ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-3-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission III, 4–11 July 2026, Toronto, Canada

Efficient Fine-Tuning For Building Damage Assessment with High-Resolution Optical Satellite Imagery: A Case Study for War Damage in Ukraine

Sebastian Gapp[1], Corentin Henry[1], Pablo d’Angelo[1], Friedrich Fraundorfer[1,2], Nina Merkle[1]

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
{sebastian.gapp, corentin.henry, pablo.angelo, nina.merkle}@dlr.de
[2] Institute of Visual Computing, Graz University of Technology, Graz, Austria
friedrich.fraundorfer@tugraz.at

Keywords: Building Damage Assessment, Fine-Tuning, Disaster Response, Segmentation, War Damage in Ukraine.

Abstract

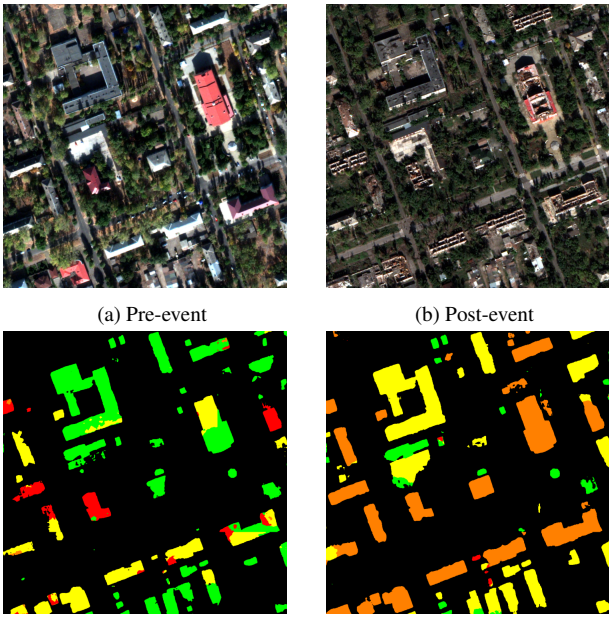
In the aftermath of a disaster, whether natural, industrial, or war-related, a rapid and accurate assessment of building damage is crucial for rescue forces to conduct an effective emergency response. Very high-resolution satellite imagery enables such assessments and serves as an important indicator for understanding the scale of destruction, supporting time-critical rescue operations, and guiding resource allocation. While deep learning models have shown promising results in automating building damage assessment (BDA) from pre- and post-disaster optical satellite imagery, they often fail to generalize to new disasters due to domain shifts. This paper studies the challenge of rapid domain adaptation for BDA in the context of the war in Ukraine. We create a new, challenging dataset annotated with damage grades across six cities in Ukraine, using pre- and post-disaster optical imagery. To facilitate rapid adaptation, we propose an efficient fine-tuning workflow using Low-Rank Adaptation. Our experiments show that this approach substantially improves performance in both out-of-domain and in-domain settings, presenting a practical and data-efficient study for deploying BDA models in time-critical emergency scenarios.

1. Introduction

Building damage assessment (BDA) using optical satellite imagery enables emergency responders to rapidly identify and prioritize affected areas, facilitating timely rescue operations and efficient allocation of resources. Around 167 M people were impacted by natural disasters worldwide in 2024, among whom more than 16 700 died (CRED, 2025). However, emergency mapping services such as the United Nations Satellite Centre (UNOSAT) and the Copernicus Emergency Management Service (CEMS Mapping) still rely primarily on costly manual annotation of optical satellite imagery to evaluate building damage. Valentijn *et al*. estimated that an annotator can label at most 100 buildings per hour with their damage grade (Valentijn et al., 2020). This rate is insufficient for large-scale disasters such as the 2023 Türkiye-Syria earthquake, during which an estimated 160 000 buildings were reported as damaged or destroyed (Toksabay and Butler, 2023). Faced with a similar situation, an AI model could have effectively supported the emergency response efforts by providing rapid initial evaluation and re-evaluations of the building damage in the case of aftershocks.

While AI models offer a scalable alternative, they often fail to generalize to new disasters due to domain shifts caused by variations in geography, building architecture, and disaster characteristics. A common mitigation strategy involves fine-tuning a trained model on a small, rapidly-labeled sample from the target region (Yang et al., 2021, Robinson et al., 2023, Yuan et al., 2021, Bouchard et al., 2022, Nex et al., 2019, Hertel et al., 2025).

In this study, we focus on the ongoing conflict in Ukraine. Prior work on conflict-related BDA in this region has primarily relied on Synthetic Aperture Radar (SAR) data (Bachmann-Gigl and Dabiri, 2024, Dietrich et al., 2025, Rodger and Guida, 2023, Aimaiti et al., 2022), while only a few studies have leveraged very-high-resolution (VHR) optical imagery (Risso et al., 2025, Holail et al., 2024). Moreover, to the best of our knowledge, no publicly available VHR optical BDA dataset exists for this conflict. To address this gap, we created a new BDA dataset com-



(a) Pre-event (b) Post-event

(c) Before fine-tuning (d) After fine-tuning

Figure 1. Qualitative comparison for out-of-domain generalization on the Sievierodonetsk region of the Ukraine building damage dataset before and after fine-tuning. Four-class building damage color coding: no damage, minor damage, major damage and destroyed, with unclassified buildings. Image sources: 13.10.2020 / 21.09.2022, WorldView-3, supplied by European Space Imaging © 2020 / 2022 Vantor.

prising 5075 manually annotated buildings across six Ukrainian regions: Bakhmut, Hostomel-Irpin, Lysychansk, Mariupol, Rubizhne, and Sievierodonetsk.

We demonstrate an effective and efficient approach for rapid domain adaptation, integrating Low-Rank Adaptation (LoRA) (Hu et al., 2022) to efficiently fine-tune a state-of-the-art model trained on large-scale public datasets. Our experiments show a substantial domain gap between existing benchmarks and our Ukraine dataset, underscoring the critical need for domain-specific fine-tuning. Our approach achieves significant performance gains in two settings. On the one hand, in out-of-domain scenarios, it generalizes effectively to unseen Ukrainian cities in our dataset, as illustrated in Fig. 1. On the other hand, in in-domain scenarios, it fine-tunes successfully on few samples from target cities. Notably, fine-tuning improves damage classification F1-scores by up to $40\,\%$ in the out-of-domain setting and demonstrates substantial gains with as few as 12 labeled buildings per class, requiring only $30\,\mathrm{min}$ of annotation time.

2. Related Work

Combining pre- and post-disaster satellite imagery with semantic segmentation techniques to extract building outlines and predict their damage classes is a standard approach for BDA (Gupta and Shah, 2020, Khvedchenya and Gabruseva, 2021, Zheng et al., 2021, Chen et al., 2022, Xiao et al., 2023, Lu et al., 2024, Chen et al., 2024, Hertel et al., 2025). The damage classes, following the standard established for optical satellite datasets (Gupta et al., 2019), are typically divided into six categories: *no building*, *no damage*, *minor damage*, *major damage*, *destroyed* and *unclassified*. We adopt this classification scheme throughout this paper.

