduction (UNDRR). Hazard Definition and Classification Review: Technical Report. https://www.undrr.org/publication/hazard-definition-and-classification-review-technical-report, 2024. Accessed on 26 November 2024.
3. Defense Innovation Unit (DIU). Assessing Building Damage from Satellite Imagery. https://www.diu.mil/latest/assessing-building-damage-from-satellite-imagery, 2024. Accessed on 26 November 2024.
4. xView2 Challenge Team. Computer Vision for Building Damage Assessment. https://xview2.org/, 2024. Accessed on 26 November 2024.
5. Gupta, R.; Goodman, B.; Patel, N.; Hosfelt, R.; Sajeev, S.; Heim, E.; Doshi, J.; Lucas, K.; Choset, H.; Gaston, M. Creating xBD: A dataset for assessing building damage from satellite imagery. In Proceedings of the

 doi:10.20944/preprints202603.2114.v1

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 10–17.
6. Melamed, Z.; Shahar, Y.; Waisglass, N.; Kliger, M.; Averbuch-Elor, H. Bias in Building Damage Detection from Satellite Imagery. *arXiv* **2021**. arXiv:2104.06533.
7. Wiguna, S.; Adriano, B.; Mas, E.; Koshimura, S. Evaluation of deep learning models for building damage mapping in emergency response settings. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 5651–5667.
8. Ryan-Mosley, T. How AI can actually be helpful in disaster response. https://www.technologyreview.com/2023/02/20/1068824/ai-actually-helpful-disaster-response-turkey-syria-earthquake/, 2023. Accessed on 26 November 2024.
9. Duarte, D.; Nex, F.; Kerle, N.; Vosselman, G. Satellite image classification of building damages using airborne and satellite image samples in a deep learning approach. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *4*, 89–96.
10. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
11. 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*, 905–927.
12. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS J. Photogramm. Remote Sens.* **2013**, *84*, 85–99.
13. Kim, D.; Won, J.; Lee, E.; Park, K.R.; Kim, J.; Park, S.; Yang, H.; Cha, M. Disaster assessment using computer vision and satellite imagery: Applications in detecting water-related building damages. *Front. Environ. Sci.* **2022**, *10*.
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*.
15. 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.
16. Tsai, F.J.; Lin, S.Y. A class distance penalty deep learning method for post-disaster building damage assessment. *KSCE J. Civ. Eng.* **2024**, pp. 1–15.
17. Jakubik, J.; Muszynski, M.; Vössing, M.; Kühl, N.; Brunschwiler, T. Toward foundation models for earth monitoring: Generalizable deep learning models for natural hazard segmentation. In Proceedings of the Proceedings of the IGARSS 2023—2023 IEEE International Geoscience and Remote Sensing Symposium, 2023, pp. 5638–5641.
18. Melamed, D.; Johnson, C.; Gerg, I.D.; Zhao, C.; Blue, R.; Hoogs, A.; Clipp, B.; Morrone, P. Uncovering bias in building damage assessment from satellite imagery. In Proceedings of the Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, 2024, pp. 8095–8099.
19. Gevaert, C.M.; Buunk, T.; van den Homberg, M.J.C. Auditing geospatial datasets for biases: Using global building datasets for disaster risk management. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 12579–12590.
20. Adedeji, O.; Owoade, P.; Ajayi, O. Image augmentation for satellite images. *arXiv* **2022**. arXiv:2207.14580.
21. Hao, X.; Liu, L.; Yang, R. A review of data augmentation methods of remote sensing image target recognition. *Remote Sens.* **2023**, *15*, 827–867.
22. Ghaffar, M.A.A.; McKinstry, A.; Maul, T.; Vu, T.T. Data augmentation approaches for satellite image super-resolution. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2019**, *4*, 47–54.
23. Takahashi, R.; Matsubara, T.; Uehara, K. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2917–2931.
24. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125.
25. DeVries, T. Improved regularization of convolutional neural networks with Cutout. *arXiv* **2017**. arXiv:1708.04552.
26. Kshirsagar, S.R.; Falk, T.H. Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Transactions on Affective Computing* **2022**, *4*, 1892–1905.
27. Kshirsagar, S.; Falk, T.H. Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *Sensors* **2022**, *22*, 6445.

