

Comparative analysis of vision-language models for infrastructure damage classification: a case study of Ukraine*

Oleksii Ivanov^{1,*}, Kyrylo Bieliaiev^{1,†}, Vadym Oliinyk^{1,†} and Olena Arsirii^{2,†}

¹ Odesa Polytechnic National University, Shevchenko Ave. 1, 65044, Odesa, Ukraine

² Odesa State Maritime University, Mechnikova Str., 34, 65029, Odesa, Ukraine

Abstract

The large-scale destruction of civil infrastructure caused by the war in Ukraine has created an urgent need for automated, reliable, and scalable methods for structural damage assessment. This study presents a comparative evaluation of state-of-the-art Vision-Language Models (VLMs) and a hybrid computer vision approach for multi-view infrastructure damage classification. A specialized dataset was constructed using ground-level images of war-damaged buildings collected from multiple regions of Ukraine, where each infrastructure object is represented by two to eight photographs, including pre- and post-damage views when available. Damage severity is categorized into five classes ranging from undamaged to completely destroyed. The proposed methodology integrates a multi-view reasoning framework with semantic segmentation based on a U-Net architecture employing a ResNet encoder and a rule-based heuristic model that estimates structural damage using pixel-level metrics. Multi-view evidence is aggregated to produce a single object-level assessment, reducing ambiguity caused by occlusions and limited viewing angles. The proposed approach was experimentally compared with several advanced multimodal VLMs, including GPT-5.5, Claude Sonnet 5, Gemini 3.1 Pro, and Grok 4, using Accuracy, Precision, Recall, F1-score, and overestimation/underestimation ratios. Experimental evaluation conducted on a dedicated dataset of 108 infrastructure objects demonstrates that advanced VLMs outperform the baseline segmentation model in overall classification accuracy while providing robust reasoning across complex structural damage scenarios. At the same time, the proposed hybrid segmentation-based framework offers interpretable pixel-level damage estimation and remains suitable for resource-constrained deployments. The developed multi-view assessment methodology constitutes a core analytical component of an intelligent multifunctional GIS designed to support construction and demolition waste inventory, optimize recycling logistics, and facilitate sustainable post-war reconstruction based on circular economy principles.

Keywords

Infrastructure damage assessment, Vision-Language Models, multi-view image analysis, semantic segmentation, construction and demolition waste, artificial intelligence, machine learning

1. Introduction

The scale of war-induced infrastructural destruction in Ukraine has generated unprecedented volumes of construction and demolition waste (CDW), requiring rapid, automated inventory systems to replace slow and hazardous manual assessments. This paper presents an intelligent image analysis framework developed as a core component of an intelligent multifunctional GIS for CDW inventory. To train and validate the system, we constructed a rare, highly specialized dataset comprising multi-view imagery (2 to 8 angles per object, including before-and-after states) of war-damaged infrastructure objects across Ukraine. The objects are classified into five distinct damage levels (0 – Undamaged to 4 – Destroyed) based on consistent visual criteria. We propose a hybrid classification method that integrates deep learning feature extraction via a U-Net architecture with a ResNet backbone and a rule-

* ADT-ISMDCC 2026: *Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control*, September 22–24, 2026, Odesa, Ukraine.

* Corresponding author.

† These authors contributed equally.

✉ lesa.ivanoff@gmail.com (O. Ivanov), kirillbelyaev2921@gmail.com (K. Bieliaiev); vadimol081@gmail.com (V. Oliinyk), elena.arsirii@gmail.com (O. Arsirii)

ORCID: 0000-0002-8620-974X (O. Ivanov), 0009-0001-7135-3562 (K. Bieliaiev), 0009-0000-5527-7750 (V. Oliinyk), 0000-0001-8660-6916 (O. Arsirii)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

based heuristic model. This heuristic model estimates pixel-volume damage ratios for three key regions of interest: roofing, structural/facade integrity, and glazing. The performance of this hybrid U-Net-ResNet-heuristic approach is compared against state-of-the-art Vision-Language Models (VLMs) using a multi-view aggregation strategy that determines the overall object status based on the maximum visible damage. The evaluation is conducted using standard metrics: Accuracy, Precision, Recall, F1-score and Overestimation/Underestimation ratio. The results demonstrate that combining deep learning segmentation with rule-based heuristics yields highly interpretable and accurate classifications. This automated damage assessment pipeline serves as a critical foundation for the intelligent multifunctional GIS platform, enabling optimized CDW recycling logistics and supporting circular economy principles during the sustainable reconstruction of Ukraine.

2. Literature overview

The integration of artificial intelligence (AI), computer vision, and geospatial technologies has recently emerged as a pivotal domain in sustainable construction and disaster recovery. This section reviews existing methodologies for building damage assessment and CDW management, structuring them into three core areas.

