Optimizing Latent Representations for Robust Building Damage Assessment Onboard Earth Observation Satellites

Anonymous CVPR submission

Paper ID 9

Abstract

Rapid identification of damaged buildings after natural disasters or on war areas is crucial to support emergency response and prioritize interventions. Earth Observation constellations provide timely, large-scale coverage, but actionable information is often delayed by data downlink constraints, on-ground processing, and human interpretation. Reducing this latency is essential to improve decision-making responsiveness. In this work, we propose an original AI-based system that enables object-level building damage assessment (localization and damage classification) directly onboard satellites from pre-disaster and post-disaster high-resolution optical imagery. Available pre-disaster images are encoded on ground into compact latent representations, transmitted to the satellite, and compared on-board with newly acquired post-event observations. Leveraging AI interpretation capabilities and increasing processing capabilities on-board satellites, the proposed design enables processing directly at the data source, reducing the amount of information to be downlinked while preserving task-relevant content and improving overall system responsivity. We explore the design space through a systematic benchmark of on-board-compatible variants, analyzing the impact of siamese processing, cross-attention, latent-space compression, and robustness-oriented data augmentation. Experiments on xBD dataset demonstrate reliable and robust damage assessment under misalignment, with minimal performance degradation under strong compression.

1. Introduction

Timely assessment of building damage after natural disasters or armed conflicts is critical to support emergency response, allocate limited resources, and prioritize field interventions. In the first hours following a catastrophic event, situational awareness over affected urban areas is essential, yet ground-based inspections are often slow, hazardous, and logistically constrained [9]. Rapidly localizing buildings and estimating their damage level therefore plays a central role in operational disaster management.

High-resolution (HR) Earth Observation (EO) satellites provide large-scale, repeatable coverage of impacted regions and have become a key source of post-disaster information. In parallel, advances in deep learning have significantly improved automated building damage assessment from satellite imagery [9, 22], enabling localization and classification of damaged structures over wide areas.

Most high-performing methods formulate building damage assessment as a change detection or bi-image problem, jointly exploiting pre- and post-disaster imagery to reduce ambiguities between newly destroyed and pre-existing structures [17, 25]. Although robust, such pipelines remain largely ground-centric, requiring full image downlink and offline processing before actionable products are generated [18], resulting in operational latency.

Recent advances in intelligent satellite systems and embedded AI accelerators enable moving part of the processing closer to the sensor [24]. Onboard object detection has been demonstrated for several EO tasks [12, 18], confirming the feasibility of embedded deep learning in orbit. Extending this concept to bi-temporal building damage assessment, however, introduces additional challenges including limited uplink bandwidth for transmission of pre-disaster reference data, restricted onboard memory and compute resources, robustness to imperfect pre-/post-disaster image co-registration, and compatibility with raw or minimally processed data.

In this work, we propose a detection-based building damage assessment architecture from pre- and post-disaster HR optical EO imagery, specifically designed for embedded deployment. Our approach separates pre- and post-disaster processing while preserving the benefits of bi-temporal reasoning. Pre-disaster imagery is encoded on the ground into compact latent representations that are uplinked and stored on board. Newly acquired post-disaster images are processed directly in orbit and compared with the stored latent features to perform object-level damage assessment, substantially reducing communication requirements.

Contributions. (i) We introduce an innovative ground/onboard architecture for bi-temporal building damage assessment that decouples pre- and post-event processing to overcome ground-to-space communication constraints. (ii) We provide a systematic ablation study of design choices—including early fusion versus siamese processing, feature-level cross-attention, and latent-space compression. (iii) We analyze robustness and deployability under realistic operational conditions, including severe pre/post misregistration and bandwidth constraints in uplink and downlink volumes.

We next position our approach with respect to prior work in bi-temporal change detection, attention-based architectures, object-level damage assessment, and intelligent onboard EO systems.

2. Related Work

To position our contribution, we review prior work along four complementary axes: (i) bi-temporal formulations for building damage assessment and siamese change detection paradigms, (ii) the evolution from dense pixel-level segmentation to object-level damage modeling, (iii) transformer-based strategies for multi-temporal feature interaction, and (iv) onboard and embedded change detection systems. This structured analysis clarifies how methodological advances in multi-temporal reasoning and object-level representation have progressed, while highlighting the remaining gap between high-performing ground-based approaches and operationally deployable embedded systems.

Bi-Temporal and Siamese Change Detection. Building damage assessment from EO imagery is commonly formulated as a bi-temporal change detection problem [1]. In disaster assessment scenarios, the use of paired pre- and post-event imagery is frequently adopted, as relying solely on post-disaster observations may cause confusion between damaged buildings and background structures [17]. Existing methods are implemented either as cascade pipelines, where building localization and damage classification are performed in successive stages, or as unified end-to-end architectures. Cascade formulations often suffer from a knowledge gap between localization and classification stages, as the two components are optimized separately, limiting joint feature learning [25]. To reduce this limitation, fully convolutional siamese networks, introduced for remote sensing change detection by Daudt et al. [2], process pre- and post-event images through weight-sharing encoders, enabling temporally consistent feature extraction and direct comparison in a shared representation space.

