Journal of Applied Automation Technologies

Research Paper

Improved Post-Disaster Building Assessment from UAV Inspection Images by Integrating Adaptive Patch Vision Transformer

Hsiao-ting Yeh[1, *], Yi-wen Tseng[1] and Shu-chun Kuo[2]

[1] Department of Computer Science and Information Engineering, National Central University, Taoyuan, 32031, China
[2] College of Information Technology, Chien Hsin University of Science and Technology, Taoyuan, 32097, China
*Corresponding author: hsiaot.ye@ncu.edu.tw

Abstract. Unmanned aerial vehicle (UAV) inspection photos used for rapid post-disaster building evaluation must properly identify structural damage under various imaging settings. The long-distance spatial linkages between roofs, walls, trash, shadows, and nearby facilities are typically overlooked by the earlier convolutional techniques, despite their effectiveness in capturing local texture. While standard Vision Transformer models enhance global reasoning, they employ fixed patch partitioning, which might lead to the fragmentation of tiny damaged regions or the introduction of duplicate tokens from the background. creation of a damage-adaptive vision transformer model for building assessment using UAV images. The model creates adaptive patches with various spatial resolutions after first predicting the local visual complexity and damage likelihood. Roof-level, building-level, and block-level context will be combined via a three-tier Token aggregation module. The distinction between undamaged, moderately damaged, severely damaged, and collapsed structures is further improved by a disaster-aware classification head. Accuracy, macro-F1, recall of high-risk damage, computational cost, and cross-disaster resilience are the foundations of quantitative appraisal. The new framework is still appropriate for emergency UAV inspection scenarios and can accurately detect damage.

Keywords: *Adaptive Patch Vision Transformer, UAV Remote Sensing, Building Damage Assessment, Disaster Response, Hierarchical Token Aggregation*

Received on 29 November 2023, Accepted on 02 September 2024, Published on 13 September 2024

Introduction

Unmanned aerial vehicle (UAV)-based inspection may quickly provide high-resolution photos over hazardous or inaccessible places, making it a viable sensing method for post-disaster building evaluation. Emergency crews must quickly locate collapsed structures, partially damaged buildings, blocked highways, and high-risk rescue regions following earthquakes, floods, storms, landslides, and explosions. In order to differentiate between intact structures and those with slight or severe damage, UAV photos are required since they have a greater resolution than satellite images and may display the texture of roofs, exposed walls, debris borders, local shadow patterns, etc. [1]. As a result, deep learning has been extensively used in disaster response processes; one such example is the computerized evaluation of building damage using aerial photos taken after the occurrence [2]. Additionally, studies on post-disaster damage mapping have shown that the timeliness of emergency decision support has increased and that the amount of manual screening may be greatly reduced by using high-resolution visual evidence and automated picture interpretation [3]. Unmanned aerial vehicle (UAV) platforms are growing in number, and as a result, UAV-based building evaluation has progressively made its way into the realm of intelligent disaster management systems thanks to real-time communication linkages and cloud-edge computing resources [4].

In post-disaster UAV image analysis, the benefits of these are not completely realized. Flight altitude, viewing angle, light and shadow, roof material, surrounding vegetation, and kind of catastrophe may all have a big impact

on how a building looks. A collapsed corner, a fractured roof, an exposed interior, or a pile of rubble may only cover a tiny portion of the image, while the majority of the pixels belong to intact backdrop or visually comparable buildings [5]. Damage signals are frequently uneven and sparse. Although their fixed receptive fields may not be able to grasp long-range structural links among roof edges, wall pieces, and surrounding debris, traditional convolutional neural networks are capable of learning strong local texture patterns [6]. Vision Transformer variations have demonstrated high performance in remote sensing scene interpretation, and transformer-based visual models provide a more robust method for global dependency modeling [7]. Because consistently tiny patches add considerable token redundancy and processing complexity, and big patches may dilute minor damaged components, a normal fixed-patch transformer is not naturally suited for UAV catastrophe scenarios [8]. Adaptive visual modeling may be necessary for high-resolution remote sensing applications, according to recent research on Transformer backbones, hierarchical attention, and effective token representation [9].

By including an adaptive patch Vision Transformer into the processing of UAV inspection photos, this article presents an enhanced approach for post-disaster building evaluation. To put it briefly, locations with high structural complexity or damage saliency are assigned more granular visual tokens, whereas homogenous backdrop or low-risk building sections are assigned coarser tokens. In order to increase the recognition accuracy of intact, moderately damaged, severely damaged, and collapsed buildings without significantly increasing the computational load on the emergency deployment platform, adaptive patch generation, hierarchical spatial context modeling, and disaster-aware classification are combined. Preserving local damage cues under different imaging circumstances, modeling the spatial connections among damaged building components, and accomplishing effective inference in the presence of numerous redundant background areas in high-resolution disaster photos are the issues that need to be resolved. The issue stems from the inability of manual interpretation to meet the pace demands of disaster relief activities. Thousands of overlapping photos with several buildings at varying visibility levels can be produced by a single UAV flight, and human annotation is challenging since it requires professional judgment, fatigue risk, inconsistent standards, and poor communication circumstances. Because it requires simultaneous accuracy, speed, robustness, and explainability, an automated model that provides the first damage ranking and interpretable evidence aids in the more rational allocation of inspection resources for emergency teams. This model differs from traditional remote sensing classification. Additionally, UAV-based post-disaster assessment should be cautious to use both local and global visual evidence: subtle indicators, like small cracks or missing roof tiles, may indicate moderate damage, but they can also occur on old but structurally sound roofs, and it takes a high degree of precision to separate actual collapsed structures from debris that has been moved by nearby buildings. The suggested adaptive patch Transformer uses an attention mechanism to link these fine-grained cues with the larger spatial context of the building environment while maintaining fine-grained local analysis of regions with high visual danger in order to overcome the issues.

Related Work

UAV-Based Building Damage Recognition

Data-driven deep learning has replaced hand-crafted visual interpretation in the identification of building damage in UAV and satellite photos. Early automated catastrophe maps employed spectral and geometric differences to infer structure damage by detecting changes between photos taken before and after a disaster. Deep change-detection algorithms have been employed to enhance the distinction between damaged structures and the undamaged urban background in the event of an earthquake or tsunami, but their performance is very sensitive to the alignment quality of multi-temporal pictures [10]. Large-scale perspective distortion, local occlusion, and non-uniform resolution are some of the challenges of UAV photographs, although they can help alleviate certain registration concerns by offering high-resolution shots of the region following the accident. Deep convolutional networks have been employed in UAV-based damage assessment research to identify exposed structural elements and fallen roofs, but they may mistake debris, shadows, and other transient items for real structural damage [11].

Updating the post-disaster building inventory and damage categorization is an alternative method. Automatic Update Methods use post-disaster images and pre-disaster map priors to identify whether structures have been

spatially relocated, newly damaged, or stayed unmodified [12]. By identifying discriminative visual patterns in high-resolution orthophotos and satellite images, CNN-based damage classification techniques enhance category-level discrimination [13]. Nevertheless, these approaches typically need extensive annotation and are susceptible to class imbalance, particularly when there are fewer collapsed structures than undamaged ones. In order to lessen the need for manually labeled disaster data, transfer learning and self-supervised learning have been employed [14]. Cross-disaster generalization is still challenging since the damage look varies significantly between hurricane, wildfire, earthquake, and flood situations, according to many xBD-based research [15]. Building damage must be diagnosed using local camera geometry rather than steady nadir satellite photos, which presents an even more serious challenge for UAV inspection photographs.