2.1 BDA for Conflicts

Existing works on conflict-related BDA have focused on the use of medium-resolution Sentinel-1 SAR data due to its open access, all-weather capabilities, and independence from daylight, leveraging it to detect and assess building damage caused by the war in Ukraine (Bachmann-Gigl and Dabiri, 2024, Dietrich et al., 2025, Rodger and Guida, 2023). Other approaches combine medium-resolution SAR with optical data; for instance, Aimaiti *et al*. used both Sentinel-1 SAR and Sentinel-2 optical images for BDA (Aimaiti et al., 2022). At a higher resolution, the BRIGHT dataset (Chen et al., 2025) provides a testbed for BDA method development by pairing VHR pre-disaster optical with VHR post-disaster SAR imagery. The dataset includes building damage annotations for the city of Mariupol during the war in Ukraine.

While SAR is a highly robust modality, VHR optical images provide complementary information about texture, color, and fine-grained structural details absent in radar data. Therefore, leveraging all available modalities is crucial to gain a comprehensive understanding of the situation on the ground. Leveraging pre- and post-disaster VHR optical imagery, Risso *et al*. performed BDA for the city of Mariupol (Risso et al., 2025). Similarly, Holail *et al*. conducted a time-series analysis of building damage in the Gaza Strip affected by the Israel-Hamas war (Holail et al., 2024). While these studies confirm the viability of using VHR optical imagery for BDA in conflict zones, it remains unclear how well these methods generalize across cities and how fine-tuning can be made efficient to minimize annotation and training times. Our work aims to fill this gap.

2.2 BDA Datasets

There are two large-scale manually-annotated datasets available for BDA currently: xBD (Gupta et al., 2019) and BRIGHT. The approach of the xBD dataset relies solely on pre- and post-disaster optical satellite imagery. The dataset itself contains 11 034 pairs of pre- and post-disaster optical satellite images collected from 19 natural disasters spanning six sources of disasters: volcanoes, floods, wind, earthquakes, tsunamis, and fires. Although it covers events from 19 regions worldwide, the dataset is primarily centered around disasters that struck areas in the United States. On the other hand, the BRIGHT dataset aims at enabling all-weather and day–night–independent BDA by replacing post-disaster optical images with SAR data. BRIGHT encompasses seven disaster types across 14 regions and contains 4246 pairs of pre- and post-disaster satellite images. Unlike the fully manual annotations used in xBD, BRIGHT integrates manually annotated building outlines with existing damage annotations provided by Copernicus and UNOSAT.

To address the high cost and limited scope of manually annotated real-world datasets, SynthBD[1] was introduced as a large-scale synthetic dataset for BDA. It contains 12 500 pairs of rendered pre- and post-disaster optical RGB images with corresponding ground truth damage maps. The dataset focuses on modeling localized structural damage, similar to that found in wind-related disasters and earthquakes. By augmenting real-world training data with synthetic samples, SynthBD increases scene variability. Internal evaluations demonstrate that this augmentation strategy can significantly improve out-of-domain (OOD) generalization. Notably, this approach enhances model performance even in challenging OOD scenarios, such as conflict-related damage in Ukraine, a disaster type not represented in the xBD training set.

2.3 Architectures for BDA

A variety of approaches have been developed for BDA over the last years. Early CNN-based methods (Gupta and Shah, 2020, Khvedchenya and Gabruseva, 2021, Zheng et al., 2021, Risso et al., 2025, Holail et al., 2024) adopted encoder-decoder designs inspired by U-Net (Ronneberger et al., 2015) which uses skip connections to allow fine-grained spatial details to flow from the encoder to the decoder. These models typically process pre- and post-disaster images independently before fusing their feature maps during decoding. With the introduction of Vision Transformers (ViTs), which effectively capture long-range dependencies through self-attention (Dosovitskiy et al., 2021), several transformer-based BDA architectures achieved notable performance improvements on the xBD dataset (Chen et al., 2022, Xiao et al., 2023, Lu et al., 2024). However, the quadratic computational complexity of self-attention in ViTs poses efficiency challenges. To address this limitation, Liu *et al*. proposed VMamba (Liu et al., 2024), a vision-adapted State Space (VSS) model derived from the Mamba architecture (Gu and Dao, 2024). VMamba extends the one-dimensional sequential data processing of Mamba, originally designed for natural language processing, to the spatial domain through a 2D Selective Scan (SS2D) mechanism that employs four-way directional scanning to process visual data effectively. Building on this foundation, ChangeMamba (Chen et al., 2024) further advances BDA performance, surpassing previous methods on the xBD dataset.

[1] Internal dataset currently under development.

2.4 Fine-tuning

Recent BDA methods have demonstrated promising results on established benchmarks. However, they exhibit a significant generalization gap when deployed on out-of-distribution datasets (Benson and Ecker, 2020, Wiguna et al., 2024, Yang et al., 2021, Robinson et al., 2023, Hertel et al., 2025, Yuan et al., 2021, Risso et al., 2025, Holail et al., 2024). This limitation arises from the inherent domain gap between individual disasters, where differences in geographical locations, building architecture, damage features, and sensor characteristics create distinct visual distributions. Due to the relatively small number of past disasters represented in current datasets, existing methods often fail to generalize effectively to unseen disasters. This limits their practical application to real-world emergency response, where rapid deployment is critical. To reduce the distribution shift between training data and a novel disaster, a common approach is to fine-tune a trained model on a small amount of data from the new target disaster (Yang et al., 2021, Hertel et al., 2025, Robinson et al., 2023, Yuan et al., 2021, Bouchard et al., 2022, Nex et al., 2019, Risso et al., 2025, Holail et al., 2024). For instance, Bouchard *et al*. (Bouchard et al., 2022) evaluated out-of-domain generalization across various data splits on xBD, while Hertel *et al*. (Hertel et al., 2025) investigated the transition from semi-supervised to supervised learning methods for domain adaptation in flood-induced BDA, examining different building subsets as training data. Existing works on supervised learning commonly adopt standard full fine-tuning, where all model parameters are updated (Yuan et al., 2021, Bouchard et al., 2022, Hertel et al., 2025).

Fine-tuning the Mamba Architecture: Given the novelty of the Mamba architecture, the literature on its fine-tuning for vision applications remains sparse. A prominent method for parameter-efficient fine-tuning (PEFT) is LoRA, initially proposed for domain adaptation of transformer-based large language models (Hu et al., 2022). While LoRA can be applied to any neural network with dense layers, its optimal configuration for Mamba architectures, specifically, which linear layers to adapt and which hyperparameter values to select, remained unclear. Yoshimura *et al*. (Yoshimura et al., 2025) benchmark PEFT methods for Mamba, proposing architecture-specific adaptations distinct from transformer-based approaches. The authors investigate partial LoRA configurations, applying reparameterization to selected layers. Their findings demonstrate that partially applying LoRA, specifically to the input projection layer, achieves the best performance, particularly when training data is sparse. Moreover, they show that LoRA substantially outperforms full fine-tuning while remaining robust to overfitting even as the number of trainable parameters increases.

3. Method

Our method is based on the state-of-the-art BDA model, ChangeMamba, using LoRA for efficient fine-tuning. The following subsections detail the ChangeMamba architecture and the LoRA integration.

3.1 BDA model

We build our approach on the publicly available VSS model ChangeMamba (Chen et al., 2024). ChangeMamba uses an encoder-decoder architecture, consisting of two encoders and decoders. The shared-weight encoder processes pre- and post-disaster images separately through four hierarchical stages, each comprising a downsampling layer (except the first stage) and a stack of VSS blocks. Subsequently, each decoder stage fuses skip connections from the encoder with feature maps from the previous decoder stage before upsampling to produce a building localization and damage classification map. As a last step, the building localization map is used to mask out non-building pixels in the damage classification map to produce the final damage map.