From Dense Segmentation to Object-Level Damage Detection. While bi-temporal architectures improve feature alignment and change identification, most building damage assessment methods are formulated as dense pixel-level segmentation tasks, as exemplified by the xBD benchmark [9]. However, pixel-wise modeling may lead to semantic inconsistencies at the building scale, especially when damage affects only part of a structure, resulting in mixed predictions within a single instance [25]. Object-based approaches instead consider buildings as the fundamental unit of analysis rather than individual pixels, enabling more coherent semantic predictions and reducing noise and spurious detections [10]. In this formulation, damage assessment can be cast as an object-level detection problem, where buildings are localized and assigned a single damage state. Although originally developed for mono-temporal scenarios, standard object detection frameworks such as SSD [11], Faster R-CNN [15], or YOLO [14] provide suitable architectural backbones for such modeling when extended to multi-temporal inputs.

Transformer-Based Multi-Temporal Modeling. Transformer architectures based on self-attention have demonstrated strong ability to model long-range dependencies [20]. Their adaptation to visual tasks through Vision Transformers confirmed the effectiveness of global attention for capturing rich spatial context [5]. Recent works have incorporated transformer-based designs into bi-temporal remote sensing pipelines, including siamese transformer architectures that leverage attention mechanisms to enhance feature interaction between pre- and post-event imagery [13], as well as multi-scale attention frameworks specifically designed for building damage assessment [3]. These approaches highlight the importance of global contextual reasoning for multi-temporal modeling, particularly in complex urban scenes where long-range spatial dependencies can provide critical cues for damage interpretation.

Onboard Deployment and Operational Constraints. The integration of embedded AI within EO satellites enables onboard processing to reduce latency and communication load [24]. Deep learning models have been deployed in orbit, including convolutional change detection systems [16] and object detection frameworks adapted to spaceborne hardware [8, 18], while system-level demonstrators have explored reactive ground–space architectures [6] that shorten the decision–action loop. However, most existing approaches address either mono-temporal detection or pixel-level change detection, without jointly considering bi-temporal reasoning, object-level damage modeling, and the practical constraints of embedded deployment. In realistic scenarios, coping with limited onboard memory and compute resources, uplink/downlink bandwidth restrictions, residual pre/post co-registration errors, and the handling of raw sensor data remain critical [4]. The explicit integration of pre-event reference information in embedded multi-temporal pipelines is still largely underexplored.

Summary and positioning. Enabling efficient multi-temporal building detection under strict bandwidth, compute,

and memory limits, as well as robustness to onboard raw data inaccuracies (e.g., geolocation uncertainty), remains an open operational challenge. Addressing this gap requires the development of new AI-based architectures capable of localizing and classifying individual damaged buildings from two or more temporally distinct images, while explicitly accounting for ground–space communication constraints and embedded deployment limitations. Our work makes a step in this direction.

3. Method

To address this gap, we introduce a ground/on-board architecture specifically designed for object-level building damage assessment under realistic operational constraints.

3.1. Ground/On-board System Concept

We consider an operational Earth Observation scenario in which rapid assessment of building damage is required to support post-disaster response. In this setting, algorithmic performance and system-level constraints must be addressed jointly.

From an algorithmic standpoint, we adopt a bi-temporal architecture taking as input pre-/post-disaster images, consistently with prior work [17] showing its superior performance over post-event analysis alone. From an operational standpoint, improving responsiveness requires moving inference closer to the sensor [24], thereby shortening the decision loop and reducing downlink data volume with respect to systematic full-resolution imagery. However, this system-level shift introduces strict constraints: uplink bandwidth and onboard memory prevent transmitting complete pre-disaster reference scenes, while limited onboard geolocation accuracy induces residual misregistration between pre- and post-event observations.

To reconcile these constraints, pre-disaster scenes are encoded on the ground into compact latent representations that are uploaded and stored onboard. These representations are designed to (i) significantly reduce uplink volume and (ii) improve robustness to onboard data imperfections, particularly residual misregistration due to limited geolocation accuracy. When post-disaster imagery is acquired, the onboard module compares it with the stored latent features to perform object-level building damage assessment.

The onboard module outputs concise products consisting of localized building instances (bounding boxes) and associated damage levels. Optionally, image crops over detected areas of interest can be generated. By transmitting only structured detection results and targeted patches rather than full scenes, the proposed concept substantially reduces downlink volume while preserving actionable information. The overall processing concept is illustrated in Fig. 1.

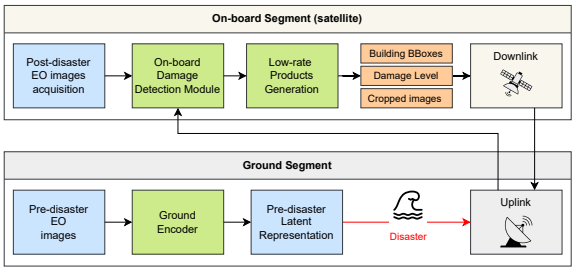


Figure 1. Ground/on-board concept. Pre-disaster scenes are encoded on the ground into compact latents uplinked to the satellite. Post-disaster images are compared on-board to detect damage and generate low-rate products.