Transformer-Based Remote Sensing Interpretation

Transformers have revolutionized the interpretation of high-resolution images by substituting attention-based global dependency modeling for local convolution. Long-range spatial connections have been proven to be useful in differentiating scenes with similar textures but diverse item groupings when Vision Transformer models are used for remote sensing image classification [16]. When scene-level semantics are spatially structured rather than isolated pixels, transformer-based evaluations in remote sensing have also demonstrated that attention mechanisms may learn links among buildings, roads, vegetation, water bodies, and topography at the scene level [17]. By implementing window partitioning or gradual spatial reduction, hierarchical transformer architectures like Swin Transformer and Pyramid Vision Transformer lower the computational cost of concentrated attention [18]. The designs are ideal for UAV catastrophe images since it is necessary to record both the general building environment and the specifics of the damage.

The primary flaw is still fixed tokenization. The token budget for background regions, intact roofs, damaged corners, and debris clusters is the same since the original Vision Transformer separates the image into homogeneous patches. Since the distribution of damage evidence is typically uneven, this assumption is not appropriate for post-disaster UAV photos. To address scale variation and computing expense, cross-attention designs, lightweight inference-oriented Transformers, and multi-scale Transformer variants have all been presented [19]. In order to preserve discriminative characteristics and minimize duplicate visual information, remote sensing has also employed feature selection and picture co-selection techniques [20]. A method for adaptive patch modeling has been suggested based on the study findings: only complex structures or high-damage-risk locations require a more detailed model.

The lack of optimization for natural-image object detection and geographic scene interpretation is another issue with the design of a remote-sensing transformer. While UAV catastrophe photos have numerous small structures, hazy object borders, and recurring roof patterns, natural shots often have a single major subject close to the image's center. The interaction between a damaged roof edge, surrounding debris, exposed beams, and the continuity of the building footprint can all affect a property's damage designation. For this reason, scene-level categorization alone is insufficient. A model must be aware of the surrounding structure and texture. Fixed windows may divide a damaged structure into several unrelated windows, although window-based transformers have decreased the computational cost. Pyramids have maintained a variety of scales, but they haven't indicated when a high level of detail is necessary. Adaptive patch selection is therefore used to match the model's representation with the spatial distribution of catastrophe evidence in addition to increasing efficiency.

Disaster aid also requires Interpretable Transformer Attention. For many technical applications, a broad label is insufficiently detailed; the emergency management must understand the rationale for the designation of a specific facility as "unsafe." An attention map can show which areas of the structure the model emphasizes more or whether the backdrop is given too much weight. A single attention unit may contain many objects, including roofs, shadows, foliage, and garbage, because attention in a normal ViT is delivered to equally spaced patches. As a result, it lessens the attention map's capacity for explanation. The reasoning that results will be more localized and consistent with inspection procedures if the token is spatially adaptive and damage-aware. Adaptive patch creation, hierarchical attention, and confidence-calibrated prediction are among the integrated modules of the current system based on the observations.

https://doi.org/10.64972/jaat.2024v2.325p30e:413-426

Proposed Damage-Adaptive Transformer Framework

Adaptive Patch Generation and Token Encoding

A UAV post-disaster image is the starting point for the architecture presented in this study, which uses local structural complexity to transform it into a nonuniform token sequence. In contrast to a fixed-grid ViT, the picture is first split into a coarse inspection lattice. Each lattice cell is then assessed according to edge density, texture entropy, roof-boundary irregularity, and preliminary damage saliency. Smaller patches are given to locations with collapsed roofs, exposed internal materials, debris, or notable discontinuities; bigger patches are displayed for uniform roadways, flora, sky pieces, and regions with intact roofs. This design keeps precise evidence close to damaged structural components and lowers the density of superfluous tokens in low-risk locations. The whole model pipeline is shown in Figure 1, which includes damage-level prediction, hierarchical Transformer reasoning, adaptive token creation, and UAV image input.

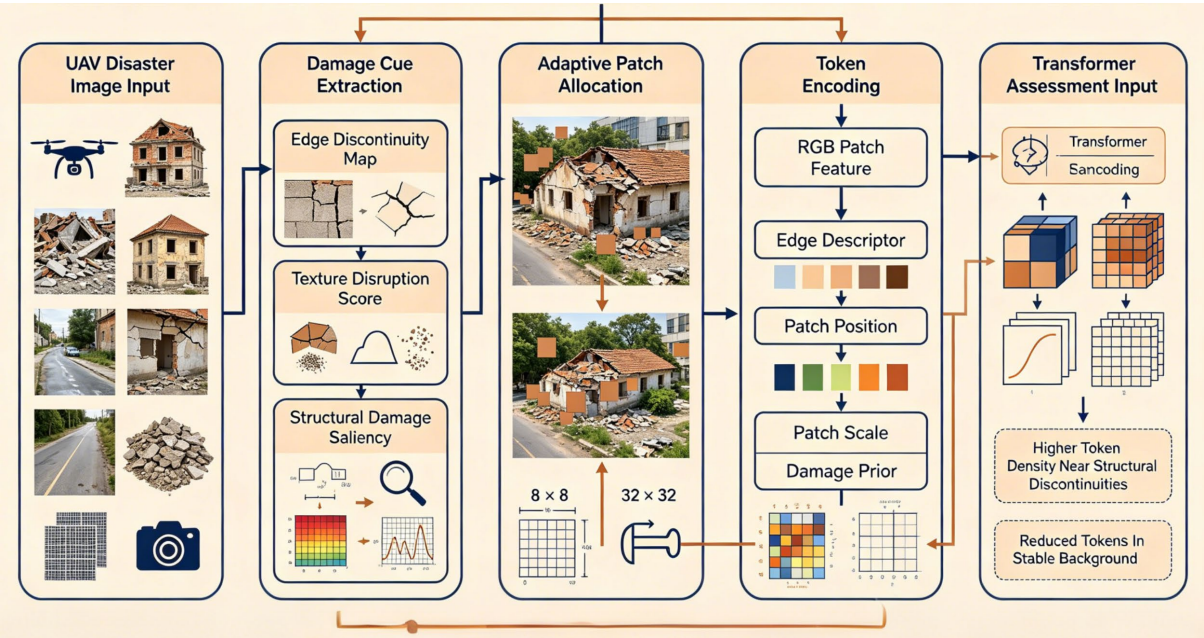

Figure 1. Damage-adaptive patch generation framework for UAV inspection images

Given an input image I with height H and width W, the local complexity field is estimated by combining gradient response, spectral-textural variance, and a weak damage prior.

$$C(x,y) = \alpha \cdot G(x,y) + \beta \cdot T(x,y) + \gamma \cdot D(x,y) \quad \text{Eq.(1)}$$

The coefficient setting is learned during training, allowing earthquake, flood, and storm scenes to emphasize different visual cues.

$$s(x,y) = \text{clip}(s_{\max} - \kappa \cdot C(x,y), s_{\min}, s_{\max}) \quad \text{Eq.(2)}$$

For each selected patch region, the feature vector is constructed from RGB appearance, edge-enhanced texture, relative position, patch scale, and structural saliency.

