UDC 004.93
https://doi.org/10.20535/2786-8729.7.2025.341475

EVALUATION OF THE EFFECTIVENESS OF TWO APPROACHES TO BUILDING DAMAGE DETECTION WITH SATELLITE IMAGERY

Yurii Oliinyk
https://orcid.org/0000-0002-7408-4927

Oleksii Rumiantsev*
https://orcid.org/0009-0005-7223-3633

National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukrainee

*Corresponding author: rumiantsev.oleksii@gmail.com

This study addresses the approaches for satellite image analysis to assess infrastructure damage. The main aim is to conduct a comprehensive comparative analysis of the effectiveness of two key machine learning approaches: specialized semantic segmentation based on the `U-Net` architecture and generalized visual analysis using large vision-language models. The object of the research is the process of quantitatively benchmarking these two distinct approaches to determine their practical applicability for multi-class damage classification.

The research material is the publicly available `xView2` dataset. The methods involved two parallel experiments. For the semantic segmentation approach, a `U-Net` model with an `EfficientNet-B4` encoder was implemented and trained on 6-channel input data ("before" and "after" images) using a combined `Dice` and `Focal` loss function. For the vision-language models approach, the open-source `LLaVA-1.5-7B` model was evaluated in a zero-shot mode using advanced prompt engineering for an aggregative counting task. To enable a direct comparison, the standard *Jaccard index* was calculated based on the aggregated object counts for each damage class.

The results of the experiments revealed a significant performance disparity. The specialized `U-Net` model demonstrated high effectiveness, achieving an intersection over union score of 0.6141 on the test set. In contrast, the `LLaVA` model proved unsuitable for accurate quantitative analysis, yielding an extremely low *Jaccard index* of approximately 0.063, primarily due to its systemic failure to correctly identify and count objects (***Recall*** $\approx$ 0.07). The scientific novelty lies in being the first study to quantitatively document this order-of-magnitude capability gap, confirming that for tasks requiring high-precision mapping, specialized segmentation models remain the indispensable tool.

Keywords: satellite image analysis, damage detection, semantic segmentation, U-Net, large vision-language model.

1. Introduction

In an era increasingly defined by widespread natural disasters and geopolitical conflicts, the ability for rapid and accurate assessment of infrastructure damage is a critical challenge in the modern world [1]. The speed and objectivity of such assessments directly influence the effectiveness of humanitarian response, economic recovery planning, and the documentation of potential war crimes. However, traditional approaches that rely on manual satellite image interpretation by experts are too slow and resource-intensive for large-scale crisis situations, creating a dangerous bottleneck when lives are at stake.

These challenges have stimulated the development of automated, AI-based methods capable of processing large volumes of remote sensing data within hours instead of weeks. Automation not only accelerates the delivery of humanitarian aid but also provides an objective and quantitative foundation for damage assessment and accountability. Therefore, the creation of reliable automated damage assessment systems has become a key research direction at the intersection of computer vision and crisis management. This establishes the field as not merely an academic pursuit, but a vital component of modern emergency response and international law, underscoring the profound relevance of advancing this technology.

2. Literature review and problem statement

Semantic segmentation forms the foundation of most modern image analysis approaches, with the U-Net architecture emerging as the dominant framework. U-Net's encoder–decoder structure with skip connections effectively combines deep semantic features with high-precision spatial information [2], which is essential for accurate object delineation. Over time, the architecture has evolved through numerous extensions, such as U-Net++ [3], and adaptations incorporating stronger encoder backbones like ResNet and EfficientNet, which have become standard choices for remote sensing tasks.

For the specific task of building damage assessment, U-Net-based models are among the most widely used, particularly in studies utilizing the xView2 benchmark dataset [4]. This dataset, which provides paired images of disaster-affected areas before and after an event, has driven the development of change detection methods that represent both images as a multi-channel input. In addition to conventional CNN-based approaches, more advanced transformer-based architectures such as DAHiTrA [5] have recently been proposed, employing hierarchical attention mechanisms to model spatial dependencies and detect structural changes more effectively.

At the same time, there are studies that focus on alternative strategies, such as analyzing damage based on a single image after the event. Although this approach loses the context of comparison, it is more flexible because it does not require the availability of "before" archive images. In particular, in the work [6], the authors demonstrated that by transfer learning a model previously trained on xView2 on specific data of damage caused by combat operations, high Intersection over Union (IoU) can be achieved on the target dataset. This highlights the importance of adapting models to local conditions and the specifics of damage.

In parallel with the progress of specialized models, the emergence of large Vision-Language Models (VLMs) such as LLaVA [7] and proprietary systems like Gemini [8] has introduced new opportunities for image analysis. Trained on massive web-scale datasets, these models can perform zero-shot analysis through natural language queries. Although still an emerging field, the application of such models to remote sensing has gained significant research attention. Recent works, including GeoRSMLLM [9] and ChangeGPT [10], have demonstrated the potential of fine-tuned VLMs for tasks such as land cover classification and textual change description.