3.2. Base Detection Architecture

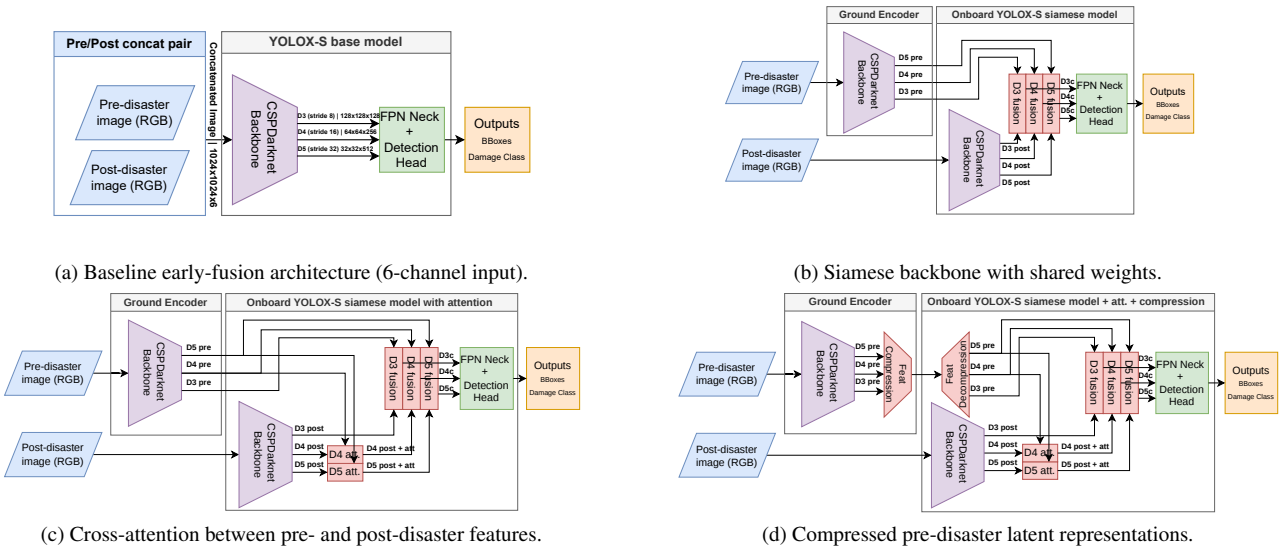
Building damage assessment is formulated as an object detection task, where buildings are localized and assigned a discrete damage level. To satisfy latency, memory, and implementation constraints associated with onboard deployment, we adopt a single-stage detection architecture as the common backbone for all experiments.

Our base detector follows the standard design of modern single-stage detectors and consists of a convolutional backbone, a feature pyramid network (FPN), and a detection head. In practice, we instantiate this architecture using YOLOX-S [7], selected for its favorable trade-off between detection accuracy and inference speed, as well as its compatibility with embedded processing targets [21]. The detection framework itself is not a contribution of this work and serves as a controlled reference architecture.

The CSPDarknet backbone produces three intermediate feature maps (D3, D4, D5) at spatial strides of 8, 16, and 32 pixels. These multi-scale representations define the feature space in which pre- and post-disaster information is combined and compared. The FPN aggregates these features across scales, and a shared detection head predicts bounding boxes and damage classes at each level. The FPN and detection head remain architecturally unchanged across all variants, although all components are trained end-to-end.

As illustrated in Fig. 2a, the baseline configuration performs early fusion by channel-wise concatenation of pre- and post-disaster images, resulting in a 6-channel input to the backbone. All architectural variants introduced in the next section—including siamese processing, cross-attention mechanisms, and latent-space compression—operate on the backbone feature maps (D3–D5) while preserving the same detection pipeline.

This early-fusion configuration assumes joint availability of pre- and post-disaster imagery at inference time, without accounting for high-bandwidth communication constraints. It therefore serves as a reference model for quantifying the impact of the proposed deployment-oriented modifications.



(a) Baseline early-fusion architecture (6-channel input).

(b) Siamese backbone with shared weights.

(c) Cross-attention between pre- and post-disaster features.

(d) Compressed pre-disaster latent representations.

Figure 2. Overview of the baseline architecture and the proposed architectural variants. All variants share the same backbone, neck and head but differ in the processing of pre- and post-disaster latent features.

3.3. Architectural Variants

The following section introduces a set of architectural variants that progressively relax system constraints and enable controlled evaluation of on-board–compatible processing strategies.

3.3.1. Siamese Processing and Feature Fusion

We extend the baseline detector with a siamese formulation inspired by bi-temporal change detection paradigms. Instead of fusing pre- and post-disaster imagery at the input level, the two images are processed independently by identical backbones with shared weights (Fig. 2b). The shared-weight design ensures that features extracted on both images lie in a common representation space, enabling consistent comparison while keeping the number of parameters unchanged.

At each backbone stage (D3, D4, D5), feature maps from the pre- and post-disaster branches are fused prior to the FPN. For each stage $n \in \{3, 4, 5\}$, we concatenate the pre-disaster features, the post-disaster features, and their element-wise difference to form the fused representation D_{nc}. This formulation exposes both absolute contextual information and explicit change cues to the detection pipeline.