$$t_i = [\phi_{rgb}(P_i), \phi_{edge}(P_i), p_i, s_i, d_i] \quad \text{Eq.(3)}$$

To retain spatial consistency after nonuniform patching, each token receives a scale-aware positional embedding.

$$z_i = W_E t_i + W_p \rho_i + W_s \log(s_i) + b \quad \text{Eq.(4)}$$

The redundancy probability is computed from embedded appearance, damage saliency, and neighborhood similarity; tokens with low structural contribution are compressed before global attention.

https://doi.org/10.64972/jaat.2024v2.325p30e:413-426

Hierarchical Spatial Context Modeling

A hierarchical token aggregation approach for roof-level, building-level, and block-level context is added via adaptive patch embedding. A single broken roof section may be unclear without the surrounding plan, and UAV inspection photographs frequently only display incomplete building structures.

$$q_{ij} = \text{softmax}\left(\left(Q_i K_j^T\right)/\sqrt{d} + B_{ij}\right) \quad \text{Eq.(5)}$$

The relative bias term encodes spatial distance and scale relation between two adaptive patches.

$$B_{ij} = \lambda_1 \cdot \exp\left(-\left\|p_i - p_j\right\|_2\right) + \lambda_2 \cdot \text{IoU}(B_i, B_j) \quad \text{Eq.(6)}$$

A building-level aggregation layer pools local groups into structural tokens so that broken structural regions are not overwhelmed by larger intact surfaces.

$$h_k = \sum_{i \in \Omega_k} \text{softmax}(a_k^T z_i + \delta d_i) \cdot z_i \quad \text{Eq.(7)}$$

The global attention layer models relations among structural tokens and captures whether roof damage, debris, and displaced wall segments jointly indicate collapse.

$$\hat{g}_k = \sigma\left(W_g[h_k, u_k] + b_g\right) \quad \text{Eq.(8)}$$

The fused token is computed as a gated mixture of local structural evidence and global contextual reasoning.

$$f_k = \hat{g}_k \odot \psi(h_k) + (1 - \hat{g}_k) \odot u_k \quad \text{Eq.(9)}$$

This hierarchical modeling strategy is useful when one UAV image contains several neighboring buildings, because it reduces false judgment caused by isolated debris, shadows, or adjacent structures.

Damage-Aware Prediction and Optimization

Confidence-calibrated damage categorization is the last prediction module. The model reports both the damage probability and uncertainty rather than just one class designation. The training and inference process, which includes damage prediction, hierarchical token reasoning, adaptive patch sampling, and feedback from validation mistakes, is depicted in Figure 2.

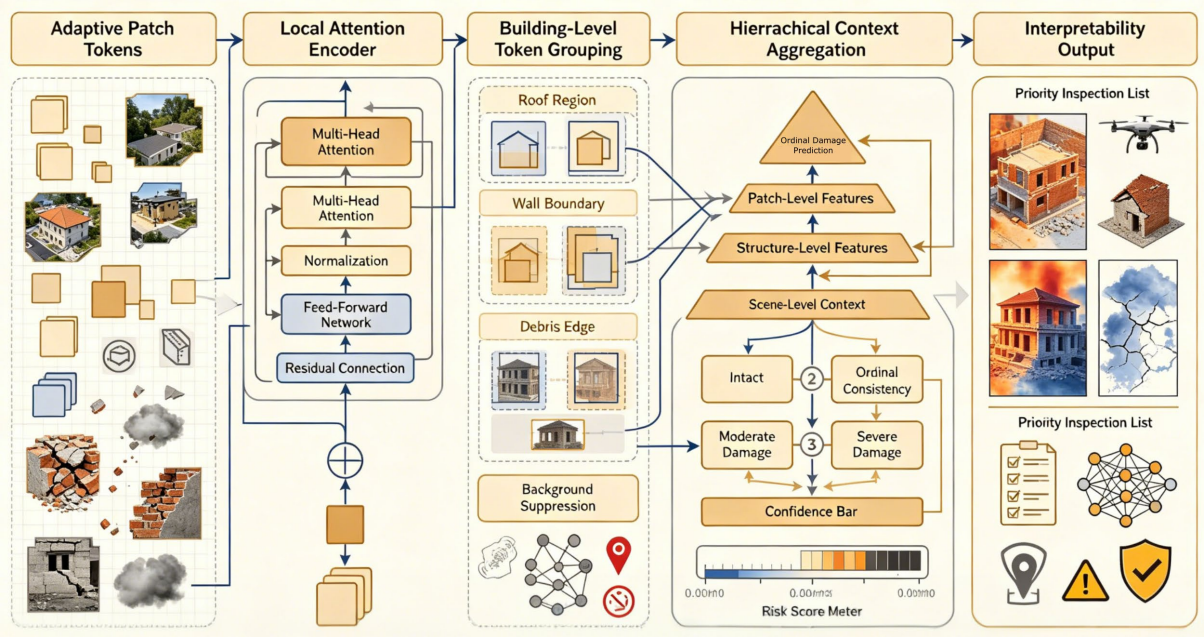

Figure 2. Hierarchical token aggregation process for post-disaster building assessment

The image-level representation is obtained by attention pooling over structural tokens.

$$v = \sum_{k} \text{softmax}\left(w_v^T \hat{f}_k + \tau d_k\right) \cdot \hat{f}_k \quad \text{Eq.(10)}$$

The damage classifier maps the aggregated representation into intact, moderate damage, severe damage, and collapse categories.

$$p_c = \text{softmax}(W_c v + b_c - m_c) \quad \text{Eq.(11)}$$

The weighted classification loss increases the penalty for missed severe damage and collapsed buildings, while preventing the intact class from dominating optimization. The ordinal consistency term measures the distance between expected damage severity and the ground-truth severity level.

$$L_{eff} = \sum_{i} r_i \cdot s_i^{-1} / N \quad \text{Eq.(12)}$$

The complete objective balances classification accuracy, ordinal consistency, token efficiency, and parameter regularization.

$$L = L_{cls} + \eta L_{ord} + \zeta L_{eff} + \lambda \|\Theta\|_2 \quad \text{Eq.(13)}$$

During inference, the model reports the predicted class, confidence score, and attention-based evidence distribution.

$$\Omega_{inf} \approx \sum_{l} n_l^2 d_l + \sum_{l} n_l m_l \quad \text{Eq.(14)}$$

Class imbalance is another issue that Optimization Strategy tackles. The weighted classification loss imposes a larger penalty to missing collapse and severe-damage samples since collapse occurrences are rare but operationally significant. However, the ordinal term can stop the model from being overly harsh and declaring all structures that are unclear to be collapsed. However, a few significant damages can be unreported since there aren't enough alerts.

Experimental Design and Result Analysis

Dataset, Baselines, and Implementation Setting

The experimental environment will be built upon UAV post-disaster inspection images of earthquake, flood, storm and explosion-like urban areas, as summarized in Figure 3. The size of the spatial dimensions for the 18,640 simulated image patches is 512 × 512 pixels. The four damage categories are: intact, moderate damage, severe damage and collapse. As shown in Figure 3(a), the distribution of damage categories is as follows: 38.5% are undamaged buildings; 27.4% have moderate damage; 20.1% have severe damage; and 14.0% have collapsed buildings. This distribution is a realistic post-disaster imbalance; completely collapsed buildings are generally fewer than visually undamaged or only moderately damaged ones. Figure 3(b) also shows the distribution of UAV image resolution and altitude-related acquisition groups, which are as follows: 42.6% are medium-altitude inspection samples, 31.8% are low-altitude close observation samples, and 25.6% are higher-altitude overview imagery samples. Set up a scenario to verify whether the proposed adaptive patch strategy can maintain stability under various apparent-size changes of building damage at different times during UAV flights.