Despite the rapid advancement of large vision–language models and their impressive performance on general-purpose tasks, their effectiveness in highly specialized domains – such as satellite image-based damage assessment – remains largely unexplored. Most existing studies either qualitatively showcase the potential of VLMs or compare different models within this category, without conducting direct, quantitative comparisons against established semantic segmentation architectures.

Consequently, a notable gap exists in the current scientific literature: the absence of a systematic, quantitative, and qualitative comparison between specialized segmentation models (e.g., U-Net) and general-purpose VLMs (e.g., LLaVA, Gemini) in the context of complex multi-class damage assessment. The magnitude of the accuracy gap between these two approaches, the characteristic error types, and the potential trade-off between flexibility and precision remain unclear. This lack of benchmarking limits the practical adoption of VLMs in emergency response workflows and introduces uncertainty regarding the selection of the most appropriate analytical approach for specific operational needs.

3. The aim and objectives of the study

The aim of this work is to address the identified gap in the literature by conducting a comprehensive comparative analysis of the effectiveness of two fundamentally different approaches to satellite image analysis – specialized semantic segmentation and generalized visual analysis – for the task of multi-class building damage assessment.

1. To develop and train an effective semantic segmentation model based on the U-Net architecture, establishing a quantitative performance benchmark for a specialized, fully-supervised approach.

2. To establish an evaluation framework for a generalized, zero-shot approach using the LLaVA Vision-Language Model, including the development of adapted metrics for its aggregative counting task.

3. To conduct a direct comparative analysis of the results to quantitatively measure the performance gap, identify the characteristic error types for each approach, and provide clear conclusions on their respective practical applications.

4. The study materials and methods for comparing damage detection approaches

4.1. The object, subject, and hypothesis of the study

The object of this study is the process of automated, multi-class building damage assessment from satellite imagery. The subject of the study is the evaluation of the effectiveness of two distinct computational approaches: a specialized, fully-supervised semantic segmentation approach and a generalized, zero-shot VLM analysis approach. The central hypothesis is that the specialized U-Net model, trained on domain-specific data, will demonstrate significantly higher quantitative accuracy and reliability compared to the general-purpose LLaVA model, revealing a fundamental performance gap between the two approaches for this specialized task.

4.2. Dataset and preprocessing

All experiments in this study were conducted using the publicly available xView2 benchmark dataset. This dataset was specifically developed for building damage assessment following natural disasters and serves as a widely recognized standard for model comparison in this domain.

The xView2 dataset comprises thousands of high-resolution satellite images (0.3–0.5 m/pixel) collected by WorldView satellites. Its key characteristic lies in its structure: for each geographic location, two images are provided – one captured before the disaster (pre-disaster) and one after (post-disaster). This dual-image design enables both single-image state analysis and change detection between paired images.

Each post-disaster image is accompanied by detailed vector annotations in JSON format. For every building instance, both its polygonal footprint and corresponding damage class are provided. Within the xView2 competition framework, four primary damage levels were defined: background, no damage, minor damage, major damage, and destroyed.

In addition to these categories, the dataset includes an “unclassified” label for buildings whose condition cannot be reliably determined. The dataset’s large scale and diversity – encompassing 19 disaster types ranging from hurricanes and earthquakes to fires and volcanic eruptions – make xView2 an ideal benchmark for training and objectively evaluating model generalization performance.

To ensure the reproducibility and objectivity of all experiments, a unified preprocessing pipeline was developed and consistently applied across all models.

All experiments in this study were conducted using a custom subset of the publicly available xView2 benchmark dataset, totaling 8,213 unique image pairs. The dataset was divided into three independent subsets:

- training set: 4,853 pairs, used to train the model weights;
- validation set: 1,680 pairs, used to tune hyperparameters and monitor the training process (e.g., to select the best epoch);
- test set: 1,680 pairs, used only once for the final, objective evaluation of the trained models’ performance.

The division was performed at the level of unique images to avoid data leakage when parts of the same image could end up in different sets.

Since semantic segmentation models require a large amount of video memory, full-size images (1024x1024 pixels) were cut into smaller fragments (tiles or patches). This process, known as patching, was implemented "on the fly" using a custom DataLoader. The following parameters were used for the training and test sets:

– tile size: 512x512 pixels;

– overlap: 25%, which allowed the model to see objects at the tile boundaries in different contexts.

For each tile, a corresponding pixel mask was generated, where each pixel was labeled with an integer from 0 to 4, corresponding to five classes: background, no-damage, minor-damage, major-damage, and destroyed.