Beyond improved multi-temporal reasoning, this siamese structure naturally supports the ground/on-board separation introduced in Section 3.1. Pre-disaster feature maps can be computed on the ground and reused onboard, enabling later extensions where only compact latent representations are uplinked.

3.3.2. Cross-Attention between Pre and Post Features

In operational onboard settings, residual spatial misregistration between pre- and post-disaster images is unavoidable due to acquisition geometry differences and limited onboard registration capabilities. Under such conditions, strictly local feature comparison may fail to retrieve relevant pre-disaster information at the correct spatial locations. To mitigate this issue, we introduce cross-attention mechanisms that enable spatially adaptive interaction between pre- and post-disaster feature maps.

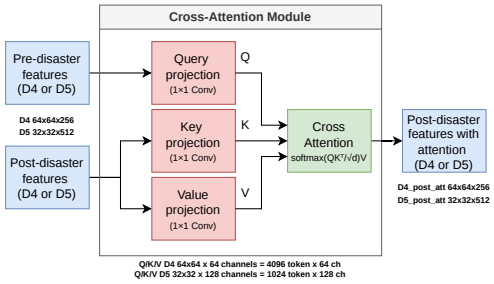


Figure 3. Principle of cross-attention between pre- and post-disaster features. Pre-disaster features act as queries to retrieve relevant contextual information from post-disaster features (keys and values), enabling robust comparison under spatial misregistration before feature fusion.

We implement an asymmetric cross-attention module following the query–key–value formulation introduced by Vaswani et al. [20]. In our design, pre-disaster backbone features act as queries, while post-disaster features provide keys and values (Fig. 3). Query, key, and value embeddings are obtained through 1×1 convolutions applied to the backbone feature maps. The 1×1 projections also reduce the channel dimensionality by a factor of four, limiting the computational cost of the attention operation while preserving sufficient representational capacity. This mechanism allows each spatial location in the pre-disaster representation to

attend over the entire post-disaster feature map and retrieve a weighted combination of features from potentially shifted locations, without requiring explicit geometric alignment.

The resulting attention output is used to refine the post-disaster representation prior to the feature fusion described in Sec. 3.3.1. The integration of the cross-attention module within the overall siamese ground/on-board architecture is illustrated in Fig. 2c. This refinement enables more robust comparison through subsequent concatenation and differencing.

To balance robustness and computational efficiency, cross-attention is applied only to higher-level feature maps (D4 and D5), where spatial resolution is lower and semantic abstraction is higher. The finest-scale features (D3) are excluded in the default configuration, as they encode low-level spatial details and would incur a significantly higher computational cost due to the quadratic scaling of attention with spatial resolution. The impact of applying attention at different feature levels is analyzed experimentally in Sec. 4.

3.3.3. Compressed Pre-Disaster Latent Representations

In the proposed ground/on-board system, pre-disaster information is represented by multi-scale backbone feature maps computed on the ground and uplinked for later comparison with post-disaster observations. Although this latent representation enables direct feature-level comparison onboard, its raw size can exceed that of the original input image. For a 1024×1024 input, the uncompressed feature maps (D3, D4, D5) amount to approximately 3.5 MB, compared to 3.0 MB for the RGB image, making uplink compression a critical system-level requirement.

To address this issue, we apply channel-wise compression prior to uplink. At each scale, the backbone features are projected into a lower-dimensional space through a lightweight 1×1 convolution that reduces the number of channels by a factor r, with $r \in \{8, 64\}$ in our experiments. This operation preserves the original spatial resolution of each feature map while significantly reducing memory footprint. The compressed feature maps are uplinked and, upon reception onboard, expanded through a symmetric lightweight projection trained jointly with the overall detection network to match the channel dimensionality expected by the fusion pipeline (Fig. 2d).

In addition to channel reduction, we explore configurations that remove the finest-resolution feature level (D3) from both pre- and post-disaster branches. Since D3 primarily captures fine-grained spatial details (e.g., edges and textures) and contributes significantly to memory and computation cost, its removal further reduces uplink volume and onboard processing requirements. The quantitative impact of this design choice is analyzed in Sec. 4.3.

Tab. 1 summarizes the resulting memory footprint of the pre-disaster latent representation under different compression configurations, reported as equivalent int8 sizes to reflect quantized onboard inference targets. The quantitative effects of these compression strategies on detection accuracy and robustness are evaluated in Sec. 4.3.

Table 1. Size of pre-disaster latent representations under different channel compression configurations for an input image of size $1024 \times 1024 \times 3$. Sizes are reported as equivalent int8 memory footprint.

Configuration	D3	D4	D5	Size (MB)
Input image (1024×1024×3)	–	–	–	3.15
Latent (no comp.)	128	256	512	3.67
Latent (comp. r=8)	16	32	64	0.46
Latent (comp. r=64)	2	4	8	0.057
Latent (comp. r=64, no D3)	–	4	8	0.025

Channel counts per feature map; spatial resolutions: 128×128 (D3), 64×64 (D4), and 32×32 (D5) for 1024×1024 input.