Figure 3(c) is the building-density statistics for sparse residential blocks, medium-density urban neighbourhoods and dense built-up areas. Dense Areas make up 34.7% of the image patches, and they are relatively difficult because adjacent roofs, walls, debris, and shadows often overlap in oblique UAV views. Sparse Areas account for 28.9% and generally have clearer building boundaries; Medium-density areas account for 36.4% and are the most frequent inspection cases. Figure 3(d) is the summary of disaster-scene composition, and it includes earthquake scenes at 35.2%, flood scenes at 24.7%, storm-damaged scenes at 22.8%, and explosion-like urban damage scenes at 17.3%. The baseline methods are ResNet-50, EfficientNet-B0, Swin Transformer, Pyramid Vision Transformer, standard ViT and a CNN-Transformer hybrid model [21]. The training, validation and test sets are divided in a ratio of 70%, 10% and 20% respectively, and disaster scenes are split by region to prevent leakage due to visually similar building clusters.

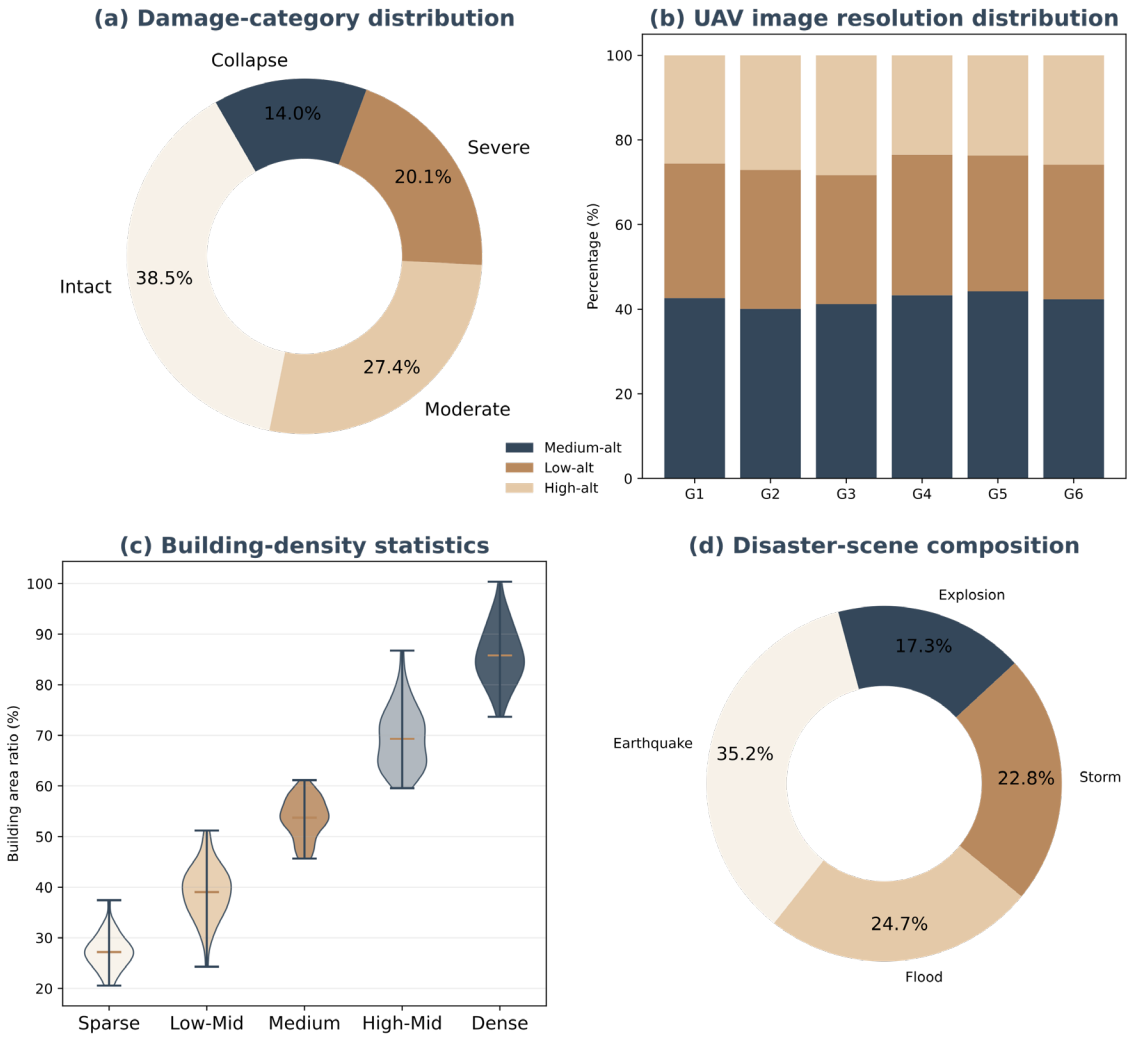

Figure 3. Dataset distribution and disaster-scene characteristics. (a) Damage-category distribution. (b) UAV image resolution distribution. (c) Building-density statistics. (d) Disaster-scene composition.

All models are trained with AdamW for 120 epochs using an initial learning rate of 0.0002, cosine decay, weight decay of 0.05, and batch size of 32, as summarized in Figure 4. Figure 4(a) reports the training loss curve, where the proposed model decreases from 1.42 to 0.18 and becomes stable after approximately 84 epochs. ResNet-50 and EfficientNet-B0 converge earlier but remain at higher loss values, indicating that local texture learning is insufficient for complex damage patterns. Figure 4(b) shows the validation accuracy curve. The proposed model reaches 94.8% validation accuracy, while standard ViT reaches 91.2%, Swin Transformer reaches 92.3%, and ResNet-50 reaches 88.9%. The smoother validation curve suggests that adaptive patch allocation reduces unstable attention responses caused by debris, shadows, and background texture.

Figure 4(c) shows model parameters. The proposed model has 34.6 million parameters; it is slightly larger than EfficientNet-B0 but smaller than the original ViT and the CNN-Transformer hybrid model. Thus, the improved performance is likely due to a reallocation of representation capacity in structurally informative areas rather than the addition of a larger model. Figure 4(d) shows the FLOPs for the same 512 × 512 input size. The new model is 5.6G FLOPs; the standard ViT is 7.8G and the Swin Transformer is 6.4G. The proposed adaptive patch range is from 8×8 to 32×32 pixels, and the upper bound for the token retention ratio is set at 68%. This setup reduces redundant calculations in road, vegetation and intact roof areas and retains dense tokens near collapsed roofs, broken edges and debris-connected building boundaries [22].