To improve the generalization ability of the U-Net model and prevent overfitting, real-time geometric augmentations were applied to the training tile set. The following set of transformations was used:

– random horizontal and vertical reflections;

– random 90-degree rotations.

These transformations allowed us to artificially increase the diversity of the training data without changing its semantic content.

4.3. Methodology for the semantic segmentation approach

For the semantic segmentation task, the U-Net architecture from the segmentation-models-pytorch library was used, which is a high-level wrapper over PyTorch. Three different training strategies were explored, the results of which are compared in this paper.

U-Net with a powerful and efficient EfficientNet-B4 encoder [11], pre-trained on the ImageNet dataset, was chosen as the main architecture for the experiments. This allowed us to take advantage of transfer learning for faster convergence and better learning of visual features. For comparative analysis, the results obtained with the ResNet34 encoder in previous studies are also presented. The number of output classes of the model was set to 5 (four damage classes and one background class).

A combined loss function consisting of Dice Loss and Focal Loss [12] was chosen for training the model. This combination has proven itself well for segmentation tasks with pronounced class imbalance, since Dice Loss optimizes the IoU metric at the object shape level, and Focal Loss focuses on pixels that are difficult to classify.

AdamW [13] was used as the optimizer with a learning rate set to $1e-4$. For dynamic adjustment of the learning rate during training, the CosineAnnealingLR scheduler was used, which smoothly decreased it over epochs, contributing to more stable model convergence.

Training was conducted using mixed precision computing (torch.cuda.amp), which significantly accelerated the process and reduced video memory usage. The batch size was set to 64. The models were trained for 50 epochs on the training set, and the best model for the final evaluation was selected based on the highest IoU metric on the validation set.

4.4. Methodology for the visual-language models approach

To evaluate the effectiveness of the visual questioning approach, one of the leading open-source VLM, LLaVA-1.5 (7B), was selected. The model was tested in zero-shot mode, i.e., without any additional training on the xView2 dataset, to evaluate its generalizability "out of the box."

After a series of preliminary experiments with different types of queries, a final, highly detailed prompt was developed and used for both models. This prompt is based on the techniques of "few-shot learning" [14] and "chain-of-thought" [15] and contains the following key elements:

– role: the model was given the role of "satellite image analysis expert";

– context: it was clearly stated that the model was given two images – pre-disaster and post-disaster – and its task was to compare them;

– class definitions: a detailed visual description was provided for each of the four damage classes to minimize ambiguity;

– clear task: the model was tasked with counting the total number of buildings in the "after" image and providing a list of their classifications;

– format example: an ideal example of a response in JSON format was provided to demonstrate the expected output structure.

The `LLaVA` model was evaluated on a random sample from the test dataset consisting of 200 examples. The transformers library with 4-bit quantization was used to run the model in an environment with limited computing resources. The generation parameters were set to deterministic (`do_sample=False`) to ensure reproducibility of the results.

The text response obtained from the model was processed by a special parser, which extracted a vector of quantities for each damage class from it. For reliability and to prevent data loss during long-term evaluation, the results for each example were stored incrementally.

4.5. Framework for comparative analysis

For an objective comparison of the performance of two different approaches (semantic segmentation and aggregative visual questioning), two sets of metrics were used, adapted to the specifics of the input data for each approach.

The quality of pixel masks generated by the `U-Net` model was evaluated using standard metrics for segmentation tasks. Calculations were performed for all damage classes, ignoring the background class to obtain a more relevant assessment.

IoU: the main metric that measures the degree of overlap between the predicted and reference masks for each class.

F1-score (dice coefficient): the harmonic mean between *Precision* and *Recall* at the pixel level.

Since the VLM model generates aggregated data in the form of a vector of quantities ($[number\ of\ no-damage, \ldots]$) rather than pixel masks, standard segmentation metrics cannot be applied directly. To evaluate its performance on this aggregative task, standard classification metrics (*Jaccard Index*, *F1-score*, *Precision*, and *Recall*) were calculated based on object counts accumulated across the entire test set.

First, the aggregated indicators *True Positives* (TP), *False Positives* (FP), and *False Negatives* (FN) were calculated for the entire dataset:

– TP was defined as the sum of element-wise minimums between the true and predicted quantity vectors;

– FN was calculated as the difference between the total number of real objects and TP;

– FP was calculated as the difference between the total number of predicted objects and TP.

Based on these aggregated metrics, the following metrics were calculated:

– *Jaccard Index*: the main metric calculated as $TP/(TP+FP+FN)$. It evaluates how well the distribution of predicted classes corresponds to the true one;

– *F1-score*, *Precision*, and *Recall*: standard metrics, calculated from the aggregated TP, FP, FN, to evaluate the accuracy and completeness of predictions at the distribution level;

– accuracy: shows the model's ability to count, regardless of classes. It is calculated as the ratio of the total sum of all predicted objects to the total sum of all actual objects (or vice versa, so that the value does not exceed 1.0). This metric shows how prone the model is to "invent" objects or, conversely, "miss" them.