Deep learning in post-disaster damage assessment and CDW classification. Traditional methods of managing construction and demolition waste have heavily relied on manual inspections, which are inherently slow, labor-intensive, and prone to human error. To address these limitations, researchers have increasingly turned to machine learning. A comprehensive foundation for these approaches is provided by Samal et al. [1], who presented a critical review of various machine learning methods used for the estimation, classification, and prediction of CDW volumes, emphasizing their role in enabling sustainable waste management. At the scale of satellite remote sensing, the landmark work by Gupta et al. [2] introduced the xBD dataset, which remains a gold standard in change detection and building damage assessment. By providing pre- and post-event multi-band satellite imagery with over 700,000 building polygons and ordinal damage labels across 15 countries, xBD demonstrated the viability of passive, large-scale remote sensing to expedite disaster recovery and lower risk for first responders. In [3] author proposed a two-step machine learning pipeline for automated damage assessment that combines semantic segmentation for building detection with classification and regression models for severity estimation. By training on the xBD dataset and fine-tuning on custom data from Ukraine, the study demonstrates an effective, adaptable solution for rapid infrastructure monitoring in post-conflict urban environments. Building upon these predictive models, computer vision has been deployed to classify waste materials directly from visual data. For instance, Molchanova et al. [4] proposed a specialized method for CDW classification using a two-factor neural network image analysis. By utilizing robust convolutional neural network (CNN) architectures such as YOLO and ResNet, their approach demonstrates how deep learning can automate the identification of various waste categories from site imagery, laying the groundwork for automated sorting and recycling pipelines. Guo et al. [5] present a novel zero-shot crack segmentation framework for structural civil infrastructure monitoring that utilizes image-level supervision and the Segment Anything Model to eliminate the need for labor-intensive, pixel-level manual annotations. Zheng et al. [6] introduced ChangeOS, a deep object-based semantic change detection framework that integrates object localization with damage classification to enhance building damage assessment using bitemporal satellite imagery. By combining the strong feature representation of deep learning with the semantic consistency of object-based analysis, the proposed end-to-end model achieves superior accuracy and generalization in mapping damage from both natural disasters and man-made explosions. Duarte et al. [7] proposed three multi-resolution CNN feature fusion approaches to improve building damage assessment by integrating imagery from different spatial resolutions, such as satellite and aerial platforms. The study demonstrates that fusing feature maps from intermediate network layers enhances both classification accuracy and localization capabilities, offering a more robust alternative to traditional mono-resolution methods for post-disaster analysis.

Emerging VLMs and generative AI for visual diagnostics. While traditional CNNs excel at narrow classification and segmentation tasks, the rapid evolution of VLMs has introduced new paradigms. Generative AI and VLMs offer promising capabilities by understanding the context of images through natural language prompts and tokenization. Estêvão [8] evaluated the effectiveness of generative AI models for post-earthquake damage assessment, highlighting how these advanced systems can tokenize structural features to identify damage levels and classify material types with minimal task-specific training. The broader potential of these intelligent systems is further explored by Alimi et al. [9]. Their literature review examines how AI, machine learning, and natural language processing can be combined to optimize CDW management workflows, predict waste generation rates, and actively identify circular opportunities for material reuse and recycling. The transition to such multi-modal AI systems allows for more interpretable and flexible diagnostic tools in structural engineering.

War-induced destruction assessment and circular economy challenges. The visual patterns of damage resulting from military conflicts differ significantly from those caused by natural disasters like earthquakes or hurricanes. Military damage is characterized by localized ballistic impacts, intense fragmentation, concrete pulverization, and mixed debris. The study [10] proposes a comprehensive strategy for managing construction waste resulting from hostilities in Ukraine, with a focus on implementing circular economy principles and establishing a market for secondary materials. Addressing this specific challenge within the context of the war in Ukraine, a highly relevant project described by Aouf [11] demonstrates how edge-based AI can analyze drone footage captured over bombed buildings. This system successfully identifies reusable materials within the rubble, actively facilitating on-site circularity during immediate reconstruction efforts.

Aimaiti et al. [12] evaluate the effectiveness of medium-resolution Sentinel-1 (SAR) and Sentinel-2 (optical) satellite imagery for rapid building damage assessment in the Kyiv region during the ongoing war. By combining intensity change detection and texture analysis with built-up area masking, the authors demonstrate that open-source satellite data provides a reliable, scalable method for initial damage mapping where ground-level verification is limited.

To capture critical structural details that remain invisible in satellite imagery, ground-level building damage assessment has become highly relevant. A major contribution to this niche was made by Lysanets et al. [13] who proposed an Embedding-enhanced U-Net framework utilizing a patch-based approach and introduced the first war damage dataset for ground-level building segmentation. Their work targets side-view images of war-affected Ukrainian buildings with pixel-wise annotations across six classes (including Damaged Roof and Broken Window). By incorporating global ConvNeXt-Large embeddings into the classic U-Net encoder, they achieved substantial performance gains (reaching a 0.8982 F1-score with a ResNet-50 backbone), validating the combination of U-Net-based deep learning with advanced feature extraction for complex, real-world structural damage tasks.

This research builds upon the conceptual framework proposed in our previous work [14], which introduced an intelligent information technology for construction and demolition waste management. That study established the foundational GIS knowledge base integrating spatial and attribute data for a "Sinks-Sources" mapping system, utilizing LLMs to interpret visual cues and optimize transportation routes for secondary resources. Despite these advancements, a critical knowledge gap remains: most existing open-source datasets (such as xBD) focus on natural disasters. There is an urgent need for multi-view, high-resolution datasets reflecting actual war-induced structural damage to validate and refine AI models for real-world deployment in conflict zones.

3. Research objective statement

The objective of this study is to develop and experimentally evaluate a method for the automated classification of infrastructure damage severity based on multi-view image analysis using state-of-the-art VLMs. To achieve this, the following tasks are addressed:

- to create and prepare a specialized dataset of Ukrainian infrastructure objects represented by multi-view image sets (from 2 to 8 photographs per object);
- to develop a logical framework for an intelligent multi-view classification technology that includes data preparation, contextual image analysis using a multimodal model, result aggregation, and validation of the obtained predictions;
- to implement a multi-view damage assessment method that combines image analysis performed by Vision-Language Models with pixel-based characteristics of damaged structural components and a result integration algorithm across multiple viewpoints;
- to conduct a comparative experimental evaluation of state-of-the-art Vision-Language Models for infrastructure damage classification using the Accuracy, Precision, Recall, and F1-score metrics, as well as to assess the models' tendency to overestimate or underestimate the severity of damage.