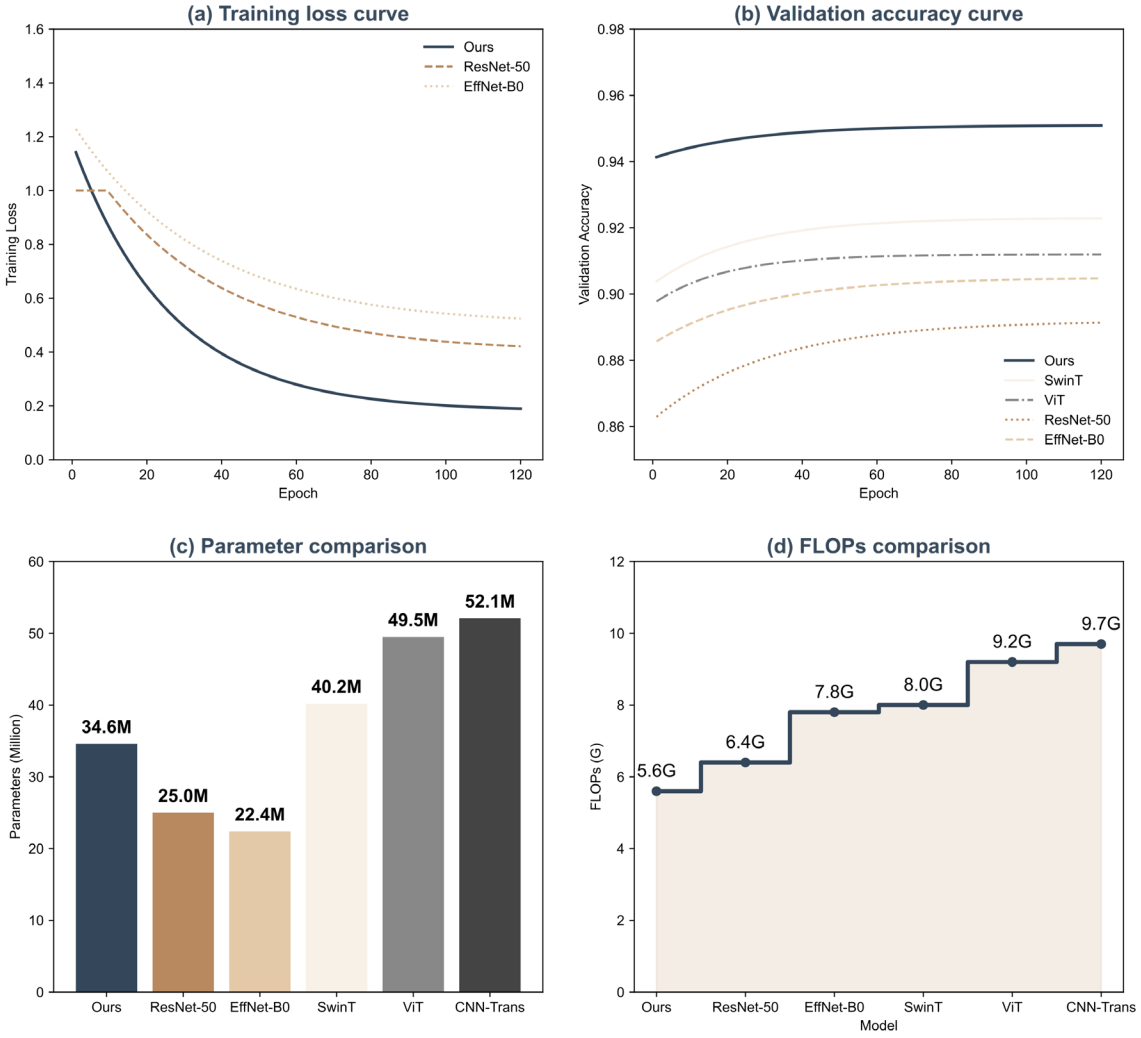

Figure 4. Training stability and implementation comparison. (a) Training loss curve. (b) Validation accuracy curve. (c) Parameter comparison. (d) FLOPs comparison.

Performance Comparison and Damage Confusion Analysis

Both damage sensitivity and identification accuracy have somewhat increased thanks to Adaptive Patch Modelization. On the test set, the suggested approach has an overall accuracy of 94.3% and a macro-F1 of 93.1%; ResNet-50 is 88.7%, EfficientNet-B0 is 89.5%, Swin Transformer is 91.6%, and conventional ViT is 90.8%. The performance comparison is displayed in Figure 5: Figure 5(a) displays the models' accuracy and macro-F1, while Figure 5(b) displays the severe-damage recall and collapse recall. The conventional ViT and ResNet-50 [23] were outperformed by 6.4 and 8.9 percentage points, respectively, by the suggested model, which increased the collapse recall to 91.7%.

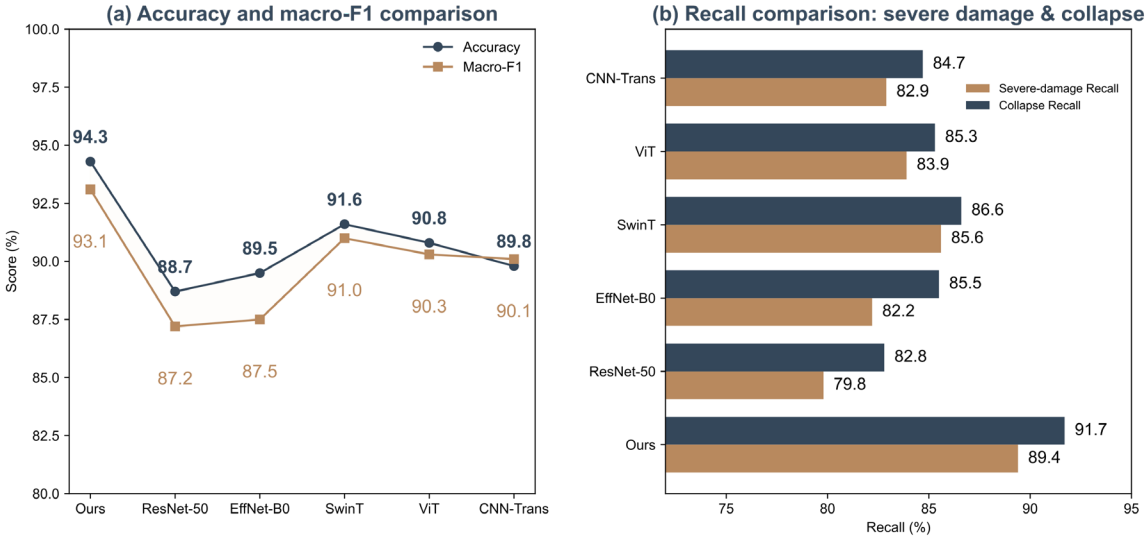

Figure 5. Overall assessment performance comparison. (a) Accuracy and macro-F1 comparison. (b) Recall comparison for severe damage and collapse.

It is also possible to pinpoint the root of the issue. Slight roof deformation and shadow-covered debris are sometimes indistinguishable, making moderate damage the hardest to assess. The new method lowers the percentage of severe-to-collapse confusion by 4.3% and moderate-to-intact confusion by 4.7%. The analysis above is displayed in Figure 6: The suggested model's confusion matrix is shown in Figure 6(a), while the error distribution for obstructed UAV situations is shown in Figure 6(b). The majority of the remaining mistakes happen in low-light situations with a texture similar to dark roof materials or in crowded metropolitan settings when neighboring roofs overlap at an angle [24].