5. Results of comparing damage detection approaches

5.1. Performance of the semantic segmentation approach

As part of this study, a semantic segmentation model based on the `U-Net` architecture was developed and trained for multi-class damage assessment. The model was trained on the `xView2` training dataset using a change detection approach, where 6-channel images (before and after the event) were fed into the input.

The training process lasted 45 epochs. The dynamics of the loss function and IoU metrics on the training and validation datasets during training are shown in Fig. 1 and Fig. 2. The graphs demonstrate the stable convergence of the model: the loss function decreased monotonically, while the IoU metric increased, reaching a plateau at the end of training. This indicates the absence of significant overfitting and the success of the chosen optimization strategy.

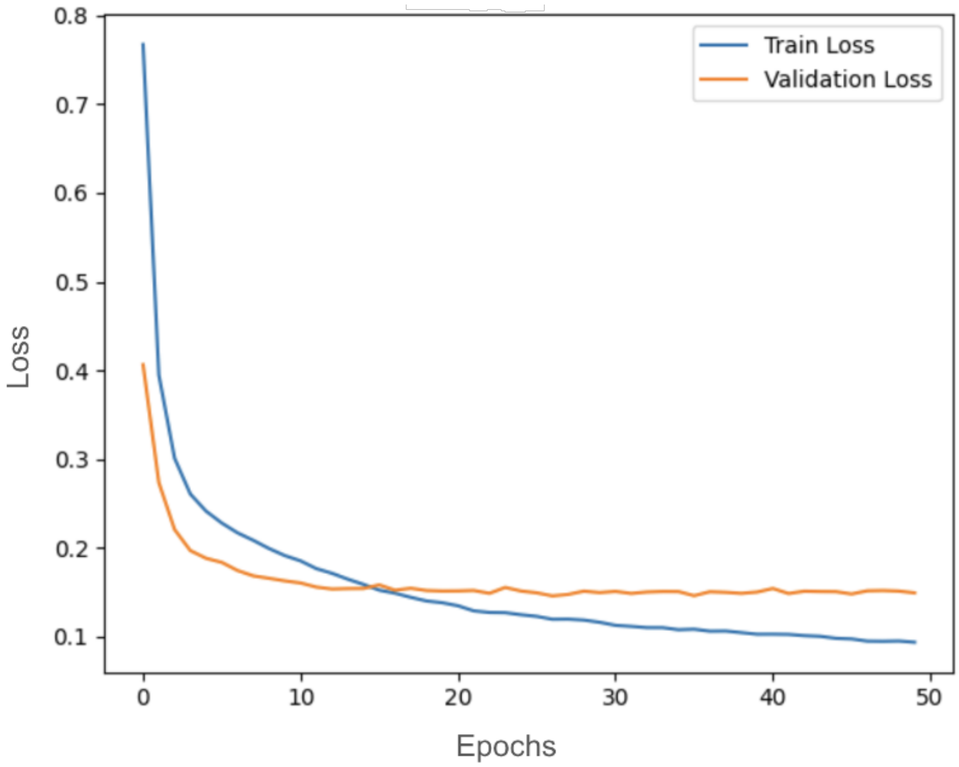


Fig. 1. Loss function graphs on training and validation datasets for the `U-Net` model

For a final, objective performance evaluation, the best version of the model (saved with the highest IoU score on the validation set) was tested on a deferred test dataset. The quantitative evaluation results are presented in Table 1. The presented metrics are macro-averaged across all four damage classes (`no damage`, `minor damage`, `major damage`, `destroyed`), providing a comprehensive assessment of the model's overall performance.

Table 1. Final performance metrics of the `U-Net` model on the test set

Metric	**Value**
mIoU (Jaccard Index)	0.6141
F1-score	0.7610
Accuracy (pixel)	0.9935

For a qualitative analysis, Fig. 3 shows an example of model visualization on an image from the test set, where the mask generated by the model is compared with the corresponding reference mask.

The visual comparison confirms the model's high level of accuracy. As shown, the predicted mask (c) closely matches the ground truth mask (b), correctly identifying both the location and the damage class of the building.