4. Methodology

4.1. Multi-view AI-based damage assessment workflow

The proposed methodology is designed for automatic damage assessment of civil infrastructure objects using multimodal artificial intelligence. Unlike conventional image classification approaches, where each image is processed independently, the proposed workflow considers an infrastructure object as a single analytical entity represented by multiple photographs captured from different viewpoints. This multi-view strategy reduces the probability of incorrect assessment caused by partial visibility of structural damage and enables more reliable classification of the overall condition of the object. The overall flowchart of the proposed Intelligent building damage assessment information system workflow is illustrated in Figure 1.

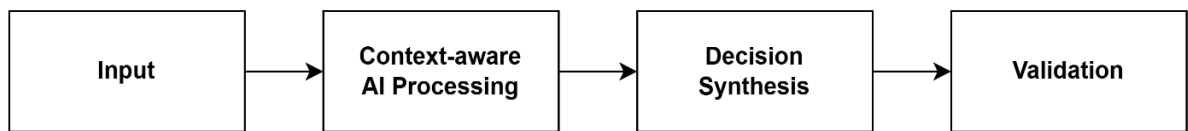


Figure 1: Flowchart of the AI-based building damage assessment workflow.

The first step of the workflow is input. The input to the system consists of a package of photographs representing the same infrastructure object. Each object is described by a set of k images captured from different viewpoints. The image package should include photographs acquired before and after the damaging event, allowing the model to compare structural changes over time.

The second step is context-aware AI Processing. Instead of analyzing each image independently, the complete package of photographs is simultaneously provided to a multimodal Vision-Language Model together with a task-specific prompt. The prompt instructs the model to jointly analyze all available views and identify the highest observed level of structural damage. By considering spatial consistency between viewpoints, the model can detect damage that may not be visible from a single camera angle.

The third step is decision synthesis. The multimodal model aggregates the visual evidence extracted from all viewpoints and produces a single damage label representing the overall condition of the infrastructure object. This process follows the principles of multi-view classification, where the final prediction is generated at the object level rather than at the individual image level.

The final step of the workflow is validation of the results. The predicted damage class is compared with the reference annotation (ground truth) assigned by human annotators. Model performance is evaluated using standard classification metrics, including Accuracy, Precision, Recall, and the F1-score.

4.2. Data aggregation and dataset preparation

The proposed methodology considers an infrastructure object, rather than an individual image, as the fundamental unit of analysis. Each object is represented by a collection of photographs captured from multiple viewpoints, enabling a comprehensive assessment of structural damage. This representation follows the principles of multi-view image classification, where the final prediction is assigned to the object after jointly analyzing all available visual observations.

Formally, each infrastructure object is represented as

$$O_i = \{I_1, I_2, \dots, I_j, \dots, I_k\}, k \geq 2, \quad (1)$$

where O_i denotes the i -th infrastructure object, I_j represents an individual image captured from a particular viewpoint, and k is the number of available views.

When available, the image set may contain photographs acquired both before and after the damaging event. In this case, the object representation can be expressed as

$$O_i = \{I_1^{before}, \dots, I_m^{before}, I_1^{after}, \dots, I_n^{after}\}, \quad (2)$$

where $m + n = k$. The "before" images provide reference information about the original appearance of the structure, while the "after" images contain evidence of possible damage. If reference images are unavailable, the assessment is performed solely using post-event observations.

Unlike conventional image classification datasets, where each image constitutes an independent sample, the proposed dataset organizes images into object-level groups. Consequently, every prediction corresponds to an entire infrastructure object rather than to a single photograph. This design substantially reduces ambiguity caused by occlusions, limited viewing angles, or locally invisible structural defects and provides a more reliable estimate of the actual damage level.

4.3. Data classification scheme

To ensure consistent annotation and evaluation, all infrastructure objects are assigned to one of five mutually exclusive damage classes. The proposed classification follows a progressive scale from intact structures to complete destruction and is based solely on visible structural characteristics observed across all available viewpoints. During annotation, the final class is determined according to the highest level of damage visible in any image belonging to the object.

- Class 0 – Undamaged. The building remains visually intact without observable structural defects. No visible cracks are present on the facade, the roof is undamaged, window glazing remains intact, and the overall geometry of the building is preserved.

- Class 1 – Slight Damage. Minor superficial damage is observed while the structural integrity remains unaffected. Typical indicators include broken window glass with intact frames, small facade defects such as chipped plaster or shrapnel marks, and limited roof covering damage without through-holes or structural deformation.
- Class 2 – Moderate Damage. The building exhibits localized structural damage but remains stable. Visible indicators include through-holes in exterior walls or the roof, complete loss of window or balcony units, and localized damage to structural elements. The primary load-bearing frame and the overall geometry of the building remain preserved, with no collapse of entire rooms or floors.
- Class 3 – Severe Damage. Significant structural failure is present. Observable characteristics include partial collapse of walls, exposure of interior rooms, collapse of floor slabs, pronounced deformation of the structural frame, and substantial loss of roof sections. Although parts of the building remain standing, its structural stability is critically compromised.
- Class 4 – Destroyed. The original building geometry is no longer recognizable. The structure has collapsed into rubble, leaving only fragments of lower walls, isolated structural remnants, or the foundation. The object is considered completely destroyed and no longer operational.

4.4. Multi-image prompting strategy

Unlike conventional deep learning classifiers that process individual images independently, Vision-Language Models enable simultaneous reasoning over multiple images within a shared conversational context. In the proposed methodology, all photographs corresponding to the same infrastructure object are submitted to the model in a single prompt. This allows the model to aggregate visual evidence from different viewpoints and produce a single object-level assessment. The prompt instructs the model to perform the role of an independent structural engineering expert and to evaluate only the visible evidence present in the provided photographs. The assessment explicitly excludes assumptions regarding hidden structural elements or the cause of the damage. Furthermore, the model is instructed to consider all available images jointly and assign one final damage class corresponding to the overall condition of the object.