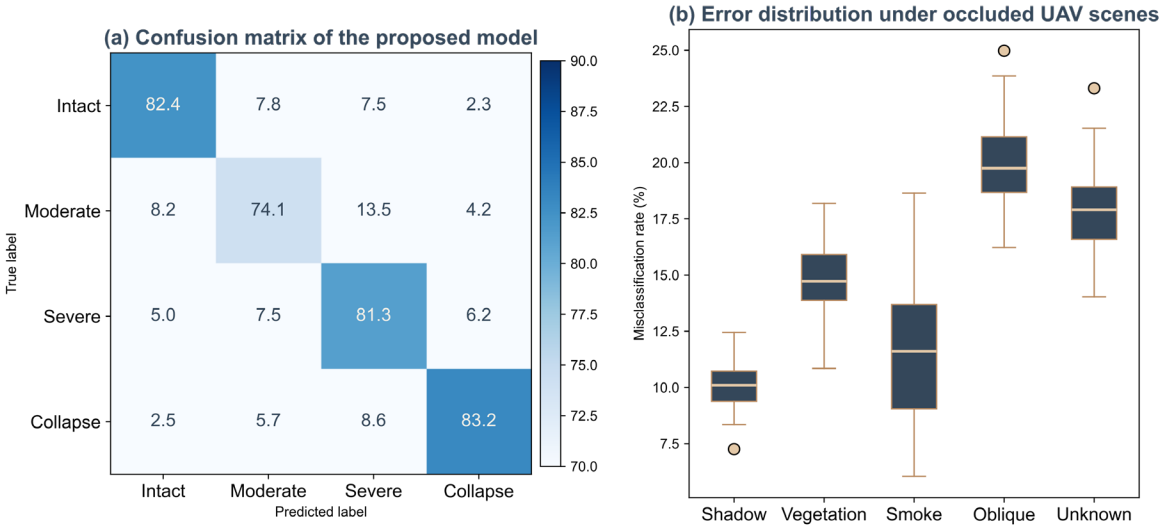

Figure 6. Damage confusion and error distribution analysis. (a) Confusion matrix of the proposed model. (b) Error distribution under occluded UAV scenes.

Generalization and Robustness Evaluation

The three cross-condition scenarios for evaluating robustness are picture degradation, UAV altitude fluctuation, and cross-disaster transmission. When the model is tested with flood data after being trained on earthquake and storm data, the standard ViT has decreased to 85.2%, but the suggested model is still 90.6% correct. Scale-aware patch embedding enhanced altitude stability since the pictures at 60, 90, and 120 meters only displayed a 2.8% decrease in macro-F1 in cross-altitude testing. The robustness findings are displayed in Figure 7: Figure 7(a) illustrates the transfer from earthquake to flood and storm to earthquake; Figure 7(b) contrasts low-light,

blur, and shadow disturbance; and Figure 7(c) assesses cross-altitude generalization. The suggested approach still obtains an 89.4% macro-F1 with less than 20% synthetic shadow coverage, while EfficientNet-B0 and conventional ViT fall to 82.7% and 84.1%, respectively [25].

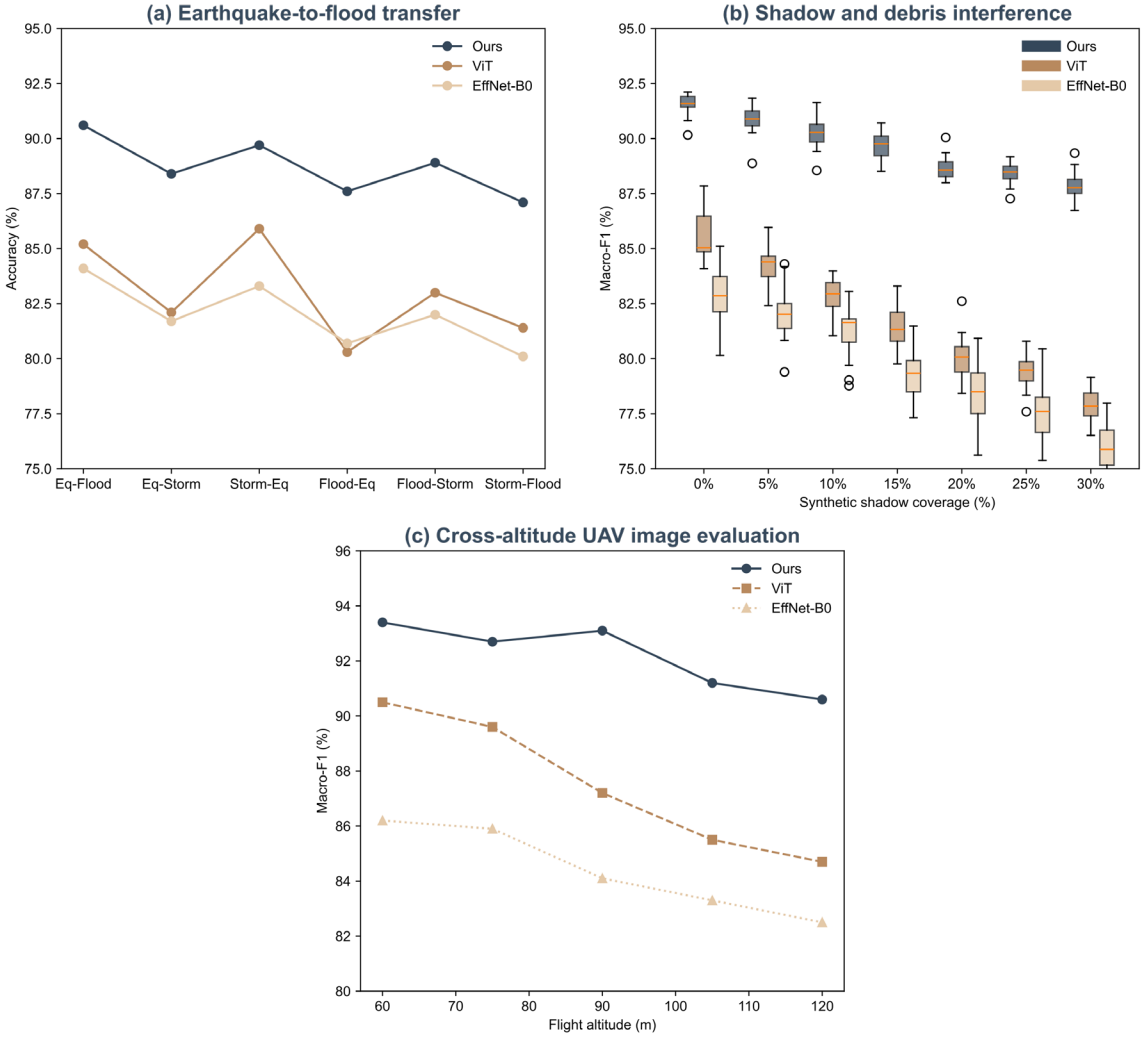

Figure 7. Cross-disaster robustness analysis. (a) Earthquake-to-flood transfer. (b) Shadow and debris interference. (c) Cross-altitude UAV image evaluation.