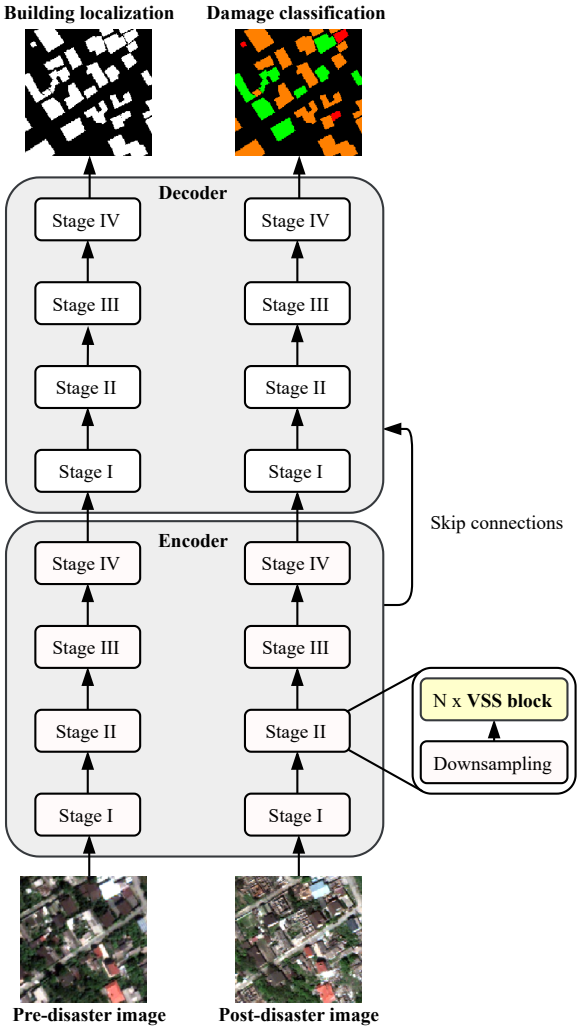


Figure 2. Schematic of the overall ChangeMamba architecture with individual stages. The Visual State Space (VSS) block position within the architecture is highlighted in yellow.

With less than 1 % performance gain between the medium and large ChangeMamba variants (*Small* and *Base*), the latter entailing a substantial increase in number of parameters, we opted for the more efficient *Small* architecture in our experiments. Fig. 2 provides a schematic overview of the ChangeMamba architecture.

3.2 Adapting LoRA for BDA Fine-tuning

Yoshimura *et al*. (Yoshimura et al., 2025) demonstrated that LoRA can surpass full Mamba fine-tuning. Motivated by these findings, we use LoRA to efficiently adapt the ChangeMamba model. LoRA keeps pre-trained weights $\mathbf{W_0}$ frozen and repara-

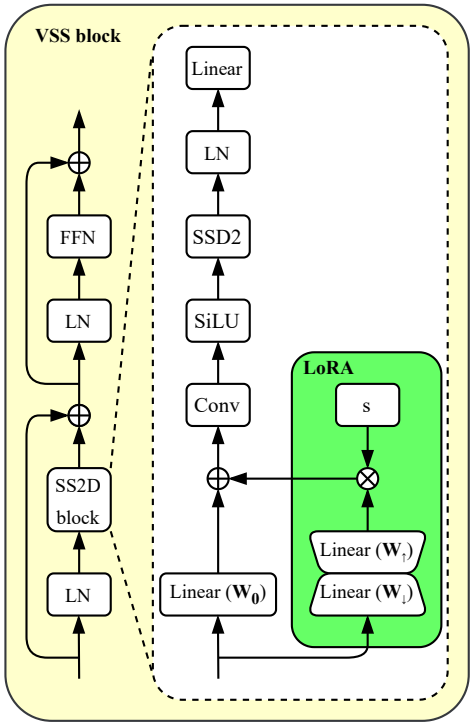


Figure 3. Detailed structure of the Visual State Space (VSS) block. Integrated Low-Rank Adaptation (LoRA) modules, pre-trained weights $\mathbf{W_0}$ and introduced trainable low-rank decomposition matrices ($\mathbf{W}_\downarrow, \mathbf{W}_\uparrow$) are highlighted in green. Abbreviations: FFN (feed-forward network), Conv (depth-wise convolution), and LN (LayerNorm).

metrizes the linear weight matrices as

$$\mathbf{W} = \mathbf{W_0} + s \cdot \mathbf{W}_\downarrow \mathbf{W}_\uparrow \tag{1}$$

where s is a scaling parameter and only the low-rank matrices $\mathbf{W}_\downarrow \in \mathbb{R}^{d \times r}$ and $\mathbf{W}_\uparrow \in \mathbb{R}^{r \times d}$ are trainable, significantly reducing the number of trainable parameters and enabling hardware-efficient training. The rank $r \ll d$ determines the capacity-efficiency trade-off, where d represents the original feature dimension.

While LoRA can be applied to any dense layer, the optimal choice of which layers to adapt is model-dependent. We follow the best practices of (Yoshimura et al., 2025), applying LoRA to the input projection layer. Unless stated otherwise, we adopt the hyperparameters from their work. To the best of our knowledge, this is the first application of LoRA for fine-tuning a multi-class BDA model. Fig. 3 illustrates the integration of LoRA modules within the VSS block of the ChangeMamba architecture.

3.3 Pre-processing

Before BDA algorithms can be applied, the satellite imagery needs to be co-registered. This is typically performed by matching tie points between the original satellite scenes, performing a bundle block adjustment to estimate consistent image orientations (d’Angelo, 2013), followed by orthorectification using a digital elevation model (DEM). Due to the large changes and different image acquisition conditions, the tie point matching procedure was not always successful, even when applying recent deep learning based feature matchers (Morelli et al., 2024). In these cases, tie points were manually measured and used for bundle block adjustment of the image scenes. In order to obtain good co-registration of the ortho images, the bundle block adjustment minimizes both the image tie point residuals as well as their distance to the DEM. The Copernicus DEM used for orthorectification contains smoothed version of larger buildings and vegetation, which would lead to deformed buildings in the orthorectified images. Thus, a morphological filtering is applied to obtain a digital terrain model (DTM), which does not include above-ground objects, and consequently avoids distortion of larger buildings. To enable efficient processing and compatibility with common deep learning architectures, we converted the input data to 8-bit integers. We use $2\,\%$ and $98\,\%$ quantile truncation for this conversion to create a robust mapping that is insensitive to outlier pixel values, thereby preserving the dynamic range of relevant scene features.

4. Experiments

To assess the real-world applicability of automated BDA on our Ukraine BDA dataset, we evaluated the generalization capability of ChangeMamba across two settings: out-of-domain and in-domain. Out-of-domain generalization evaluates whether a model can perform well in geographic regions it has never seen during training. This is highly relevant in practice because it corresponds to applying a model trained on annotated data from various regions to a new target region where no labeled data are available. In-domain generalization evaluates how effectively the model adapts to a target region when we provide a small amount of training data from that specific region. This scenario mirrors disaster response, where practitioners must rapidly fine-tune models for an affected area. For both out- and in-domain generalization analysis, we employ a two-stage training strategy. The ChangeMamba model is first trained on the large-scale, public BDA datasets xBD and SynthBD. Subsequently, we use LoRA to fine-tune the model on the target domain.