 doi:10.20944/preprints202603.2114.v1

28. Kshirsagar, S.; Pendyala, A.; Falk, T.H. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Frontiers in Computer Science* **2023**, *5*, 1039261.
29. Xu, Q.; Shi, Y.; Yuan, X.; Zhu, X.X. Universal domain adaptation for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–15.
30. Hu, Y.; Tang, H. On the generalization ability of a global model for rapid building mapping from heterogeneous satellite images of multiple natural disaster scenarios. *Remote Sens.* **2021**, *13*, 984–1008.
31. Parupati, B.C.R.; Kshirsagar, S.; Bagai, R.; Dutta, A. Towards robust building damage detection: Leveraging augmentation and domain adaptation. In Proceedings of the Proceedings of the 2025 IEEE Green Technologies Conference (GreenTech), 2025, pp. 163–167.
32. Mouradi, A.; Kshirsagar, S. Robust Building Damage Detection in Cross-Disaster Settings Using Domain Adaptation. *arXiv preprint arXiv:2603.14694* **2026**. https://doi.org/10.48550/arXiv.2603.14694.
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009, pp. 248–255.
34. Chen, Y.; et al. Advancing Self-Supervised Learning for Building Change Detection... *Remote Sensing* **2025**.
35. Canny, J. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, *PAMI-8*, 679–698.
36. Zuiderveld, K. Contrast limited adaptive histogram equalization. *Graphics gems* **1994**, pp. 474–485.
37. Polesel, A.; Ramponi, G.; Mathews, V.J. Image enhancement via adaptive unsharp masking. *IEEE transactions on image processing* **2000**, *9*, 505–510.
38. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2020**, *11*, 1–46.
39. Sun, B.; Saenko, K. Deep CORAL: Correlation alignment for deep domain adaptation. In Proceedings of the Proceedings of the Computer Vision—ECCV 2016 Workshops, Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, 2016, pp. 443–450.
40. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in neural information processing systems, 2014, pp. 3320–3328.
41. Durnov, V. xView2 First Place Solution. https://github.com/vdurnov/xview2_1st_place_solution, 2019. Accessed on 10 September 2024.
42. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In Proceedings of the Proceedings of the ECCV Workshops, Amsterdam, The Netherlands, 2016; pp. 443–450.
43. Seferbekov, S. xView2 Second Place Solution. https://github.com/selimsef/xview2_solution, 2019. Accessed on 10 September 2024.
44. Khvedchenya, E. xView2 third place. https://github.com/DIUx-xView/xView2thirdplace/tree/master, 2020. Accessed on 10 September 2024.
45. Powers, D.M. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* **2020**.
46. xView2 Access Building Damage. Computer vision for building damage assessment. https://xview2.org/, 2024. Accessed on 26 November 2024.
47. Weber, E.; Kané, H. Building Disaster Damage Assessment in Satellite Imagery with Multi-Temporal Fusion. *arXiv preprint* **2020**. arXiv:2004.05525.
48. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* **2016**, *17*, 1–35.
49. Sublime, J.; Kalinicheva, E. Impact of Vegetation Density on Automatic Building Detection in High-Resolution Satellite Imagery. *Remote Sensing Applications: Society and Environment* **2019**, pp. 1–12.
50. Hao, X.; Liu, L.; Yang, R. Building Damage Detection Using U-Net with Attention Mechanism from Pre- and Post-Disaster Remote Sensing Datasets. *Remote Sensing* **2021**, *13*, 905.
51. Ghorbanzadeh, O.; Blaschke, T.; Gholamnia, K.; Meena, S.R.; Tiede, D.; Aryal, J. Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sensing* **2019**, *11*, 196. https://doi.org/10.3390/rs11020196.
52. Ji, S.; Wei, S.; Lu, M. Fully Convolutional Networks for Multisource Building Extraction in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *56*, 1709–1721. https://doi.org/10.1109/TGRS.2017.2756858.

 doi:10.20944/preprints202603.2114.v1

53. Tuia, D.; Volpi, M.; Copa, L.; Kanevski, M.; Munoz-Mari, J. Domain Adaptation in Remote Sensing: An Overview of Recent Advances. *IEEE Geoscience and Remote Sensing Magazine* **2016**, *4*, 41–57. https://doi.org/10.1109/MGRS.2016.2549320.
54. Ji, S.; Wei, S.; Lu, M. Building Extraction in Tropical Regions from High-Resolution Satellite Imagery Using Multispectral and Shadow Information. *Remote Sensing* **2019**, *11*, 2005. https://doi.org/10.3390/rs11172005.
55. Silva, P.R.D.; Pereira, J.C.; Moreira, F.; Oliveira, T.M. Wildfire Damage Assessment in Portugal Using High-Resolution Satellite Imagery. *Remote Sensing* **2018**, *10*, 1792. https://doi.org/10.3390/rs10111792.