4. Experiments

We experimentally evaluate the design choices introduced in Sec. 3, including early pre/post fusion, siamese processing, cross-attention mechanisms, and compression of pre-disaster latent representations. All configurations are trained and evaluated under a unified protocol, enabling controlled comparisons in terms of detection accuracy, robustness to misregistration, and system-level trade-offs.

4.1. Dataset and Detection Setup

We conduct all experiments on the xBD dataset [9], introduced as part of the xView2 Challenge. The dataset provides pairs of coarsely registered pre- and post-event RGB satellite images for large-scale assessment of building damage following natural disasters.

Each sample consists of a pair of registered RGB images with a spatial resolution below 0.8 m/px and a size of 1024×1024 pixels. Buildings are annotated with four discrete damage levels: *no damage*, *minor damage*, *major damage*, and *destroyed*. The dataset covers a wide range of disaster types, including earthquakes, tsunamis, volcanic eruptions, wildfires, wind-related events, and floods, resulting in significant variability in visual appearance and acquisition conditions.

Although xBD was originally designed for pixel-wise segmentation, we reformulate the task as object detection by converting building polygons provided in the GeoJSON annotations into axis-aligned bounding boxes. All evaluations are therefore performed in a detection setting.

Following the standard xBD split (excluding Tier 3 data), we use 4,665 image pairs partitioned into 60% training, 20% validation, and 20% test sets. This fixed split is used consistently across all architectural variants to ensure fair comparison.

4.2. Training and Evaluation Protocol

All experiments are conducted using a YOLOX-S backbone [7] unless explicitly stated otherwise. All architectural variants are trained independently under identical optimization settings to enable controlled comparisons.

Post-Disaster Shift Augmentation. To improve robustness to residual pre/post misregistration expected in onboard scenarios, we introduce a post-disaster shift augmentation (Aug) during training. A random spatial displacement is applied exclusively to the post-disaster image, with independent horizontal and vertical offsets uniformly sampled in $[0, 150]$ pixels. Ground-truth annotations remain defined in the reference frame of the pre-disaster image. After shifting, only the spatial intersection between the pre- and post-disaster images is retained. Pixels outside this shared region and bounding boxes falling partially or fully outside it are discarded to preserve annotation consistency.

Evaluation and Test-Time Misregistration Protocol. Detection performance is evaluated using precision, recall, F1-score, and mAP@0.5. For robustness experiments, synthetic shifts are applied at test time to the post-disaster image only. Ground-truth annotations remain unchanged and are always defined in the reference frame of the pre-disaster image. To ensure fair comparison across shift magnitudes and directions, evaluation is performed on a fixed spatial support corresponding to the intersection induced by the maximum considered displacement. Pixels outside this shared region, as well as bounding boxes falling partially or entirely outside it, are removed consistently for all evaluated shifts. This guarantees that all robustness results are computed on an identical image subset and annotation set, independently of the applied displacement.

4.3. Quantitative Results and Ablation Study

Tab. 2 reports quantitative results for the proposed architectural variants and reference baselines on the xBD dataset. We first compare detection-based approaches to segmentation-driven pipelines adapted for detection. The xView2 challenge winning solution, originally designed for semantic segmentation, exhibits substantially lower performance when adapted to object detection by extracting bounding boxes from segmentation masks. This performance gap should be interpreted with caution, as the original method was not optimized for instance-level detection, and the mask-to-bounding-box conversion can introduce additional localization errors, particularly for small or fragmented structures.

Using early fusion of pre- and post-disaster images as a baseline (EF), we observe that introducing siamese processing consistently improves performance, yielding a gain of approximately +1.5 points in F1-score and mAP. This confirms the benefit of explicitly separating pre- and post-disaster feature extraction prior to comparison, enabling more structured temporal reasoning.

Table 2. Quantitative comparison of architectural variants for onboard building damage detection on xBD dataset.

Cfg.	**P**	**R**	**F1**	**mAP@0.5**
xView2 winner†	50.4	52.1	51.2	38.8
EF (6ch)	64.1	62.2	63.1	64.3
S	66.5	63.5	65.0	65.6
S + A	64.9	61.5	63.2	64.1
S + A^{+D3}	60.8	55.9	58.2	56.0
S + Aug	**67.8**	**67.7**	**67.7**	**68.5**
S + A + Aug	64.1	63.7	63.9	64.7
S + A + Aug + Comp(8)	64.5	62.3	63.4	64.6
S + A + Aug + Comp(64)	65.6	61.1	63.3	64.3
S + A + Aug + Comp(64)$-D3$	60.1	55.8	57.9	56.3
YOLOv12 S + A + Aug	69.3	61.4	65.1	67.0