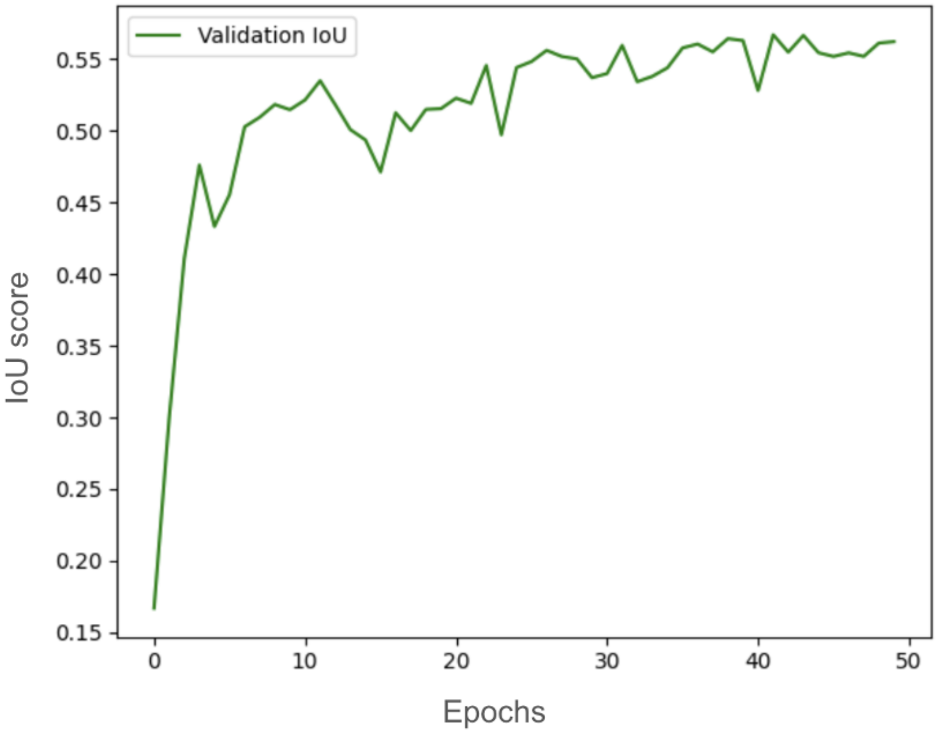


Fig. 2. *IoU* score on validation dataset for the `U-Net` model

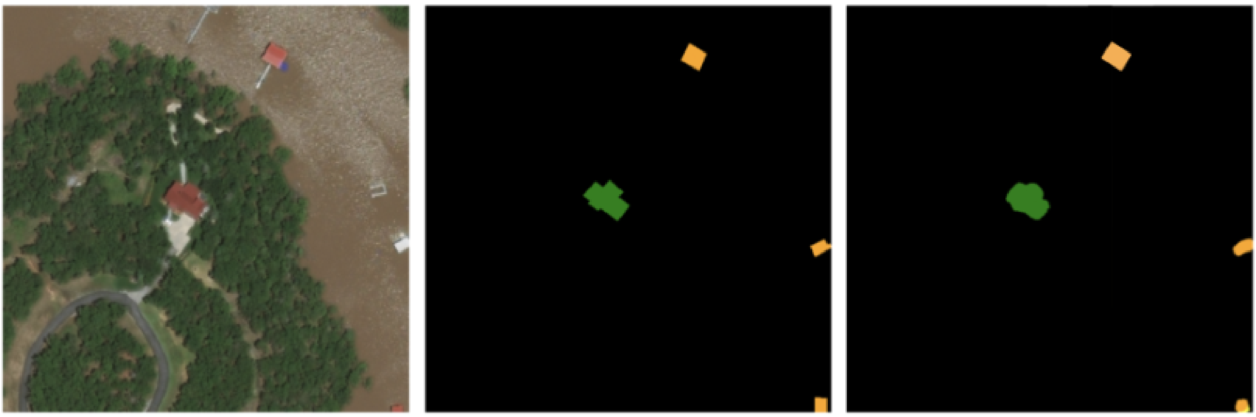


Fig. 3. Example of visualization of the `U-Net` model: *a* – input image (“after”); *b* – ground truth mask; *c* – predicted mask