The prompt used in this study is presented below.

“You act as an independent expert in structural inspection and post-disaster damage assessment of buildings and civil infrastructure. Your task is to perform an objective visual assessment of a single infrastructure object using the provided photographs acquired before and after the damaging event and determine its overall damage class. Each object may contain one or multiple photographs representing both the pre-event and post-event conditions. All photographs belonging to the same object must be analyzed jointly as a single source of information, and only one final damage class shall be assigned. Evaluate only the information directly visible in the images. Do not make assumptions about structural elements that are not visible, and do not infer the cause of the damage. Assess the building as an integrated structure. No individual visual feature should determine the damage class by itself. For example, broken windows or extensive facade damage alone do not indicate severe structural damage. The primary assessment criteria are the condition of the load-bearing structure, preservation of the building geometry, and overall structural integrity.

Classify the object according to the following scale.

Class 0 – Undamaged: The building is visually intact. No visible facade cracks, roof damage, or damage to windows or external structural elements. The original geometry is fully preserved.

Class 1 – Slight Damage: The structural geometry remains intact. Only window glass is broken while frames remain in place. Minor facade defects (e.g., shrapnel marks or localized plaster loss) and limited roof covering damage may be present. No through-holes or structural failures are observed.

Class 2 – Moderate Damage: Localized structural damage is visible. Through-holes in walls or the roof and complete loss of window or balcony units may occur. However, the primary structural frame, building corners, and overall geometry remain preserved. No collapse of floors or major structural sections is present.

Class 3 – Severe Damage: Partial structural collapse is observed. Exterior walls or floor slabs may have collapsed, interior rooms may be visible, significant roof sections may be missing, and the building may exhibit noticeable deformation or tilting. Nevertheless, a substantial portion of the structure remains standing.

Class 4 – Destroyed: The building has lost its structural form and consists primarily of rubble or collapsed debris. Only isolated foundation elements or remnants of lower walls remain.

If the observed evidence corresponds to two neighboring classes or is insufficient for an unambiguous decision, select the most probable class and reduce the confidence score accordingly.

Return the results using the following format:

1. Object identifier;
2. Visual analysis;
3. Key observations affecting the assessment;
4. Justification;
5. Final damage class (0–4);
6. Confidence score (0–100 %) with a brief explanation.”

The structured prompt standardizes the reasoning process of the multimodal model and ensures that every infrastructure object is evaluated according to identical criteria. By explicitly instructing the model to aggregate evidence from multiple viewpoints and focus on structural integrity rather than isolated visual defects, the proposed prompting strategy reduces ambiguity and improves the consistency of object-level damage classification. Furthermore, the structured textual output facilitates subsequent comparison with manually annotated ground-truth labels and enables direct computation of evaluation metrics.

5. Proposed method for damage assessment based on multi-view segmentation analysis and pixel metrics

For semantic segmentation of building damage from ground-level views, we use a deep convolutional neural network combining a symmetric U-Net decoder with a pre-trained ResNet-34 encoder. Skip connections directly transfer high-frequency spatial features – such as structural cracks and glass shards – from early encoder stages to the decoder, preventing spatial discretization losses during downsampling. The model is trained on a dedicated ground-level war damage dataset [13, 15] containing 290 high-resolution side-view images of affected buildings in Ukraine. Pixel-wise annotations cover six semantic classes: Other (class 0, 66.10 % of pixels), Building (class 1, 19.42 %), Roof (class 2), Damage (class 3), Broken Window (class 4, 2.66 %), and Damaged Roof (class 5, 1.92 %). To

preserve fine textures, images are segmented using a sliding window into 640×640 pixel patches with a stride of 160. Overlapping patch logits are averaged before applying a Softmax function to eliminate edge artifacts. A critical limitation of this dataset is that annotations are restricted strictly to "post-damage" images. Lacking "pre-damage" reference geometry, standard end-to-end models exhibit high classification bias and tend to underestimate actual structural degradation. The primary contribution of this research is a mathematical interpretation framework designed to resolve this bias. By applying optimized micro-thresholding and multi-view weighted score aggregation, the framework converts raw segmentation masks into robust, safety-constrained structural integrity classes. The six baseline classes are aggregated into three distinct Regions of interest (ROI) to extract physical area metrics:

- Building area (P_{building}): Total pixels representing active structures:

$$P_{\text{building}} = \sum \text{pixels} (\text{mask} \in [1,2,3,4,5]) \quad (3)$$

- Structural damage area ($P_{\text{structural_damage}}$): Combined pixels of damaged walls and roofs that threaten load-bearing capacity:

$$P_{\text{structural_damage}} = \sum \text{pixels} (\text{mask} \in [3,5]) \quad (4)$$

- Window damage area ($P_{\text{window_damage}}$): Pixels representing compromised glazing elements:

$$P_{\text{window_damage}} = \sum \text{pixels} (\text{mask} == 4) \quad (5)$$

These segmented pixel arrays form the input for the decision-making pipeline. The qualitative performance of the modified U-Net model in extracting the targeted structural and damage boundaries is shown in Figure 2.