4.1 Ukraine BDA dataset

Area selection: To capture variations in region-specific building typologies, we sampled six regions of interest across Ukraine: Mariupol, Bakhmut, Hostomel-Irpin, Lysychansk, Rubizhne and Sievierodonetsk. Site selection was guided by UNOSAT building-damage assessments to balance damage classes and encompass a broad spectrum of conflict-related damage features. For each region, two $500\,\mathrm{m}$-large square tiles were selected for annotation. In Sievierodonetsk, we annotated three tiles to roughly match the total building counts of the other areas and to ensure that region-specific building typologies were represented.

Annotation protocol and dataset statistics: We annotated the dataset in two stages: first, we labeled the building outlines; second, we assigned the four damage classes. To minimize bias and potential errors, we separated tasks across teams: one team generated the initial annotations, and a different team subsequently reviewed and refined them. When classification disagreements persisted for specific buildings after review, we labeled those as unclassified. Overall, we manually annotated 5075 building instances for the 6 Ukrainian cities Bakhmut (797 buildings, Hostomel-Irpin (634 buildings), Lysychansk (574 buildings), Mariupol (1116 buildings), Rubizhne (1205 buildings) and Sievierodonetsk (749 buildings). Fig. 4 provides a summary of the dataset statistics.

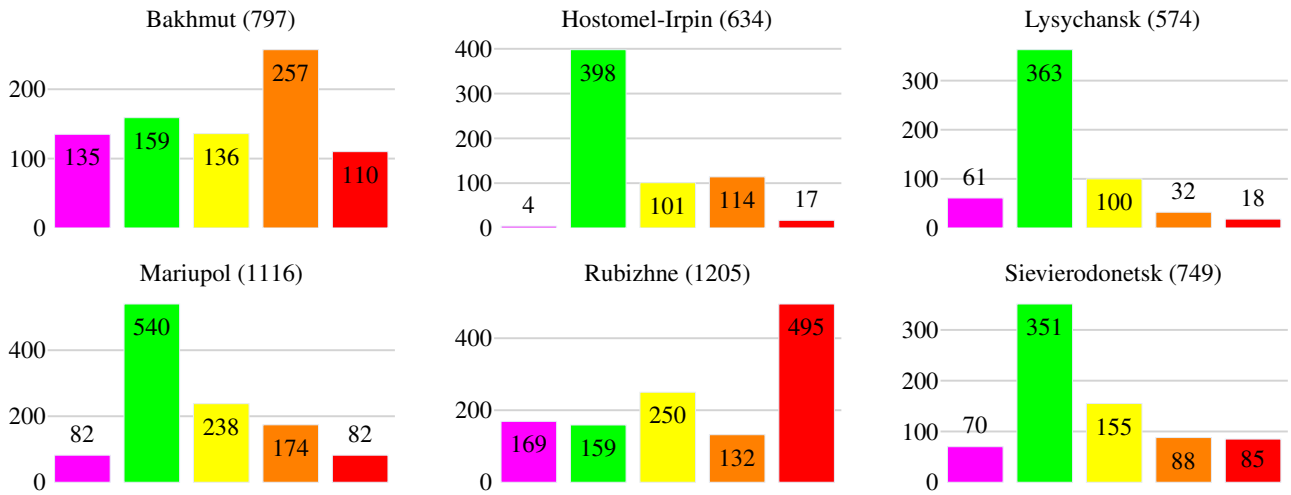


Figure 4. Statistics from our Ukraine building damage assessment dataset, with the number of instances annotated per area of interest and per class: no damage, minor damage, major damage and destroyed. Unclassified buildings are shown as .

4.2 Model Training and Evaluation

We employ a two-stage training strategy consisting of training and subsequent fine-tuning. The primary goal of training is to learn feature representations that provide a robust initialization for fine-tuning, allowing effective adaptation to the target domain during fine-tuning.

Training: We train the ChangeMamba model on the large-scale BDA datasets, xBD and SynthBD. To improve generalization to conflict-related damage scenarios absent in xBD, we augment training with SynthBD, which has shown to improve out-of-domain performance for Ukraine. For this stage, following (Chen et al., 2024), we use the AdamW optimizer with a learning rate of 1×10^{-4}, a weight decay of 5×10^{-3}, and a batch size of 10 for 40 000 iterations. In a practical application, this computationally intensive training is performed once, before any specific disaster event. We employ a combined loss function comprising cross-entropy (CE) loss and Lovász (Lov) loss (Berman et al., 2018):

$$\mathcal{L} = \mathcal{L}_{\text{CE}}^{\text{loc}} + \mathcal{L}_{\text{Lov}}^{\text{loc}} + \mathcal{L}_{\text{CE}}^{\text{clf}} + \mathcal{L}_{\text{Lov}}^{\text{clf}}. \tag{2}$$

All experiments are conducted on a single NVIDIA A100 80GB GPU.

Mitigating non-determinism in Mamba training: Due to non-deterministic atomic operations in the CUDA implementation of Mamba, identical training runs do not converge identically, producing varying model weights and consequently leading to varying performance, complicating direct experimental comparisons. To ensure the comparability and robustness of our results, we implemented a systematic model selection protocol. During the training phase we perform ten independent runs. Before fine-tuning, we evaluate each of the ten resulting models on the respective training set and select the corresponding best-performing model. This selected model is then used for a single subsequent fine-tuning run. By selecting the top-performing model from multiple runs, this approach mitigates the impact of training stochasticity and ensures that observed performance differences reflect dataset distinctions rather than random training variations. We note that during the fine-tuning stage, only the newly-injected LoRA linear layers and the single-layer segmentation head are trained, while the Mamba blocks remain frozen. Therefore, the non-deterministic variations originating from the Mamba architecture are isolated to the training phase and do not affect the fine-tuning process itself.

Data augmentation: To enhance the robustness of the model and avoid overfitting, we augment the data with horizontal and vertical flips, random rotations, and a random square crop of side length 512 px.

Evaluation: We evaluate the joint task of building localization and damage classification using the xView2 metric (Gupta et al., 2019), an often-reported metric for BDA. Here, the overall performance is measured by a weighted average of the F1-scores for localization and damage classification, given by:

$$F_1^{\text{comb}} = 0.3 \cdot F_1^{\text{loc}} + 0.7 \cdot F_1^{\text{clf}}. \tag{3}$$

Whereby F_1^{loc} is the F1-score for building localization and F_1^{clf} is the harmonic mean of the per-class damage F1-scores. Unless stated otherwise, we compare models using F_1^{comb}. We report metrics for four damage classes unless otherwise specified. Additionally, we provide results for combined minor and major damage classes (referred to as the *damaged* class) to align with prior works where distinguishing between these classes is not required or feasible.

4.3 Evaluating Out-of-Domain Generalization for BDA

Experimental setup: To evaluate the out-of-domain generalization, we perform a leave-one-region-out cross-validation using the six annotated regions of the Ukraine dataset. For each of the six resulting folds, one region serves as the unseen test set while the other five are used for fine-tuning. The models are fine-tuned for 13 000 iterations using a batch size of 10. Following (Yoshimura et al., 2025), we set the LoRA rank to 64 and use an initial learning rate of 1×10^{-3} with a linear warmup and cosine annealing schedule. Each training fold completes in approximately 30 min on an NVIDIA A100 80GB GPU.