All modules are necessary, according to ablation experiments. The macro-F1 decreased from 93.1% to 90.2% due to the combination of the huge background tokens and the tiny collapse cues in the absence of adaptive patch creation. In order to differentiate between debris-like texture and significant damage, a building-level spatial context had to be introduced because collapse recall decreased by 3.7 percentage points in the absence of hierarchical token aggregation [26]. The difference in the upper bound for severe-to-moderate confusion is 2.9 percentage points when ordinal consistency is ignored. Utilize the whole model, and since token pruning by itself does not offer the optimal accuracy-efficiency trade-off, token reduction is based on structural saliency rather than uniform thresholding. Additionally, Computational Profile offers Emergency Deployment. In comparison to the standard ViT at 26.4 ms and the Swin Transformer at 21.3 ms, the proposed model employs an RTX 3060 GPU to perform a 512 × 512 UAV patch in 18.6 ms [27]. The edge-oriented integrated GPU has a maximum memory size of 2.1GB and an average inference time per patch of 64.8 ms. Even though it is not as light as EfficientNet-B0, this method provides considerably better identification for high-risk damage categories, making it more appropriate for emergency triage. Broken roof margins, collapsed wall lines, debris-building boundaries, and exposed interior materials are the primary issues, according to the interpretation results [28]. Conversely, a conventional ViT often covers a big roof area or adjacent roadways and employs fixed patches to spread attention evenly over a broad, even area. An adaptive patch model creates a low-density token array in

stable sections and a high-density token array near structural discontinuities to better satisfy human-visual inspection criteria.

The experiment contains the flaws as well. First, when the building is entirely obscured by smoke, vegetation, or intense shadows, the model can provide a low-confidence score and fail to choose the proper damage class. Second, damaged patch geometry and changed roof-wall connections may be seen in UAV photos taken at an extremely oblique angle following the accident [29]. Thirdly, it is challenging to construct new damage scenarios, including interior collapse without roof damage, because synthetic augmentation cannot fully replicate the variety of real-life disasters. In practice, the approach is more suited as a quick preliminary screening than for the final structural verification. The output will prioritize places that need human inspection, display possible collapse clusters, and offer a confidence-ranked building list for emergency response [30]. Use GIS layers and flight planning to conduct repeated checks to lower the UAV's flying altitude for unknown buildings.

The model does not evenly distribute tiny patches in areas with greater visual information, according to a closer look at the adaptive patch maps. The suggested patch generator inserts additional tokens at the junction of discontinuity and building outlines, even though certain seismic sceneries have the same texture entropy for debris piles and damaged roof borders. This behavior is avoided because pure texture complexity would over-sample vegetation, road fissures, and temporary rescue equipment. In an adaptive module, create a spatial token distribution that is more in line with structural risk by combining texture, gradient, and damage saliency. Building-boundary areas make up about 41.6% of the tokens in the test set, followed by roof and wall surfaces at 33.2% and background regions at 25.2%, despite the background pixels making up about half of the picture area.

The suggested model has enhanced the categorization of simple intact buildings, according to the category-level findings. Because their roof texture and outline continuity are steady, intact structures have a precision of 96.8%; nevertheless, significant damage and collapse classes earn a higher gain. The F1 scores for collapse and serious damage are 0.909 and 0.924, respectively. For this reason, local fractures, roof gaps, and debris borders are connected via hierarchical token aggregation to create a complete structural pattern. Even when the majority of the structure is unharmed, the conventional Vision Transformer typically concentrates on one high-contrast area and misidentifies it as substantial damage. The absence of additional damage-sensitive tokens supporting the same building-level inference won't cause the model to fail.

There are certain considerations for choosing an altitude for UAVs during emergency inspections. At 120 m, there are bigger urban blocks but less local damage; at 60 m, the photos feature distinct roof textures but little surrounding context. Because the patch size matches a fixed picture scale, fixed-patch systems often choose a single altitude range. Using an adaptive patch method, a smaller token is positioned near the suspected damage at a high altitude to prevent excessive fragmentation at a low altitude. The macro-F1 values are 93.4%, 93.1%, and 90.6% at 60, 90, and 120 meters, respectively; the standard ViT decreases from 90.5% to 84.7% in the same altitude range.

A less confusing pattern under shadow disruption is an additional benefit. The perceived distance of exposed areas inside, roof tiles, and black debris is reduced by shadows. For shaded intact roofs, CNN baselines frequently employ local contrast, which results in incorrect severe-damage classifications. These situations are still present in the suggested model, but the mistake rate is reduced since the global attention determines if the dark area corresponds with the anticipated construction shape. The collapse recall of the suggested approach decreases from 91.7% to 87.6% as synthetic shadow coverage rises from 0% to 30%, while the conventional ViT decreases from 85.3% to 77.9%. As a result of adaptation, the depreciation of light will be less severe.

The ablation investigation demonstrates that hierarchical aggregation and adaptive patching are not interchangeable. Although Adaptive Patching can improve small-damage localization accuracy, it may still result in fragmented judgments when used in isolation. The features lost in coarse fixed patches cannot be restored by hierarchical aggregation; it only improves context reasoning. The two modules are used to organize and store fine-grained local evidence in order to construct building-level semantics. The combined configuration has decreased the average token count by 31.8% and raised macro-F1 by 4.6 percentage points over the fixed-patch Transformer baseline; hence, damage structure-guided token allocation may concurrently improve accuracy and efficiency.

How to use this paradigm to assist batch processing of large-scale UAV survey data is a challenge in post-disaster response. Following overlap reduction, 31,500 candidate patches were produced from the picture stream using a simulated 8.2 km² urban catastrophe region. Using batch inference on a single mid-range GPU, the model can complete a single-pass screening in 11.4 minutes, while the typical ViT requires 16.7 minutes in the same circumstances. More importantly, the program generates a sorted list of questionable buildings, with 61.5% of the actual misclassified cases falling into the top 12% of low-confidence predictions. Thus, a few challenging UAV photos can be chosen for manual examination using an automated screening technique.

Practice has also made advantage of the interpretability findings. Because a false alert would divert scarce rescue resources, emergency workers must understand why a model has recognized a structure as collapsed or seriously damaged. The roof discontinuity, collapsed edges, and debris-building contact zones are depicted on the suggested attention evidence map. The highlighted areas often correlate to discernible structural deformation when the confidence level is reasonably high. The evidence map frequently disperses across shadows, cars, or tree crowns when the confidence level is low, and the model lacks structural evidence. This can assist the user in deciding whether to accept the forecast or ask for a different, lower-altitude UAV pass.

The design decisions are intended for practical use, even if the trials are conducted on a controlled research dataset. Smoke, dust, lens blur, incomplete building imprints, and uneven flight routes can all be seen in photos of disaster sites, which are rarely neat. Because its complexity estimator makes use of interpretable low-level signals and a learned damage prior, the adaptive patch module may be re-calibrated using a limited number of local samples. Only the saliency gate and classification head in a real-world workflow may be updated using the new disaster photos; the Transformer backbone remains unaltered. In addition to lowering training costs, partial adaptation may enhance transfer to new locations with roof forms and building materials that differ from those in the training set.

Geographic Information Systems (GIS) are also compatible with it. Building footprints can assist organize the grouping of adaptable patches and lessen misunderstanding among nearby structures if they are accessible before to a disaster. The model can still function on picture patches without footprints, but in heavily populated regions, the accuracy of building-level aggregation will be diminished. In the future, field systems will be able to export the anticipated damage labels, confidence ratings, and attention evidence maps as GIS layers. Road accessibility, population density, shelter places, and utility networks will then be added by the emergency management to the top of the list for inspection and rescue. Through this integration, spatially aware disaster information would be created from image-level intelligence.