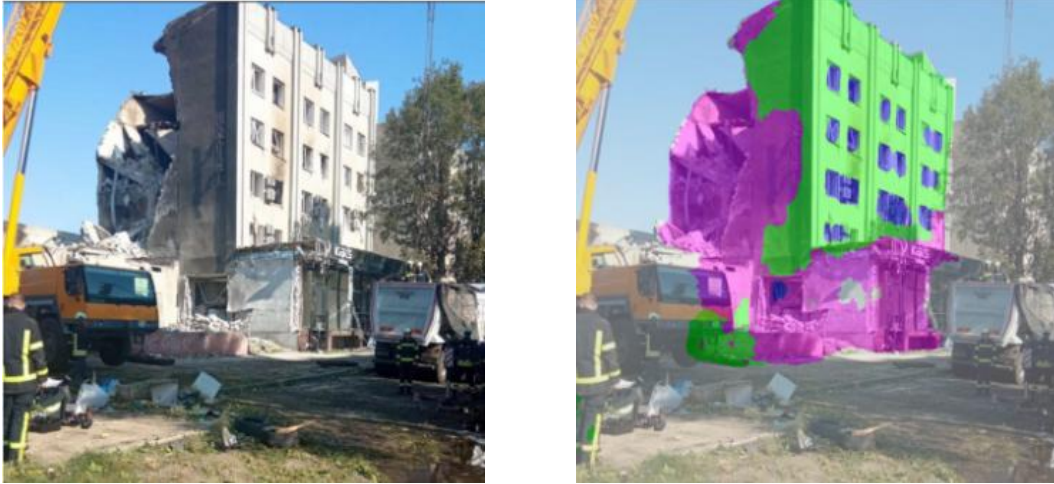


Figure 2: Visualization of the segmentation model: input facade image paired with overlaid multi-color ROI masks.

To eliminate scale variance caused by shooting distances, areas are converted into normalized ratios. The Structural damage ratio (R_{str}) and Glazing damage ratio (R_{w}) are computed relative to the total visible building silhouette:

$$R_{\text{str}} = \left(\frac{P_{\text{structural_damage}}}{P_{\text{building}}} \right) \times 100 \% \quad (6)$$

$$R_{\text{w}} = \left(\frac{P_{\text{window_damage}}}{P_{\text{building}}} \right) \times 100 \% \quad (7)$$

To counteract model bias from the post-damage training data constraint, we introduce a non-linear expert decision tree. The thresholds are calibrated to empirical engineering standards at tight structural intervals (6.0 % – 14.0 % – 35.0 % – 70.0 %):

$$\text{Class} = \begin{cases} 4, & \text{if } R_{\text{str}} > 70 \% \\ 3, & \text{if } 35 \% < R_{\text{str}} \leq 70 \% \\ 2, & \text{if } 14 \% < R_{\text{str}} \leq 35 \% \vee (R_{\text{str}} \leq 6 \% \wedge R_{\text{w}} > 20 \%) \\ 1, & \text{if } 6 \% < R_{\text{str}} \leq 14 \% \vee (R_{\text{str}} \leq 6 \% \wedge R_{\text{w}} \leq 20 \% \wedge (R_{\text{w}} > 6 \% \vee R_{\text{str}} > 2 \%)) \\ 0, & \text{else} \end{cases} \quad (8)$$

Ground-level evaluation using a single view is prone to occlusion errors (e.g., an intact front facade masking a collapsed rear wall). For multi-view sets ($2 \leq N \leq 8$), we implement a two-factor aggregation pipeline. For each individual viewpoint, a linear raw score is computed by scaling structural and glazing components with localized engineering weights:

$$\text{Score}^{(i)} = 1.0 \times R_{\text{str}}^{(i)} + 0.2 \times R_{\text{w}}^{(i)} \quad (9)$$

To balance localized, severe damage with the overall state of the entire asset, a global score (S_{object}) is formulated as a convex combination of maximum and mean perspective scores:

$$S_{\text{object}} = \alpha \times \max_{i=1 \dots N} (\text{Score}^{(i)}) + (1 - \alpha) \times \text{Mean}(\text{Score}^{(i)}) \quad (10)$$

where $\alpha = 0.35$ is the mathematically optimized weight.

The final structural class of the building ($\text{Class}_{\text{final}}$) is determined using calibrated decision boundaries ($T_1 = 6.0, T_2 = 17.0, T_3 = 25.0$):

$$\text{Class}_{\text{final}} = \begin{cases} 4, & \text{if } S_{\text{object}} > 70.0 \\ 3, & \text{if } T_3 < S_{\text{object}} \leq 70.0 \\ 2, & \text{if } T_2 < S_{\text{object}} \leq T_3 \\ 1, & \text{if } T_1 < S_{\text{object}} \leq T_2 \\ 0, & \text{if } S_{\text{object}} \leq T_1 \end{cases} \quad (11)$$

The parameters α and decision thresholds were determined through an empirical Grid Search optimization protocol over the defined parameter space (α in $[0.35, 0.50]$ with step 0.05; T_1, T_2, T_3 in $[1, 100]$ with steps 1.0 and 2.0). The objective function was set to maximize the global multi-class classification accuracy and F1-Macro metric across the dataset. This calibration procedure guarantees an optimal trade-off between prioritizing localized severe damage and suppressing isolated pixel-level segmentation noise. This dual-factor model enforces a safety constraint. If a severe failure is captured on even a single perspective, the maximum term through α pulls the global score into the corresponding hazard class. Concurrently, the mean term suppresses noise from isolated segmentation anomalies, ensuring reliability before sending data to GIS mapping and circular economy routing networks for construction waste.

6. Experiments & results

To evaluate the performance of the proposed interpretation and multi-view aggregation framework, we collected and annotated a dedicated regional test dataset. The dataset consists of damaged civil

infrastructure assets impacted by military strikes across multiple administrative regions in Ukraine. The test dataset was acquired from verified open channels (SESU official reports, regional news feeds, municipal administrations) across multiple war-affected regions of Ukraine.