Results: A comparison of the trained model’s performance on individual test regions of the Ukraine dataset, shown in Tab. 1, with the xBD holdout set, reveals a significant domain gap between the xBD dataset and the Ukraine dataset, aligning with the findings of Risso *et al.* for Mariupol (Risso et al., 2025). The initially

Test Source	Fine-tuning	F_1^{comb}	F_1^{clf}	F_1^{loc}
xBD hold	✗	**0.79**	**0.76**	**0.86**
Bakhmut	✗	0.42	0.27	0.78
xBD hold	✗	**0.79**	**0.76**	**0.86**
Hostomel-Irpin	✗	0.35	0.18	0.75
xBD hold	✗	**0.79**	**0.76**	**0.86**
Lysychansk	✗	0.48	0.42	0.61
xBD hold	✗	**0.79**	**0.76**	**0.86**
Mariupol	✗	0.43	0.29	0.73
xBD hold	✗	**0.79**	**0.76**	**0.86**
Rubizhne	✗	0.45	0.31	0.79
xBD hold	✗	**0.80**	**0.77**	**0.86**
Sievierodonetsk	✗	0.30	0.17	0.59

Table 1. Out-of-domain model performance on the xBD holdout set *vs*. individual Ukraine test regions, before fine-tuning. The low baseline performance on the Ukraine regions reveals a significant domain gap, highlighting the necessity of fine-tuning. For each region, we report the score from the best-performing model out of ten training runs, selected per-region on the respective training set. Best values are in bold.

low classification score on the Ukraine dataset highlights the need for fine-tuning to this new domain. Notably, the model's localization performance is substantially higher than its classification performance, indicating that localization features are more transferable across domains than classification-specific features.

Tab. 2 shows that fine-tuning yields substantial improvements over the non-finetuned baseline, with F_1^{comb}, F_1^{clf}, and F_1^{loc} increasing by up to $36\,\%$, $40\,\%$, and $26\,\%$, respectively. Out-of-domain performance remains largely stable across regions, with Hostomel-Irpin being an exception with the lowest scores. A qualitative review of the predictions, presented in Figs. 1 and 5, shows that this performance drop is mainly due to confusion between the classes *major damage* and *destroyed*. We attribute this to a cluster of small, densely-packed industrial buildings with metal roofs, whose visual characteristics significantly differ from the buildings in the training data. Despite this mismatch, the model successfully identifies overall damage hot spots.
While overall F_1^{comb} improves after fine-tuning, we observe that the performance on individual classes may occasionally decrease compared to the non-finetuned baseline model, although the remaining classes show consistent improvements. We find that merging the minor and major damage classes after training yields additional performance gains. In line with other studies (Yuan et al., 2021, Wiguna et al., 2024), this suggests that the classifier struggles to distinguish between minor and major damage, particularly before fine-tuning.

4.4 Evaluating In-Domain Generalization for BDA

Experimental setup: We evaluate in-domain generalization on each of the six annotated regions from the Ukraine dataset using a $50\,\%$ train-test split per region. Since model selection is made based on the initial performance on the respective training sets, out-of-domain and in-domain results differ before fine-tuning, also due to partial buildings caused by the split. To study the impact of annotation budget on performance, we fine-tune our model on curated subsets of labeled buildings from the training split similar as (Hertel et al., 2025). We define annotation budgets that reflect practical time constraints in rapid damage assessment scenarios: 12 and 24 buildings per class. Based on an estimated annotation rate of 100 buildings per hour (Valentijn et al., 2020), these budgets correspond to approximately $30\,\text{min}$ and $60\,\text{min}$ of labeling time, respectively. To provide an upper bound for performance comparison, we further provide results for a scenario using all available training data. To ensure a fair comparison, the smaller annotation budgets are proper subsets of the larger ones. To prevent artifacts caused by incomplete building features and missing context, we exclude any buildings that touch the image boundary from our selection. Consequently, some annotation budgets are unavailable for certain regions if the number of available, fully-visible buildings is insufficient; even though the number of buildings per area and class would allow for the budget.

During fine-tuning, we focus solely on the classification task. The localization loss is ignored because our building selection process provides only positive examples (where buildings are) and no negative examples (where buildings are not). This is reasonable to do so given that the model already achieves high localization performance without fine-tuning, as demonstrated in our out-of-domain experiments. To simulate a limited annotation scenario, buildings not included in the selected budget are ignored during the loss calculation. To mitigate data sparsity, we reduce the LoRA rank to 4 and decrease fine-tuning to 100 epochs, enabling the process to complete in approximately $6\,\text{min}$ on a single NVIDIA A100 80GB GPU.

Results: Given the focus on the classification task in the in-domain experiment, we subsequently analyze the F_1^{clf} score, unless otherwise stated.

Tab. 3 shows across all areas that both F_1^{comb} and F_1^{clf} increase continuously with the number of annotated buildings. With 12 buildings per class ($30\,\text{min}$ annotation time budget), the in-domain classification performance improves significantly compared to the non-fine-tuned baseline, achieving gains of up to $36\,\%$ for Mariupol. This aligns with the findings of Hertel *et al*. (Hertel et al., 2025). Despite focusing on a different disaster type, the authors similarly observe that a few samples suffice to significantly enhance initial model performance. Doubling the budget to 24 buildings per class ($60\,\text{min}$ annotation time budget) yields further improvements of up to $8\,\%$, though with diminishing returns compared to the initial 12 buildings per class. When compared to using all available annotations, we observe similar improvement trends, with Sievierodonetsk standing out with a $20\,\%$ improvement. The quantitative improvements are also reflected in the qualitative results shown in Fig. 6. While damage hotspots are partially visible even before fine-tuning, the individual damage classes are progressively refined as more samples are shown.

The F_1^{loc} score fluctuates but remains generally high across different budgets, which is expected as our fine-tuning process only optimizes the classification head. When comparing the non-fine-tuned model to one trained on all annotations, we observe a performance decrease for Bakhmut ($7\,\%$) and Sievierodonetsk ($6\,\%$), while other areas see an increase of up to $10\,\%$. Although the localization head is not explicitly optimized, its performance is nonetheless affected due to the shared encoder. We hypothesize that optimizing the classification head implicitly refines the shared encoder, yielding features that can benefit the localization task.