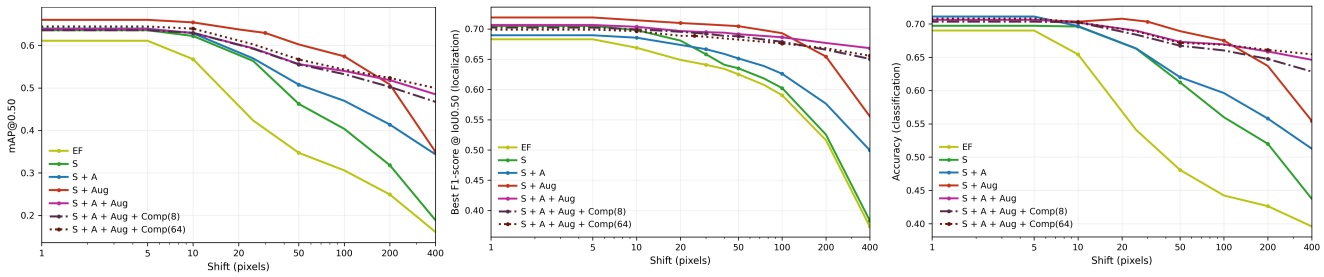
† Seg → Det adaptation. Acronyms: **EF**: early fusion; **S**: siamese processing; **A**: cross-attention on D4–D5; **A**$^{+D3}$: cross-attention extended to D3; **Aug**: post-disaster shift augmentation during training; **Comp**(r): channel-wise latent compression with ratio r; $-D3$: removal of the finest-resolution feature level from both branches.

Introducing cross-attention slightly decreases average detection performance in nominal conditions. In all reported results, attention is applied only at the coarser feature levels (D4–D5), as described in Sec. 3.3.2. Extending attention to the finest-resolution feature level (D3) leads to a strong degradation (approximately -8 mAP), supporting the design choice to exclude D3 from attention-based interactions.

From a training perspective, incorporating post-disaster shift augmentation (Aug) yields a significant improvement in average detection performance (+3% mAP), indicating that exposure to synthetic misregistration enhances generalization even under nominal alignment. While architectures without cross-attention achieve the highest mean accuracy, attention-based variants become advantageous under large pre/post misregistrations, as analyzed in Sec. 4.4.

We further evaluate system-level constraints through compression of pre-disaster latent representations. Channel-wise compression with ratios up to $r = 64$ induces only a marginal performance drop (< 1 mAP), while reducing the uplinked latent size from 3.7 MB to 57 kB (Tab. 1), corresponding to more than two orders of magnitude reduction. In contrast, removing the finest-resolution feature level (D3) results in a substantial accuracy degradation (approximately -8 mAP), highlighting the importance of high-resolution features for precise building localization.

Finally, we report results obtained with a more recent detection backbone (YOLOv12 [19]) as a reference point. This experiment aims to position the proposed design choices relative to a newer generation of detectors that already incorporate attention mechanisms. While YOLOv12 yields a moderate performance improvement compared to the YOLOX-S baseline, the gain remains limited, and the relative trends observed across architectural variants remain consistent. This



(a) mAP@0.5 vs. post-image shift (b) Localization F1-score vs. post-image shift (c) Classification accuracy vs. post-image shift

Figure 4. Robustness to pre/post misregistration on xBD. Performance as a function of the applied post-image shift magnitude. Results are averaged over four shift directions. Legend: **EF**: early fusion (6-channel input); **S**: siamese processing; **A**: cross-attention on D4–D5; **Aug**: post-disaster shift augmentation during training; **Comp**(r): channel-wise latent compression with ratio r.

suggests that the conclusions drawn from our design space analysis are not tied to a specific detector implementation.

4.4. On-board Robustness and Feasibility Analysis

We further analyze the proposed architectural variants along two complementary dimensions: robustness to pre/post image misregistration representative of on-board acquisition conditions, and system-level trade-offs related to communication, computation, and deployability on spaceborne platforms.

4.4.1. Robustness to Pre/Post Misregistration

In on-board EO scenarios, residual misregistration between pre- and post-disaster images is unavoidable due to acquisition geometry differences and limited co-registration capabilities. We evaluate robustness using the controlled test-time shift protocol described in Sec. 4.2, where synthetic displacements are applied to the post-disaster image while ground-truth annotations remain defined in the pre-disaster reference frame. Results are computed on the common spatial support defined by the maximum displacement.

Fig. 4a reports mAP@0.5 as a function of the misregistration magnitude. Architectures trained without Aug—whether early-fusion or siamese—exhibit a rapid performance drop for shifts on the order of a few tens of pixels, corresponding to typical building extents. In contrast, models trained with Aug show a markedly more gradual degradation. When the induced shift exceeds the network's effective receptive field, performance eventually collapses, as reliable spatial correspondence between pre- and post-disaster features can no longer be established. Cross-attention further mitigates degradation under large misregistration when combined with Aug. Cross-attention alone (without Aug) provides limited robustness gains, indicating that the ability to retrieve spatially displaced features must be learned during training to be effective at inference. To better understand the origin of performance degradation, we separately analyze localization and damage classification.

Localization performance (Fig. 4b) is evaluated as building detection irrespective of damage level, using the maximum F1-score optimized over the object confidence threshold at IoU ≥ 0.5. A prediction is considered correct if it matches a ground-truth building with sufficient overlap, independently of its damage label. Localization degrades with increasing shift, but much less than damage classification and overall mAP. This relative stability stems from the fact that bounding boxes are predicted in the pre-disaster reference frame and mainly rely on features from that image. In particular, architectures combining Aug and cross-attention are almost insensitive to misregistration, suggesting that they have learned a robust localization behavior under spatial shifts.