For each physical building, we developed a structured database entry to capture both spatial attributes and visual pairs. The collected metadata schemas are simplified into three functional categories:

- **Spatial and classification metadata:** Includes a unique Object ID, geospatial parameters (latitude, longitude), administrative positioning (address, region code), functional category (object type).
- **Temporal visual perspective pairs (multi-view array):** Structured to pair historical pre-damage views (Before_Front, Before_Back, Before_Left, Before_Right) with post-damage perspectives (After_Front, After_Back, After_Left, After_Right, After_Iso). The isometric projection (After_Iso) is explicitly included to support downstream depth estimation of physical rubble piles.
- **Administrative and ground truth records:** Documents data status indicators, licensing terms, data collectors, and the verified expert-assigned target class (Damage).

Following this schema, we compiled a total of 108 unique testing assets. To ensure annotation integrity, each physical asset was independently evaluated by three domain experts (structural engineers and researchers) using the standardized 5-class engineering rubric (Classes 0–4). The target label distribution among these test samples is illustrated in Figure 3.

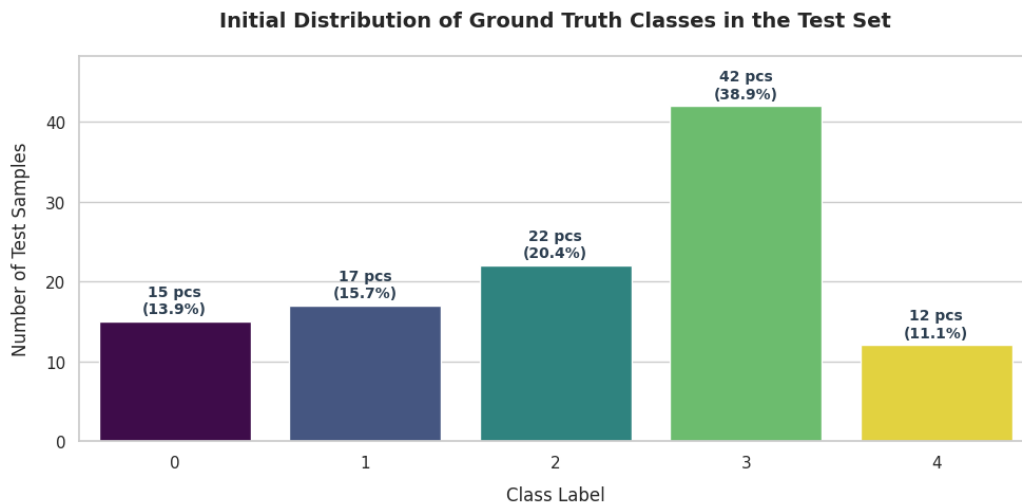


Figure 3: Initial distribution of ground truth classes in the test set.

The sample skew toward Class 3 (38.9%) and Class 2 (20.4%) reflects the physical reality of artillery and missile impacts in urban environments, where structural partial collapses are significantly more prevalent than either zero-damage scenarios (Class 0) or complete ground-level pulverization (Class 4). While this real-world distribution introduces a class imbalance bias, it serves as a representative stress-test for automated damage evaluation systems. To evaluate the effectiveness of the proposed multi-view damage assessment methodology, experiments were conducted using one conventional deep learning model and four state-of-the-art multimodal Large VLMs. The selected models represent different architectural paradigms, enabling a comparative analysis between a supervised computer vision approach and recent general-purpose multimodal AI systems. A brief overview of the evaluated models is presented in Table 1. To ensure independent reproducibility, all Vision-Language

Models (specifically Claude Sonnet, Gemini, Grok, and ChatGPT) were evaluated under standardized inference settings. Zero-shot prompting was employed, explicitly constraining the model output to return a single integer representing the structural damage class (0–4) based on predefined engineering guidelines provided in the prompt context. The performance of each evaluated model was assessed using six complementary metrics. Together, these metrics provide a comprehensive evaluation of both the overall classification performance and the types of prediction errors made by the models:

- Accuracy – measures the overall proportion of correctly classified infrastructure objects.

Table 1
Evaluated AI models used for infrastructure damage classification

Model	Category	Brief description
Unet + ResNet	Deep Learning	A specialized computer vision model featuring a symmetric encoder-decoder architecture
GPT-5.5 (High Intelligence)	Vision-Language Model	OpenAI multimodal large language model
Claude Sonnet 5 (Medium)	Vision-Language Model	Anthropic multimodal model
Gemini 3.1 Pro (High)	Vision-Language Model	Google's multimodal foundation model
Grok 4 (Fast)	Vision-Language Model	xAI multimodal model

- Weighted Precision – evaluates the reliability of predicted damage classes while accounting for class distribution.
- Weighted Recall – measures the model's ability to correctly identify objects belonging to each damage category.
- Macro F1-score – provides a balanced evaluation by treating all damage classes equally, making it particularly suitable for imbalanced datasets.
- Overestimation ratio – represents the proportion of predictions where the assigned damage class is higher than the ground-truth label.
- Underestimation ratio – represents the proportion of predictions where the assigned damage class is lower than the ground-truth label.

To perform a systematic quantitative evaluation, the prediction outputs generated across the test set were compiled into a centralized database. The example of models evaluation is available via link [16]. The comparative metrics were then calculated to assess classification accuracy, precision, recall, and balanced F1-macro values for each model. Additionally, we monitored engineering error trends, specifically tracking the rate of Overestimation (Hyperbolization) and Underestimation (Mitigation) to evaluate the safety and reliability of each approach. The comparative performance results of the evaluated models are summarized in Table 2.