Consistent with the out-of-domain experiment, we observe that while individual damage class F1-scores can fluctuate during fine-tuning, the overall F_1^{clf} demonstrates a clear upward trend. As for the out-of-domain experiment, merging the minor- and

Test region	Fine-tuning	F_1^{comb}	F_1^{clf}	F_1^{loc}	Damage F1-scores				
					No damage	Minor	Major	Destroyed	Combined
4 damage classes									
Bakhmut	✗	0.42	0.27	0.78	0.31	0.31	0.16	0.49	-
	✓	**0.63**	**0.55**	**0.82**	**0.54**	**0.52**	**0.60**	**0.55**	-
Hostomel-Irpin	✗	0.35	0.18	0.75	**0.76**	0.25	0.18	0.08	-
	✓	**0.41**	**0.24**	**0.82**	0.68	**0.29**	**0.28**	**0.12**	-
Lysychansk	✗	0.48	0.42	0.61	**0.78**	0.31	0.38	0.41	-
	✓	**0.67**	**0.60**	**0.84**	0.65	**0.61**	**0.62**	**0.53**	-
Mariupol	✗	0.43	0.29	0.73	**0.62**	0.17	0.24	0.49	-
	✓	**0.59**	**0.51**	**0.79**	0.46	**0.51**	**0.50**	**0.56**	-
Rubizhne	✗	0.45	0.31	0.79	0.45	0.34	0.16	**0.76**	-
	✓	**0.62**	**0.52**	**0.86**	**0.59**	**0.62**	**0.48**	0.43	-
Sievierodonetsk	✗	0.30	0.17	0.59	0.36	0.13	0.10	0.37	-
	✓	**0.66**	**0.57**	**0.85**	**0.58**	**0.74**	**0.63**	**0.42**	-
3 damage classes: minor + major combined									
Bakhmut	✗	0.51	0.40	0.78	0.31	-	-	0.49	0.43
	✓	**0.67**	**0.61**	**0.82**	**0.54**	-	-	**0.55**	**0.77**
Hostomel-Irpin	✗	0.35	0.19	0.75	**0.76**	-	-	0.08	0.34
	✓	**0.42**	**0.25**	**0.82**	0.68	-	-	**0.12**	**0.44**
Lysychansk	✗	0.53	0.49	0.61	**0.78**	-	-	0.41	0.43
	✓	**0.69**	**0.62**	**0.84**	0.65	-	-	**0.53**	**0.72**
Mariupol	✗	0.54	0.46	0.73	**0.62**	-	-	0.49	0.34
	✓	**0.64**	**0.57**	**0.79**	0.46	-	-	**0.56**	**0.76**
Rubizhne	✗	0.59	0.51	0.79	0.45	-	-	**0.76**	0.44
	✓	**0.64**	**0.55**	**0.86**	**0.59**	-	-	0.43	**0.70**
Sievierodonetsk	✗	0.41	0.34	0.59	0.36	-	-	0.37	0.30
	✓	**0.66**	**0.57**	**0.85**	**0.58**	-	-	**0.42**	**0.86**

Table 2. Out-of-domain inference results using leave-one-region-out cross-validation on six annotated regions of the Ukraine dataset. Each region serves as an unseen test set while the remaining five are used for training. Results shown before and after fine-tuning, with fine-tuning yielding substantial improvements over the non-finetuned baseline. Best values are in bold.

major-damage classes post-training increases the reported F1-scores. The observed improvements using this metric then are 18 % for fine-tuned models and 38 % for non-fine-tuned models.

5. Conclusion

In this work, we addressed the critical challenge of domain adaptation for automated BDA in the context of the conflict in Ukraine, for which we created a new dataset based on VHR optical imagery. We demonstrated that a model pre-trained on large-scale public datasets can be rapidly and effectively adapted to this new, challenging domain using LoRA.

The out-of-domain and in-domain experiments we conducted demonstrate that domain-specific fine-tuning is crucial for accurate damage assessment in new geographical regions. Our results confirm that LoRA facilitates quick and effective domain adoption for the domain of BDA. In the out-of-domain setting, fine-tuning on a collection of Ukrainian regions allows the model to generalize successfully to unseen areas within Ukraine. Even in challenging cases with significant visual differences, the method proved robust at identifying general damage hotspots. Furthermore, our in-domain experiments show that fine-tuning with a minimal annotation budget is a highly effective strategy for adapting the model to new geographic locations, even without prior training data from a similar region. We found that as little as 30 min of annotation effort can yield significant performance improvements, making this a practical approach for emergency response where extensive data collection is infeasible.

To further improve the efficiency of BDA, future work should focus on optimizing the annotation budget for in-domain adaptation. While initial parameter selections are promising, further studies are required to determine optimal budget allocation and sample selection strategies.

References

Aimaiti, Y., Sanon, C., Koch, M., Baise, L. G., Moaveni, B., 2022. War Related Building Damage Assessment in Kyiv, Ukraine, Using Sentinel-1 Radar and Sentinel-2 Optical Images. *Remote Sensing*, 14(24).

Bachmann-Gigl, U., Dabiri, Z., 2024. Cultural Heritage in Times of Crisis: Damage Assessment in Urban Areas of Ukraine Using Sentinel-1 SAR Data. *ISPRS International Journal of Geo-Information*, 13(9).

Benson, V., Ecker, A., 2020. Assessing Out-of-Domain Generalization for Robust Building Damage Detection. *NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (AI+HADR)*.

Berman, M., Triki, A. R., Blaschko, M. B., 2018. The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4413–4421.

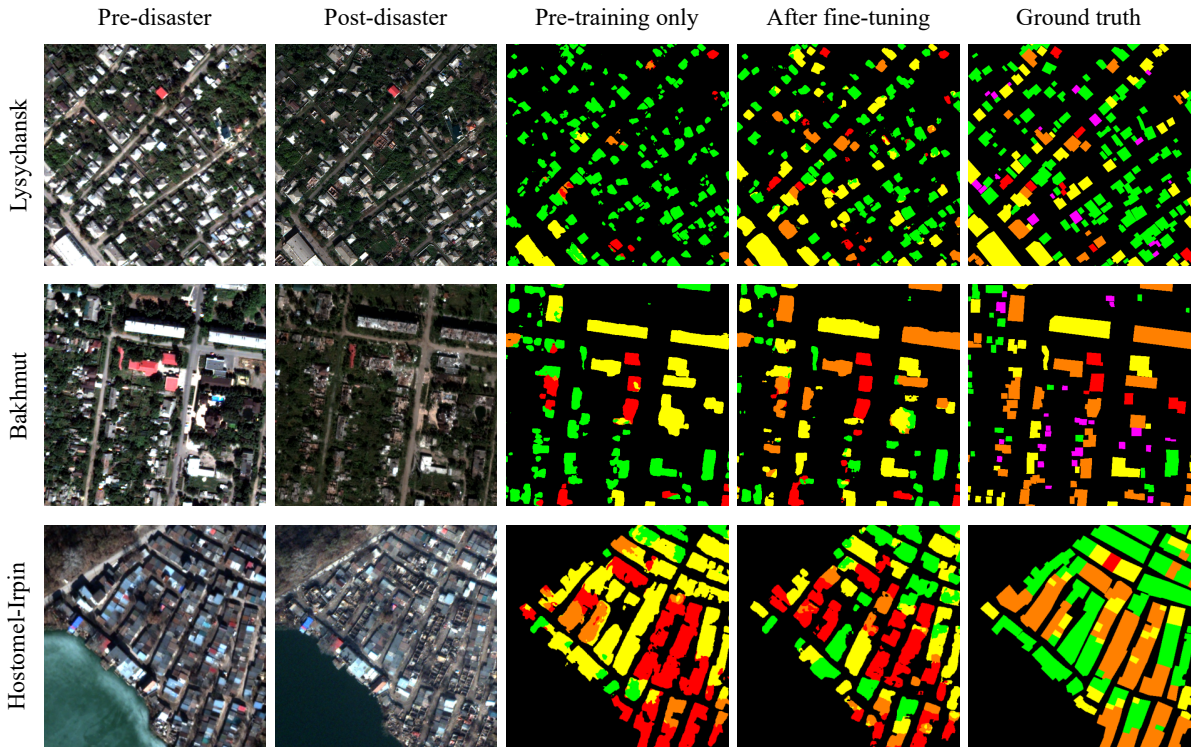


Figure 5. Qualitative results for the out-of-domain generalization experiment on the Ukraine building damage assessment dataset. Four-class building damage color coding: no damage, minor damage, major damage and destroyed. Unclassified buildings are shown as . Image sources: Lysychansk: pre: 26.08.2021, WorldView-2, © 2021 European Space Imaging; post: 21.09.2022, WorldView-3, supplied by European Space Imaging © 2022 Vantor. Bakhmut: pre: 13.07.2021, WorldView-2, © 2021 European Space Imaging; post: 05.08.2023, WorldView-2, supplied by European Space Imaging © 2023 Vantor. Hostomel-Irpin: pre: 28.02.2022, WorldView-2, supplied by European Space Imaging © 2022 Vantor; post: 31.03.2022, WorldView-3, © 2022 European Space Imaging.