Damage classification performance (Fig. 4c) is evaluated only on correctly localized buildings and measured using standard classification accuracy over the four damage levels. In contrast to localization, damage classification is substantially more sensitive to misregistration and largely explains the observed mAP drop, since accurate class prediction requires consistent comparison between pre- and post-disaster features. Architectures combining Aug and cross-attention exhibit the slowest degradation in classification accuracy under large offsets, confirming their complementary roles. Importantly, channel-wise compression of pre-disaster latent representations does not modify the observed robustness trends for localization, classification, or overall mAP, indicating that aggressive data reduction remains compatible with misregistration tolerance.

4.4.2. On-board Feasibility Analysis

Beyond detection accuracy and robustness, we assess the feasibility of deploying the proposed pipeline under realistic on-board constraints in terms of computation and communication. While the present study does not target a specific mission profile, we provide order-of-magnitude estimates to position the proposed approach with respect to existing spaceborne processing capabilities.

Previous works have demonstrated the deployment of YOLO-based object detection pipelines on embedded FPGA platforms equipped with dedicated AI accelerators, including recent Xilinx Versal devices [6, 8]. Results reported

on latest-generation embedded targets indicate processing throughputs on the order of 10^2 MPixels/s for the neural inference stage alone, excluding image I/O and pre-processing. Assuming 1024×1024 image tiles at a ground sampling distance of approximately 0.8 m/px, this corresponds to processing roughly 100 tiles per second, i.e., about 70–80 km^2/s. Under these assumptions, a representative urban disaster area of 100 km^2 (roughly equivalent to the area of Paris proper) could be processed in approximately one to two seconds of pure inference time.

From a communication perspective, uplinking full-resolution pre-disaster images would require transmitting approximately 450 MB for a 100 km^2 area. In contrast, using the proposed compressed latent representation with a compression ratio of $r = 64$ reduces the uplinked information to less than 10 MB for the same area, corresponding to more than two orders of magnitude reduction.

In addition, on-board processing enables drastic reduction of required downlink datarate by transmitting compact detection products (building bounding boxes and damage levels) instead of full-resolution imagery, with optional transmission of image patches limited to detected damaged regions. Overall, these estimates indicate that the proposed architecture offers a favorable trade-off between detection performance, robustness, and on-board resource usage.

5. Discussion, Limitations, and Future Work

Our experiments highlight several key advances enabling resource-efficient bi-temporal damage assessment. Siamese processing improves detection performance over early fusion while naturally supporting the separation between ground-encoded pre-disaster features and on-board post-disaster inference. Robustness-oriented shift augmentation—introduced to simulate realistic on-board misregistration—not only strengthens robustness to spatial offsets but also improves average detection accuracy under nominal conditions. Controlled shift experiments show that localization remains relatively stable under misalignment, whereas damage classification is more sensitive. Cross-attention helps maintain performance under large spatial shifts through non-local feature retrieval, while latent compression up to $r = 64$ yields only marginal degradation and reduces uplink data volume by more than two orders of magnitude. Taken together, these findings represent a step toward the feasibility of bi-temporal object-level damage assessment under embedded inference and constrained ground-to-space bandwidth, using a ground/on-board latent encoding strategy, paving the way toward operational on-board deployment.

Several aspects must however be addressed to further increase system maturity. Current experiments rely on pre-processed imagery, whereas real on-board acquisitions involve raw data affected by sensor noise, radiometric variability, optical distortions, and residual geolocation errors [4]. Future work should evaluate robustness under realistic raw-data simulations and integrate lightweight preprocessing strategies. In addition, harmonization between pre- and post-event imagery remains challenging due to illumination and potential sensor differences; image uniformization strategies, combined with latent-space comparison, may help mitigate these discrepancies.

Beyond these maturity improvements, several exploratory directions can be considered. Leveraging multiple pre-disaster observations (reference time series) could produce more stable latent representations, improving localization robustness and reducing sensitivity to scene variability. Increasing the computational complexity of the ground segment—through larger encoders or additional contextual processing—while maintaining a lightweight on-board branch, offers a promising asymmetric architecture compatible with strict uplink budgets. Furthermore, although this study focuses on mono-modal optical imagery, the proposed ground/on-board architecture naturally extends to multi-sensor damage assessment. In operational contexts, heterogeneous observations (e.g., optical and Synthetic Aperture Radar) could enrich latent representations, although learning consistent cross-modal embeddings remains challenging and may require dedicated fusion strategies [23].

6. Conclusion

We introduced a unified ground/onboard architecture for object-level building damage assessment from bi-temporal satellite imagery, compatible with realistic satellite communication and compute constraints. The proposed approach works by first encoding, on ground, pre-disaster images into compact latent representations. These compressed representations are then used during post-disaster inference performed directly on board the satellite. By combining these two steps, this design enables efficient multi-temporal analysis close to the sensor while reducing the need to transmit large volumes of raw data.