Table 2
Summary model performance evaluation

Model	Accuracy	Precision (Weighted)	Recall (Weighted)	F1-Macro (Balanced)	Overesti- mation Ratio	Underesti- mation Ratio
Claude	0.639	0.674	0.639	0.639	0.176	0.185

Gemini	0.639	0.691	0.639	0.570	0.231	0.130
Grok	0.546	0.563	0.546	0.527	0.343	0.111
ChatGPT	0.537	0.550	0.537	0.515	0.315	0.148
Unet	0.407	0.424	0.407	0.348	0.398	0.194

The experimental results demonstrate a clear performance gap between high-level reasoning models and the raw baseline segmentation setup:

- **Top-tier performance:** Claude and Gemini achieved the highest overall accuracy of 0.639. Claude demonstrated the most balanced classification behavior with an F1-Macro of 0.639 and a low overestimation ratio of 0.176. Conversely, Gemini achieved the highest weighted precision of 0.691 while keeping critical underestimation (mitigation) mistakes down to a ratio of 0.130.
- **Mid-tier models:** Grok and ChatGPT showed moderate performance with accuracies of 0.546 and 0.537 respectively, and F1-Macro values slightly above 0.515. However, both models suffered from elevated overestimation ratios exceeding 0.310, leading to frequent false alarms.
- **Baseline limitations and mitigation:** The baseline Unet setup underperformed significantly, registering a low accuracy of 0.407 and a balanced F1-Macro of 0.348. This baseline was highly prone to overestimating damage levels with an overestimation ratio of 0.398.

Notably, if the undamaged Class 0 (for which the baseline U-Net was fundamentally untrained) is excluded from the evaluation, the segmentation model's classification accuracy rises significantly, trailing the top-tier models by a margin of only 10 % in the F1-macro metric. All full confusion matrices and classification reports are provided in the appendix [16].

7. Conclusions and future work

This study presented a comparative evaluation of high-level VLMs and a customized semantic segmentation baseline for ground-level infrastructure damage classification. The experimental findings validate that advanced reasoning models achieve strong predictive performance on complex structural parsing tasks.

Claude demonstrated the most robust and mathematically balanced classification behavior, securing the highest accuracy (0.639) and balanced F1-Macro (0.639). Crucially, Claude proved highly capable at delineating the heavily blurred boundary between Class 3 (Heavy Damage) and Class 4 (Destroyed), achieving a high precision of 0.740 and recall of 0.690 on the dominant Class 3 subset (Support = 42).

Gemini matched Claude's peak accuracy of 0.639 and exhibited a superior weighted precision of 0.691. Gemini proved to be the most conservative engineer-aligned model, recording the lowest underestimation ratio (0.130). This safety-first behavior is evidenced by its exceptionally high recall of 0.900 for Class 3, minimizing catastrophic structural false negatives, though it showed a tendency to collapse Class 4 instances into Class 3 predictions (recalling only 0.170 of Class 4).

Grok and ChatGPT occupied a moderate tier, with accuracies of 0.546 and 0.537, respectively. Both models suffered from elevated overestimation ratios exceeding 0.310, frequently misclassifying minor aesthetic surface anomalies as heavy structural collapses.

The Baseline U-Net (ResNet Backbone) underperformed globally, with an accuracy of 0.407 and an F1-macro of 0.348. This baseline registered a high overestimation ratio of 0.398. The confusion matrix reveals this error is fundamentally tied to an inability to identify undamaged structures (Class 0 precision of 0.600 and recall of 0.200), caused directly by the dataset limitation where training masks only contained "post-damage" imagery. However, the localized pixel segmenter's underestimation ratio (0.194) remained comparable to the top-performing VLMs. When Class 0 is removed from the array, the computational baseline trails the top models by a margin of only 0.100 in F1-macro, confirming its effectiveness as a high-fidelity local feature extractor. Beyond raw classification metrics, the comparative analysis reveals critical operational trade-offs between commercial cloud-based VLMs and localized computer vision baselines. Frontier models (Claude, Gemini) were deliberately evaluated to establish the theoretical upper-bound performance limit of multimodal reasoning for structural damage assessment. However, their real-world post-disaster deployment faces major challenges, including data privacy risks when processing sensitive national infrastructure, operational dependency on unstable cloud connectivity in active conflict zones, and high per-image API costs during large-scale surveys. Open-source VLMs were intentionally excluded from this initial benchmark to preserve a clear methodological boundary between high-level cloud reasoning and lightweight edge segmentation. Despite lower visual reasoning accuracy, the custom patch-based U-Net baseline offers indispensable advantages for practical field deployment within the GIS decision support engine:

1. **Offline Autonomy & Security:** It operates locally on edge hardware, ensuring full data privacy and independence from external API infrastructure;
2. **Deterministic Pixel Quantification:** Unlike VLMs that output qualitative text labels, U-Net generates explicit semantic pixel masks required for precise downstream construction and demolition waste (CDW) volume estimation;
3. **Economic Scalability:** It eliminates pay-per-token API costs, enabling cost-effective processing of tens of thousands of urban inspection images.

The developed multi-view classification framework serves as a core decision engine for the intelligent multifunctional GIS, aligned with the objectives of project. The localized structural class predictions and raw pixel distributions generated by this model are directly ingested by the system's downstream components:

- **Automated material estimation:** the system maps the assigned damage class (0–4) to localized building archetypes, converting pixel areas into concrete volume estimates for construction and demolition (C&D) waste fractions (e.g., crushed concrete, brick masonry dust, and structural glazing scrap).
- **Logistics optimization:** these physical waste volume estimates are cross-referenced with spatial coordinates (latitude, longitude) inside the geographic database. The intelligent multifunctional GIS decision support engine then dynamically routes demolition fleets, matching estimated waste stockpiles to the nearest processing and recycling hubs to lower regional transport costs.