5.2. Performance of the visual-language models approach

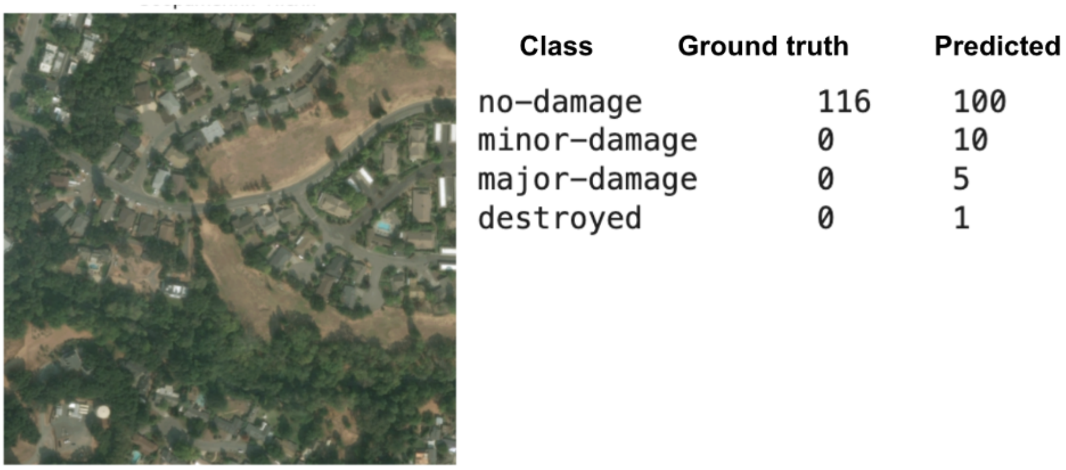
The quantitative evaluation of the `LLaVA-1.5-7B` model in zero-shot mode was performed on a random sample from the test set. The model’s performance in performing the aggregation task (counting the number of buildings for each damage class) was evaluated using a set of distribution metrics. Table 2 presents a detailed report on the model evaluation.

For a qualitative analysis of the model’s behavior, Fig. 4 and Fig. 5 show two representative examples from the test sample. For each example, the input image “after,” the “raw” text response generated by the model, and a comparison table between the reference and predicted numbers of objects are shown.

Visual analysis of examples confirms quantitative metrics, demonstrating the model’s tendency to significantly underestimate the number of objects and its difficulty in correctly classifying damage

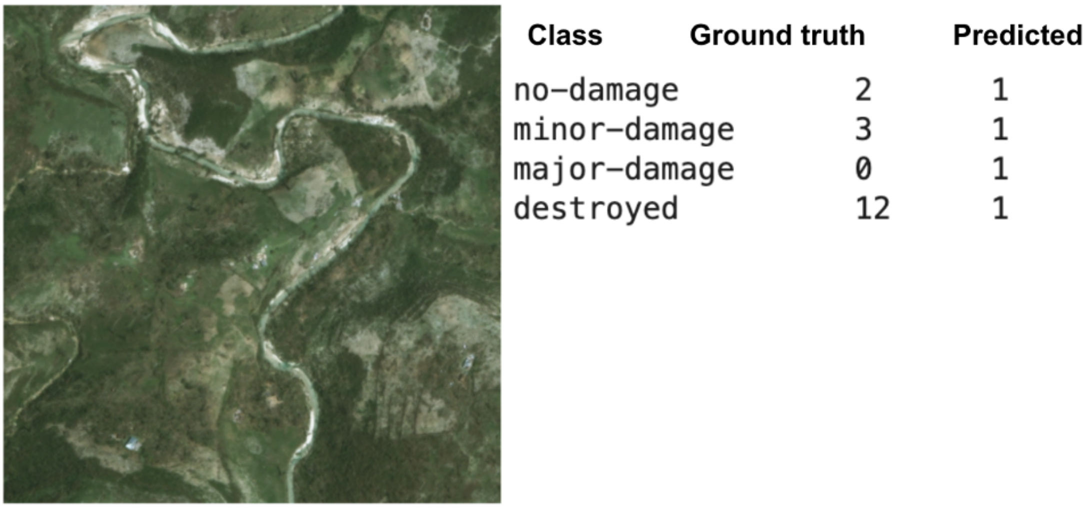
Table 2. Detailed report on the `LLaVA` model on the test subsets

Metric	Value
Jaccard Index	0.0627
F1-score	0.1181
Precision	0.4060
Recall	0.0691
Accuracy	0.1702



Class	Ground truth	Predicted
no-damage	116	100
minor-damage	0	10
major-damage	0	5
destroyed	0	1

Fig. 4. Example #1 of the `LLaVA` model in action. In an image containing 116 buildings, the model generated a response from which 116 objects were parsed, but with partially incorrect class distribution



Class	Ground truth	Predicted
no-damage	2	1
minor-damage	3	1
major-damage	0	1
destroyed	12	1

Fig. 5. Example #2 of the `LLaVA` model in action. In an image containing 17 buildings, the model generated a response from which 4 objects were parsed

levels.

5.3. Summary of comparative results

A direct comparison of the primary performance metrics reveals a significant disparity between the two approaches. The specialized `U-Net` model achieved a pixel-level *IoU* of 0.6141. In contrast, the generalized LLaVA model achieved a *Jaccard index* of 0.0627 on the analogous aggregative task. This represents a nearly tenfold difference in quantitative accuracy between the two approaches.

6. Discussion of comparative effectiveness

This study enabled a direct comparison between two fundamentally different approaches for damage assessment: specialized semantic segmentation based on the U-Net architecture and generalized visual reasoning using VLMs.

The key finding of this work is the substantial gap in quantitative accuracy between the two approaches. The specialized U-Net model, trained on paired “before” and “after” images, demonstrated strong performance, achieving an *IoU* score of 0.6141 on the test set. This result confirms the model’s capability to reliably localize and classify all four damage levels at the pixel level.