Through a systematic design-space exploration, we showed that siamese processing and robustness-oriented data augmentation primarily improve average detection performance, while the combination of data augmentation and cross-attention mechanisms enhances robustness to severe pre/post misregistration. At the same time, aggressive latent-space compression achieves more than two orders of magnitude communication reduction with only marginal performance degradation, confirming the feasibility of embedded bi-temporal reasoning under stringent uplink constraints.

Beyond specific architectural choices, this work illustrates a shift in EO systems from ground-centric processing toward onboard applications that reduce communication load, improve responsiveness, and remain compatible with embedded hardware constraints.

References

[1] Giuseppina Andresini, Annalisa Appice, Dino Ienco, and Donato Malerba. Seneca: Change detection in optical imagery using siamese networks with active-transfer learning. *Expert Systems with Applications*, 214:119123, 2023. 2

[2] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067, 2018. 2

[3] Yifan Da, Zhiyuan Ji, and Yongsheng Zhou. Building damage assessment based on siamese hierarchical transformer framework. *Mathematics*, 10(11), 2022. 2

[4] Adrien Dorise, Marjorie Bellizzi, Adrien Girard, Benjamin Francesconi, and Stéphane May. Explaining raw data complexity to improve satellite on-board processing. In *2025 European Data Handling & Data Processing Conference (EDHPC)*, pages 1–8, 2025. 2, 8

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 2

[6] Benjamin Francesconi, Luis Palluel, Thomas Goudemant, Hugo Meleiro, Benjamin Marchand, Olivier Thiery, and France Creil. From event to action: A reactive loop demonstrator for earth observation based on modular ai-driven components. In *Actes de la 7ème Conference on Artificial Intelligence for Defense (CAID 2025)*, 2025. 2, 7

[7] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 6

[8] Thomas Goudemant, Benjamin Francesconi, Houssem Farhat, Lionel Daniel, Olivier Thiery, Erwann Kervennic, Adrien Girard, and Seif Mzoughi. Détection de navires embarquable à bord de satellites. In *Actes de la 4ème Conference on Artificial Intelligence for Defense (CAID 2022)*, Rennes, France, 2022. DGA Maîtrise de l'Information. 2, 7

[9] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. 1, 2, 5

[10] Huiwei Jiang, Min Peng, Yuanjun Zhong, Haofeng Xie, Zemin Hao, Jingming Lin, Xiaoli Ma, and Xiangyun Hu. A survey on deep learning-based change detection from high-resolution remote sensing images. *Remote Sensing*, 14(7), 2022. 2

[11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015. 2

[12] N Melega, Bernardo A Carnicero, N Longepe, A Paskeviciute, V Marchese, Oriol Aragon, Irina Babkina, Alessandro Marin, Jakub Nalepa, Leonie Buckley, et al. Implementation of the ϕsat-2 on board image processing chain. In *Sensors, systems, and next-generation satellites XXVII*, pages 264–276. SPIE, 2023. 1

[13] Amir Mohammadian and Foad Ghaderi. Siamixformer: a fully-transformer siamese network with temporal fusion for accurate building detection and change detection in bi-temporal remote sensing images. *International Journal of Remote Sensing*, 44(12):3660–3678, 2023. 2

[14] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. 2

[15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. 2

[16] Chahira Serief, Youcef Ghelamallah, and Youcef Bentoutou. Deep-learning-based system for change detection onboard earth observation small satellites. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:8115–8124, 2023. 2

[17] Jinyuan Shao, Lina Tang, Ming Liu, Guofan Shao, Lang Sun, and Quanyi Qiu. Bdd-net: A general protocol for mapping buildings damaged by a wide range of disasters based on satellite imagery. *Remote Sensing*, 12(10), 2020. 1, 2, 3

[18] Yanyun Shen, Di Liu, Junyi Chen, Zhipan Wang, Zhe Wang, and Qingling Zhang. On-board multi-class geospatial object detection based on convolutional neural network for high resolution remote sensing images. *Remote Sensing*, 15(16), 2023. 1, 2

[19] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors, 2025. 6

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. 2, 4

[21] Ling Wang, Hai Zhou, Chunjiang Bian, Kangning Jiang, and Xiaolei Cheng. Hardware acceleration and implementation of yolox-s for on-orbit fpga. *Electronics*, 11(21), 2022. 3

[22] Lili Wang, Jidong Wu, Youtian Yang, Rumei Tang, and Ru Ya. Deep learning models for hazard-damaged building detection using remote sensing datasets: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:15301–15318, 2024. 1

[23] Moyang Wang, Kun Tan, Xiuping Jia, Xue Wang, and Yu Chen. A deep siamese network with hybrid convolutional feature extraction module for change detection based on multi-sensor remote sensing images. *Remote Sensing*, 12(2), 2020. 8

[24] Bing Zhang, Yuanfeng Wu, Boya Zhao, Jocelyn Chanussot, Danfeng Hong, Jing Yao, and Lianru Gao. Progress and challenges in intelligent remote sensing satellite systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:1814–1822, 2022. 1, 2, 3

[25] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021. 1, 2