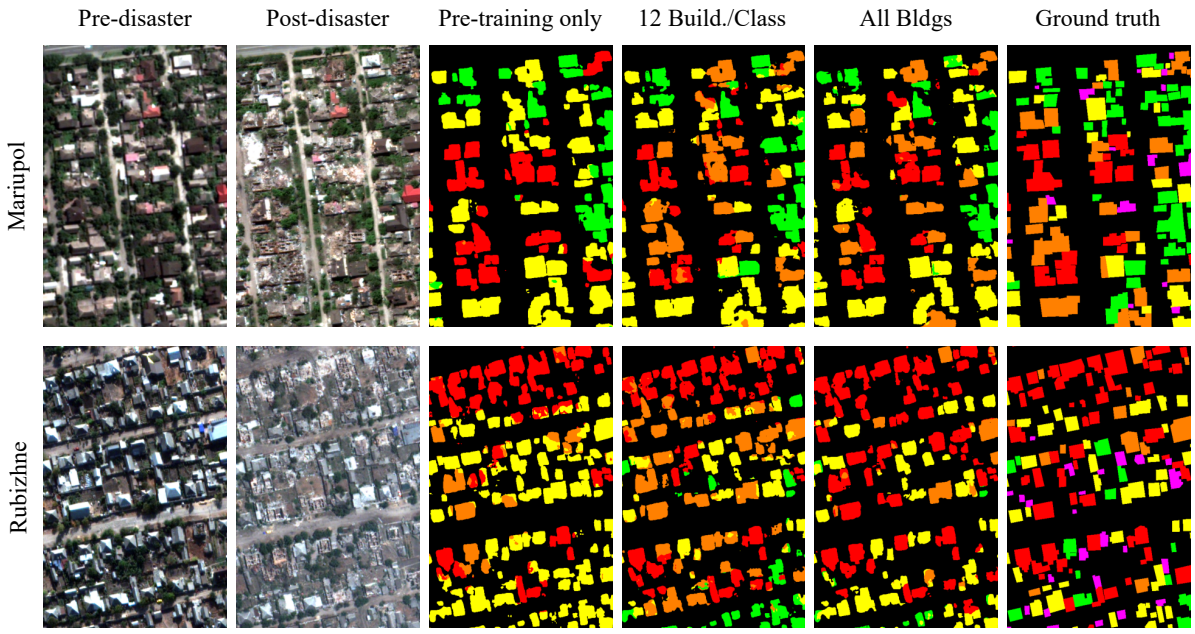


Figure 6. Qualitative results for the in-domain generalization experiment on the Ukraine building damage assessment dataset. Fine-tuning results are shown for 12 buildings per class (30 min annotation time budget) and for all buildings in the training set. Four-class building damage color coding: no damage, minor damage, major damage and destroyed. Unclassified buildings are shown as . Image sources: Mariupol: pre: 21.06.2021, WorldView-2, © 2021 European Space Imaging; post: 12.05.2022, WorldView-3, supplied by European Space Imaging © 2022 Vantor. Rubizhne: pre: 13.10.2020, WorldView-3, supplied by European Space Imaging © 2020 Vantor; post: 02.07.2022, WorldView-3, supplied by European Space Imaging © 2022 Vantor.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-539-2026 | © Author(s) 2026. CC BY 4.0 License.

Region	Fine-tuning # Build./Class	F_1^{comb}	F_1^{clf}	F_1^{loc}	Damage F1-scores No damage	Minor	Major	Destroyed	Combined
4 damage classes									
Bakhmut	✗ (0)	0.39	0.21	**0.80**	0.08	0.41	0.35	**0.58**	-
	✓ (12)	0.49	0.39	0.73	0.24	0.50	0.51	0.50	-
	✓ (24)	0.56	0.47	0.77	0.40	0.49	0.49	0.50	-
	✓ (all)	**0.59**	**0.52**	0.73	**0.43**	**0.53**	**0.60**	0.57	-
Hostomel-Irpin	✗ (0)	0.36	0.18	0.79	0.44	**0.31**	0.09	0.18	-
	✓ (all)	**0.50**	**0.37**	**0.81**	**0.86**	0.26	**0.70**	**0.23**	-
Lysychansk	✗ (0)	0.31	0.16	0.65	**0.81**	0.13	0.07	0.57	-
	✓ (all)	**0.56**	**0.48**	**0.75**	0.71	**0.40**	**0.34**	**0.67**	-
Mariupol	✗ (0)	0.31	0.13	0.75	0.54	0.25	0.04	0.62	-
	✓ (12)	0.58	0.49	**0.81**	0.63	0.41	0.37	0.64	-
	✓ (all)	**0.65**	**0.60**	0.79	**0.70**	**0.62**	**0.45**	**0.70**	-
Rubizhne	✗ (0)	0.53	0.42	0.78	0.42	0.41	0.31	0.69	-
	✓ (12)	0.58	0.49	0.78	0.50	0.43	0.42	0.69	-
	✓ (24)	0.63	0.55	0.79	**0.51**	0.52	0.50	0.73	-
	✓ (all)	**0.67**	**0.61**	**0.80**	0.49	**0.65**	**0.56**	**0.81**	-
Sievierodonetsk	✗ (0)	0.44	0.29	**0.78**	0.41	0.26	0.33	0.22	-
	✓ (12)	0.54	0.48	0.66	0.61	0.57	0.35	0.48	-
	✓ (24)	0.54	0.48	0.68	0.52	0.63	0.47	0.37	-
	✓ (all)	**0.69**	**0.68**	0.72	**0.62**	**0.83**	**0.71**	**0.60**	-
3 damage classes: minor + major combined									
Bakhmut	✗ (0)	0.38	0.20	**0.80**	0.08	-	-	**0.58**	0.78
	✓ (12)	0.50	0.40	0.73	0.24	-	-	0.50	0.77
	✓ (24)	0.60	0.52	0.77	0.40	-	-	0.50	**0.84**
	✓ (all)	**0.62**	**0.57**	0.73	**0.43**	-	-	0.57	0.82
Hostomel-Irpin	✗ (0)	0.45	0.30	0.79	0.44	-	-	0.18	0.47
	✓ (all)	**0.55**	**0.43**	**0.81**	**0.86**	-	-	**0.23**	**0.74**
Lysychansk	✗ (0)	0.55	0.50	0.65	**0.81**	-	-	0.57	0.33
	✓ (all)	**0.69**	**0.66**	**0.75**	0.71	-	-	**0.67**	**0.60**
Mariupol	✗ (0)	0.58	0.51	0.75	0.54	-	-	0.62	0.42
	✓ (12)	0.71	0.67	**0.81**	0.63	-	-	0.64	0.76
	✓ (all)	**0.75**	**0.73**	0.79	**0.70**	-	-	**0.70**	**0.82**
Rubizhne	✗ (0)	0.59	0.51	0.78	0.42	-	-	0.69	0.48
	✓ (12)	0.64	0.59	0.78	0.50	-	-	0.69	0.61
	✓ (24)	0.68	0.63	0.79	**0.51**	-	-	0.73	0.68
	✓ (all)	**0.70**	**0.65**	**0.80**	0.49	-	-	**0.81**	**0.76**
Sievierodonetsk	✗ (0)	0.47	0.33	**0.78**	0.41	-	-	0.22	0.48
	✓ (12)	0.63	0.62	0.66	0.61	-	-	0.48	**0.90**
	✓ (24)	0.57	0.52	0.68	0.52	-	-	0.37	0.86
	✓ (all)	**0.69**	**0.68**	0.72	**0.62**	-	-	**0.60**	**0.90**