In contrast, the LLaVA model, evaluated using the *Jaccard index* metric, achieved only 0.0627. This disparity is considered fundamental rather than incremental for several reasons. First, the difference in performance spans an order of magnitude, reflecting a fundamentally different level of capability. Second, the error analysis reveals that while U-Net produces local segmentation inaccuracies, LLaVA exhibits a conceptual inability to execute core tasks such as object identification and counting (***Recall*** $\approx$ 0.07). These findings indicate that, at the current stage of technological development, specialized architectures such as U-Net remain the only reliable option for producing detailed and trustworthy damage maps.

Further analysis of LLaVA’s distribution-level metrics (*Precision* and *Recall*) provides deeper insight into the causes of its low performance. The extremely low *Recall* value (0.0691) represents the primary limitation: it suggests that the model identifies only about 7% of all actual buildings present in the imagery. This finding highlights a systemic inability of the model to reliably detect and enumerate objects.

At the same time, the relatively higher *Precision* value (0.4060) reflects an interesting behavior: among the limited number of buildings that the model manages to classify, a non-zero proportion of predictions are correct. However, this marginal accuracy is entirely offset by the severe incompleteness of the analysis. In contrast, U-Net demonstrates well-balanced performance, as reflected by its high *F1-score* (0.7610).

The results clearly delineate practical niches for both approaches.

U-Net remains indispensable for applications requiring high quantitative accuracy and reliability. Its capacity to produce detailed, validated pixel-level maps establishes it as the de facto “gold standard” for precise damage assessment, mapping for rescue operations, and other high-stakes analytical tasks.

LLaVA (and similar open-source VLMs) operating in zero-shot mode are unsuitable for quantitative analysis. Nevertheless, their unique advantage – the ability to function without additional training or labeled data – gives them potential value in tasks requiring rapid, preliminary qualitative assessment. For example, they may be useful for coarse filtering or triaging large image collections to detect possible damage indicators, where accuracy is not the primary requirement.

It should be noted that the LLaVA evaluation was performed in zero-shot mode, and its performance could likely be improved through fine-tuning on a specialized dataset. Furthermore, this study examined only one open-source VLM, whereas more advanced proprietary models may yield different outcomes. Finally, both models were evaluated using data exclusively from the xView2 dataset, which limits the generalizability of the findings to other damage domains, particularly those related to conflict-induced destruction.

The findings confirm that U-Net and VLMs are complementary tools suited to different stages of crisis response. Specialized models such as U-Net are indispensable for detailed quantitative analysis, whereas VLMs retain potential for rapid qualitative reconnaissance and preliminary triage, albeit requiring substantial improvement to reach operational reliability.

Key areas for future research include fine-tuning open VLM on specialized datasets to test whether this approach can bridge the existing performance gap, as well as developing hybrid architectures that combine the strengths of both approaches.

Conclusions

This study investigated and compared two distinct computational approaches for building damage assessment. Based on the experimental results, the following conclusions were reached:

1. The specialized approach using a `U-Net` based semantic segmentation model proved to be highly effective for multi-class damage assessment. The trained model achieved a high quantitative accuracy, reaching an IoU of 0.6141 on the `xView2` test set.

2. The generalized approach using the `LLaVA` VLM in a zero-shot setting was found to be unsuitable for reliable quantitative analysis. The established evaluation methodology revealed a critically low performance, with a *Jaccard index* of 0.0627, primarily caused by a systemic failure in object identification (*Recall* of 0.0691).

3. The comparative analysis confirmed a fundamental, order-of-magnitude performance gap between the two approaches. The results quantitatively demonstrate that for tasks requiring high-precision damage mapping, the specialized, supervised `U-Net` approach is a reliable and effective tool, whereas the generalized, zero-shot VLM approach, in its current state, is not a viable alternative.

References

[1] S. Voigt et al., “Global trends in satellite-based emergency mapping,” *Science*, vol. 353, no. 6296, pp. 247–252, Jul. 2016. https://doi.org/10.1126/science.aad8728

[2] O. Ronneberger, P. Fischer, and T. Brox, “U-NET: Convolutional Networks for Biomedical Image Segmentation,” *arXiv.org*, May 18, 2015. https://doi.org/10.48550/arXiv.1505.04597

[3] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNET++: a nested U-Net architecture for medical image segmentation,” *Lecture Notes in Computer Science*, vol. 11045, pp. 3–11, Jan. 2018. https://doi.org/10.1007/978-3-030-00889-5_1

[4] R. Gupta et al., “xBD: A Dataset for Assessing Building Damage from Satellite Imagery,” *arXiv (Cornell University)*, Feb. 2022. https://doi.org/10.48550/arxiv.1911.09296