Future work will focus on expanding the ground-truth database to include an equal balance of undamaged, pre-disaster building geometries. Re-training the patch-based U-Net with these reference samples will eliminate its current overestimation bias regarding Class 0 features. Given its small footprint and low latency, this localized segment represents a highly scalable edge deployment option for real-time field operations, bypassing the high API costs and network dependencies associated with large-scale proprietary VLMs.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.5 (High Intelligence), Claude Sonnet 5 (Medium), Gemini 3.1 Pro (High) and Grok 4 (Fast) in order to: implement method of a multi-view damage assessment that combines image analysis performed by Vision-Language Models in Sections 4, 6 of the article (data in [16]). Further, the author(s) used Gemini 3.5 Flash in order to: improve text translation, paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

Acknowledgments

This research was funded by the Ministry of Education and Science of Ukraine as part of the state budget scientific research project "Development of an intelligent multifunctional geographic information system for construction waste inventory in the reconstruction of war-damaged infrastructure" (State Registration No. 0126U001737), carried out at Odesa Polytechnic National University. The authors also express their gratitude to Anastasiia Baranova, Bohdan Ipatov and Dmytro Palas for their valuable assistance in collecting the war-damage image dataset and running the VLMs benchmarking experiments.

References

- [1] C. G. Samal, D. R. Biswal, G. Udgata, S. K. Pradhan, Estimation, Classification, and Prediction of Construction and Demolition Waste Using Machine Learning for Sustainable Waste Management: A Critical Review, *Construction Materials* 5 (2025). doi:10.3390/constrmater5010010.
- [2] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, M. Gaston, Xbd: A dataset for assessing building damage from satellite imagery, *arXiv* (2019). doi:10.48550/ARXIV.1911.09296.
- [3] Y. Klym, Assessment of Damage to Ukrainian Infrastructure Using Satellite Imagery, Bachelor's thesis, Ukrainian Catholic University, Lviv, Ukraine, 2025. URL: <https://hdl.handle.net/20.500.14570/5516>.
- [4] M. Molchanova, V. Didur, O. Mazurets, O. Sobko, O. Zakharkevich, Method for Construction and Demolition Waste Classification Using Two-Factor Neural Network Image Analysis, in: *Proceedings of the 2nd International Conference on Smart Automation & Robotics for Future Industry (SMARTINDUSTRY 2025)*, Vol. 3970, *CEUR Workshop Proceedings*, (CEUR-WS.org), Lviv, Ukraine, 2025, pp. 168–182. URL: <https://ceur-ws.org/Vol-3970/PAPER14.pdf>.
- [5] S. Guo, M. J. Cockayne, M. Ortolani, R. Al-Hamd, B. Al-Bander, Zero-Shot Crack Segmentation for Structural Health Monitoring Using Class Activation Map Guided Segment Anything Model, *SSRN* (2025). doi:10.2139/ssrn.5197086.
- [6] Z. Zheng, Y. Zhong, J. Wang, A. Ma, L. Zhang, Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters, *Remote Sensing of Environment* 265 (2021). doi: 10.1016/j.rse.2021.112636.
- [7] D. Duarte, F. Nex, N. Kerle, G. Vosselman, Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks, *Remote Sensing* 10 (2018). doi: 10.3390/rs10101636.

- [8] J. M. C. Estêvão, Effectiveness of Generative AI for Post-Earthquake Damage Assessment, *Buildings* 14 (2024). doi:10.3390/buildings14103255.
- [9] K. Alimi, R. Jin, B. N. Nguyen, Q. Nguyen, W. Chen, L. Hosking, Exploring artificial intelligence applications in construction and demolition waste management: a review of existing literature, *Journal of Science and Transport Technology* 5 (2025) 104–136. doi:10.58845/jstt.utt.2025.en.5.1.104-136.
- [10] L. Dadiverina, D. Laukhin, O. Hryhoriev, D. Khalymendyk, V. Bondarenko, Management of construction waste generated by military actions: analysis and solutions, *Collection of Research Papers of the National Mining University* 79 (2024) 163–174. doi: 10.33271/crpnmu/79.163.
- [11] R. S. Aouf, Dezeen, Circularity on the Edge AI finds reusable materials in rubble of Ukrainian buildings, 2025. URL: <https://www.dezeen.com/2025/07/07/circularity-on-the-edge-ai-ukraine/>.
- [12] Y. Aimaiti, C. Sanon, M. Koch, L. G. Baise, B. Moaveni, War Related Building Damage Assessment in Kyiv, Ukraine, Using Sentinel-1 Radar and Sentinel-2 Optical Images, *Remote Sensing* 14 (2022). doi: 10.3390/rs14246239.
- [13] A. Lysanets, O. Kosukha, T. Panchenko, Y. Tereshchenko, Embedding-Enhanced U-Net with a Patch-Based Approach and a Novel Dataset for Ground-Level Building Damage Segmentation, *Applied Artificial Intelligence* 40, (2026). doi: 10.1080/08839514.2026.2685357.
- [14] O. Arsirii, O. Ivanov, K. Bieliaiev, N. Cudecka-Puri`1na, Intelligent Information Technology for Inventory and Utilization of Construction and Demolition Waste From Damaged Infrastructure Facilities, in: *CEUR Workshop Proceedings, Vol-4048: Proceedings of the 13-th International Conference on Information Control Systems & Technologies (ICST 2025)*, Odesa, Ukraine: CEUR-WS.org, Sep. 2025, pp. 276–289. doi: 10.55056/ceur-ws.org/Vol-4048/paper22.pdf.
- [15] A. Lysanets, Ukrainian Building Damage Assessment System: Ground-level Ukrainian Building Damage Segmentation Dataset, 2025. URL: <https://www.kaggle.com/datasets/andrewlysanets/ukrainian-damaged-buildings/data>.
- [16] O. Ivanov, K. Bieliaiev, V. Oliinyk, O. Arsiriy. Supplementary Material for: Comparative Analysis of Vision-Language Models for Infrastructure Damage Classification: A Case Study of Ukraine, 2026. URL: https://drive.google.com/drive/folders/1r3_nUITY1vIx_461FKUoebM4VM5OgGmW?usp=sharing.