Table 3. In-domain generalization results on Ukraine dataset using 50 % train-test split across six regions. Models are fine-tuned on annotation budgets of 12, 24, and all available buildings per damage class. Both F_1^{comb} and F_1^{clf} increase continuously with annotation budget across all areas. Best values are in bold.

Bouchard, I., Rancourt, M.-È., Aloise, D., Kalaitzis, F., 2022. On Transfer Learning for Building Damage Assessment from Satellite Imagery in Emergency Contexts. *Remote Sensing*, 14(11).

Chen, H., Nemni, E., Vallecorsa, S., Li, X., Wu, C., Bromley, L., 2022. Dual-Tasks Siamese Transformer Framework for Building Damage Assessment. *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, 1600–1603.

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

Chen, H., Song, J., Han, C., Xia, J., Yokoya, N., 2024. ChangeMamba: Remote Sensing Change Detection With Spatiotemporal State Space Model. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–20.

CRED, 2025. 2024 Disasters in numbers. Technical report, Centre for Research on the Epidemiology of Disasters (CRED).

d’Angelo, P., 2013. Automatic Orientation of large multitemporal Satellite Image Blocks. *Proceedings of International Symposium on Satellite Mapping Technology and Application 2013*, 1–6.

Dietrich, O., Peters, T., Sainte Fare Garnot, V., Sticher, V., Ton-That Whelan, T., Schindler, K., Wegner, J. D., 2025. An Open-Source Tool for Mapping War Destruction at Scale in Ukraine Using Sentinel-1 Time Series. *Communications Earth & Environment*, 6(1), 215.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D.,

Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Gu, A., Dao, T., 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., Gaston, M., 2019. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Gupta, R., Shah, M., 2020. RescueNet: Joint Building Segmentation and Damage Assessment from Satellite Imagery. *Proceedings - International Conference on Pattern Recognition*, 4405–4411.

Hertel, V., Geiß, C., Wieland, M., Taubenböck, H., 2025. Rapid Domain Adaptation for Disaster Impact Assessment: Remote Sensing of Building Damage after the 2021 Germany Floods. *Science of Remote Sensing*, 12, 100287.

Holail, S., Saleh, T., Xiao, X., Xiao, J., Xia, G.-S., Shao, Z., Wang, M., Gong, J., Li, D., 2024. Time-Series Satellite Remote Sensing Reveals Gradually Increasing War Damage in the Gaza Strip. *National Science Review*, 11(9), nwae304.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*.

Khvedchenya, E., Gabruseva, T., 2021. Fully Convolutional Siamese Neural Networks for Buildings Damage Assessment from Satellite Images. *NeurIPS Workshop on Artificial Intelligence for Humanitarian Assistance and Disaster Response (AI+HADR)*.

Liu, Y., Yue, J., Xia, S., Ghamisi, P., Xie, W., Fang, L., 2024. Diffusion Models Meet Remote Sensing: Principles, Methods, and Perspectives. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–22.

Lu, W., Wei, L., Nguyen, M., 2024. Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 4917–4935.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F., 2024. Deep-Image-Matching: A Toolbox for Multiview Image Matching of Complex Scenarios. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W4-2024, 309–316.

Nex, F., Duarte, D., Tonolo, F. G., Kerle, N., 2019. Structural Building Damage Detection with Deep Learning: Assessment of a State-of-the-Art CNN in Operational Conditions. *Remote Sensing*, 11(23).

Risso, M., Goffi, A., Motetti, B. A., Burrello, A., Bove, J. B., Macii, E., Poncino, M., Pagliari, D. J., Maffeis, G., 2025. Building Damage Assessment in Conflict Zones: A Deep Learning Approach Using Geospatial Sub-Meter Resolution Data. *2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS)*, CFP2540Z-ART, 1–6.

Robinson, C., Nsutezo, S. F., Ortiz, A., Sederholm, T., Dodhia, R., Birge, C., Richards, K., Pitcher, K., Duarte, P., Lavista Ferres, J. M., 2023. Rapid building damage assessment workflow: An implementation for the 2023 Rolling Fork, Mississippi tornado event. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3762–3766.

Rodger, M., Guida, R., 2023. Damage Assessment Mapping in Mariupol (Ukraine) with Multi-Temporal Synthetic Aperture Radar (SAR). *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 7186–7189.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Springer International Publishing, 234–241.

Toksabay, E., Butler, D., 2023. Turkey widens probe into building collapses as quake toll exceeds 50,000. https://www.reuters.com/world/middle-east/turkey-widens-probe-into-building-collapses-quake-toll-exceeds-50000-2023-02-25/.

Valentijn, T., Margutti, J., van den Homberg, M., Laaksonen, J., 2020. Multi-Hazard and Spatial Transferability of a CNN for Automated Building Damage Assessment. *Remote Sensing*, 12(17).

Wiguna, S., Adriano, B., Mas, E., Koshimura, S., 2024. Evaluation of Deep Learning Models for Building Damage Mapping in Emergency Response Settings. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 5651–5667.

Xiao, W., Su, J., Chen, Y., Cao, G., 2023. Cross-Scale-Guided Fusion Transformer for Disaster Assessment Using Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–12.

Yang, W., Zhang, X., Luo, P., 2021. Transferability of Convolutional Neural Network Models for Identifying Damaged Buildings Due to Earthquake. *Remote Sensing*, 13(3).

Yoshimura, M., Hayashi, T., Maeda, Y., 2025. MambaPEFT: Exploring Parameter-Efficient Fine-Tuning for Mamba. *The Thirteenth International Conference on Learning Representations*.

Yuan, X., Azimi, S. M., Henry, C., Gstaiger, V., Codastefano, M., Manalili, M., Cairo, S., Modugno, S., Wieland, M., Schneibel, A., Merkle, N., 2021. Automated Building Segmentation and Damage Assessment from Satellite Images for Disaster Relief. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43(B3-2021), 741–748.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building Damage Assessment for Rapid Disaster Response with a Deep Object-Based Semantic Change Detection Framework: From Natural Disasters to Man-Made Disasters. *Remote Sensing of Environment*, 265, 112636.