[5] N. Kaur, C. Lee, A. Mostafavi, and A. Mahdavi-Amiri, “Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 38, no. 15, pp. 2072–2091, Feb. 2023. https://doi.org/10.1111/mice.12981

[6] O. Rumiantsev, Y. Oliinyk, and L. Oliinyk, “Damage detection based on satellite image analysis,” *in Lecture notes on data engineering and communications technologies*, 2025, pp. 177–189. https://doi.org/10.1007/978-3-031-88483-2_9

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *arXiv (Cornell University)*, Apr. 2023. https://doi.org/10.48550/arxiv.2304.08485

[8] G. Team et al., “Gemini: a family of highly capable multimodal models,” *arXiv (Cornell University)*, Dec. 2023. https://doi.org/10.48550/arxiv.2312.11805

[9] Z. Zhang et al., “GeoRSMLLM: a multimodal large language model for Vision-Language tasks in geoscience and remote sensing,” *arXiv.org*, Mar. 16, 2025. https://arxiv.org/abs/2503.12490v1

[10] Z. Xiao and J. Ma, “LLM agent framework for intelligent change analysis in urban environment using remote sensing imagery,” *Automation in Construction*, vol. 177, p. 106341, Jun. 2025.https://doi.org/10.1016/j.autcon.2025.106341

[11] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” *arXiv (Cornell University)*, Feb. 2022.https://doi.org/10.48550/arxiv.1905.11946

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” *IEEE Xplore*, pp. 2999–3007, Oct. 2017. https://doi.org/10.1109/iccv.2017.324

[13] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv (Cornell University)*, Mar. 2022. https://doi.org/10.48550/arxiv.1711.05101

[14] T. B. Brown et al., “Language Models are Few-Shot Learners,” *arXiv (Cornell University)*, vol. 33, pp. 1877–1901, Feb. 2022. https://doi.org/10.48550/arxiv.2005.14165

[15] J. Wei et al., “Chain-of-Thought prompting elicits reasoning in large language models,” *arXiv.org*, Jan. 28, 2022. https://doi.org/10.48550/arXiv.2201.11903

УДК 004.93

ОЦІНКА ЕФЕКТИВНОСТІ ДВОХ ПІДХОДІВ ДО ВИЯВЛЕННЯ РУЙНУВАНЬ БУДІВЕЛЬ ЗА ДОПОМОГОЮ СУПУТНИКОВИХ ЗНІМКІВ

Юрій Олійник
https://orcid.org/0000-0002-7408-4927

Олексій Румянцев
https://orcid.org/0009-0005-7223-3633

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна

У цьому дослідженні розглядаються підходи до аналізу супутникових знімків для оцінки пошкоджень інфраструктури. Основна мета — провести комплексний порівняльний аналіз ефективності двох ключових підходів машинного навчання: спеціалізованої семантичної сегментації на основі архітектури `U-Net` та узагальненого візуального аналізу з використанням великих зорово-мовних моделей. Об'єктом дослідження є процес кількісного порівняння цих двох різних підходів для визначення їхньої практичної придатності для багатокласової класифікації пошкоджень.

Матеріалом для дослідження слугував загальнодоступний набір даних `xView2`. Методи включали два паралельні експерименти. Для підходу семантичної сегментації було реалізовано та навчено модель `U-Net` з енкодером `EfficientNet-B4` на 6-канальних вхідних даних (зображення "до"та "після") з використанням комбінованої функції втрат `Dice` та `Focal`. Для підходу із зорово-мовними моделями, модель з відкритим кодом `LLaVA-1.5-7B` оцінювалася в режимі "zero-shot"з використанням передової інженерії запитів для задачі агрегованого підрахунку. Для прямого порівняння був розрахований стандартний *індекс Жаккара* на основі агрегованого підрахунку об'єктів для кожного класу пошкоджень.

Результати експериментів виявили значну розбіжність у продуктивності. Спеціалізована модель `U-Net` продемонструвала високу ефективність, досягнувши показника IoU 0.6141 на тестовому наборі. На противагу цьому, модель `LLaVA` виявилася непридатною для точного кількісного аналізу, показавши надзвичайно низьке значення *індексу Жаккара* близько 0.063, переважно через її системну неспроможність коректно ідентифікувати та підраховувати об'єкти (повнота для розподілів ≈ 0.07). Наукова новизна полягає в тому, що це перше дослідження, яке кількісно задокументувало цей розрив у можливостях на порядок, підтверджуючи, що для завдань, які вимагають високоточного картографування, спеціалізовані моделі сегментації залишаються незамінним інструментом.

Ключові слова: : аналіз супутникових знімків, оцінка руйнувань, семантична сегментація, U-Net, великі зорово-мовні моделі.