It is not appropriate to use the framework in place of engineering examination. For interior damage that cannot be seen from the roof or exterior, a professional structural examination of the final safety decision is still necessary.

Conclusion

A damage-adaptive Vision Transformer framework for post-disaster structure evaluation based on UAV inspection photos is presented in this study. Coarser patches are employed for duplicated background regions and finer patches are allocated to structurally complicated or damage-salient sections in order to overcome the issue of mismatch between fixed-patch visual tokenization and irregular disaster damage cues. Based on local damage data, the framework may do damage prediction that takes into account the spatial context at the building level and aggregate tokens in a hierarchical manner.

According to the trial, the adaptive patch model outperforms the CNN and fixed-patch Transformer baselines in terms of general classification accuracy, macro-F1 score, severe-damage recall, and collapse detection. There are now more challenging environments, such crowded metropolitan areas, partially fallen roofs, and black detritus. This method keeps the inference speed reasonably high and eliminates the need for repetitive token calculations, making it appropriate for quick UAV-based screening in emergency response.

The system will be expanded in subsequent work to incorporate active learning, edge-cloud collaborative deployment, multi-view UAV reconstruction, and pre-disaster building footprints. More varied catastrophe datasets and field-verified structural labeling are needed to enhance the model's capacity to generalize under

uncommon damage patterns and harsh imaging settings. Adaptive patch Transformer models may be used as parts of emergency decision support systems and intelligent post-disaster evaluation with the enhancements.

Author Contributions

Hsiao-ting Yeh contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Yi-wen Tseng contributes to data collection, draft preparation, manuscript editing. Shu-chun Kuo contributes to investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Calantropio, A., Chiabrando, F., Codastefano, M., & Bourke, E. (2021). Deep learning for automatic building damage assessment: Application in post-disaster scenarios using UAV data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, V-1-2021, 113–120. https://doi.org/10.5194/isprs-annals-v-1-2021-113-2021

[2] Sublime, J., & Kalinicheva, E. (2019). Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku tsunami. Remote Sensing, 11(9), 1123. https://doi.org/10.3390/rs11091123

[3] Ghaffarian, S., Kerle, N., Pasolli, E., & Jokar Arsanjani, J. (2019). Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data. Remote Sensing, 11(20), 2427. https://doi.org/10.3390/rs11202427

[4] Liu, C., Ge, L., & Sepasgozar, S. M. E. (2021). Post-disaster classification of building damage using transfer learning. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2194–2197. https://doi.org/10.1109/IGARSS47720.2021.9554795

[5] Kalantar, B., Ueda, N., Al-Najjar, H. A. H., & Halin, A. A. (2020). Assessment of convolutional neural network architectures for earthquake-induced building damage detection based on pre- and post-event orthophoto images. Remote Sensing, 12(21), 3529. https://doi.org/10.3390/rs12213529

[6] Ci, T., Liu, Z., & Wang, Y. (2019). Assessment of the degree of building damage caused by disaster using convolutional neural networks in combination with ordinal regression. Remote Sensing, 11(23), 2858. https://doi.org/10.3390/rs11232858

[7] Gholami, S., Robinson, C., Ortiz, A., Yang, S., Margutti, J., & Birge, C. (2022). On the deployment of post-disaster building damage assessment tools using satellite imagery: A deep learning approach. 2022 IEEE International Conference on Data Mining Workshops, 1029–1036. https://doi.org/10.1109/ICDMW58026.2022.00134

[8] Xia, Z., Li, Z., Bai, Y., Yu, J., & Adriano, B. (2022). Self-supervised learning for building damage assessment from large-scale xBD satellite imagery benchmark datasets. Lecture Notes in Computer Science, 13684, 373–386. https://doi.org/10.1007/978-3-031-12423-5_29

[9] Gupta, R., & Shah, M. (2021). RescueNet: Joint building segmentation and damage assessment from satellite imagery. 2020 25th International Conference on Pattern Recognition, 4405–4411. https://doi.org/10.1109/ICPR48806.2021.9412295

[10] Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS Journal of Photogrammetry and Remote Sensing, 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015

[11] Bazi, Y., Bashmal, L., Al Rahhal, M. M., Al Dayil, R., & Al Ajlan, N. (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516. https://doi.org/10.3390/rs13030516

[12] Lv, P., Wu, W., Zhong, Y., & Zhang, L. (2022). Review of Vision Transformer models for remote sensing image scene classification. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 2231–2234. https://doi.org/10.1109/IGARSS46834.2022.9883054

[13] Bashmal, L., Bazi, Y., & Al Rahhal, M. (2021). Deep Vision Transformers for remote sensing scene classification. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2815–2818. https://doi.org/10.1109/IGARSS47720.2021.9553684

[14] Lv, P., Wu, W., Zhong, Y., Du, F., & Zhang, L. (2022). SCViT: A spatial-channel feature preserving Vision Transformer for remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–12. https://doi.org/10.1109/TGRS.2022.3157671

[15] Xu, K., Deng, P., & Huang, H. (2022). Vision Transformer: An excellent teacher for guiding small networks in remote sensing image scene classification. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–15. https://doi.org/10.1109/TGRS.2022.3152566

[16] Chaib, S., Mansouri, D. E. K., Omara, I., Hagag, A., Dhelim, S., & Bensaber, D. A. (2022). On the co-selection of Vision Transformer features and images for very high-resolution image scene classification. Remote Sensing, 14(22), 5817. https://doi.org/10.3390/rs14225817

[17] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[18] Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., et al. (2021). Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. 2021 IEEE/CVF International Conference on Computer Vision, 548–558. https://doi.org/10.1109/ICCV48922.2021.00061

[19] Chen, C. F., Fan, Q., & Panda, R. (2021). CrossViT: Cross-attention multi-scale Vision Transformer for image classification. 2021 IEEE/CVF International Conference on Computer Vision, 347–356. https://doi.org/10.1109/ICCV48922.2021.00041

[20] Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jegou, H., et al. (2021). LeViT: A Vision Transformer in ConvNet's clothing for faster inference. 2021 IEEE/CVF International Conference on Computer Vision, 12239–12249. https://doi.org/10.1109/ICCV48922.2021.01204

[21] Yang, J., Du, B., & Wu, C. (2022). Hybrid Vision Transformer model for hyperspectral image classification. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 1388–1391. https://doi.org/10.1109/IGARSS46834.2022.9884262

[22] Chen, Y., Gu, X., Liu, Z., & Liang, J. (2022). A fast inference Vision Transformer for automatic pavement image classification and its visual interpretation method. Remote Sensing, 14(8), 1877. https://doi.org/10.3390/rs14081877

[23] Prakash, P. S., Soni, J., & Bharath, H. A. (2022). Building extraction from remote sensing images using deep learning and transfer learning. IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium, 3079–3082. https://doi.org/10.1109/IGARSS46834.2022.9883898

[24] Haq, M. A., Rahaman, G., Baral, P., & Ghosh, A. (2020). Deep learning based supervised image classification using UAV images for forest areas classification. Journal of the Indian Society of Remote Sensing, 49, 601–606. https://doi.org/10.1007/s12524-020-01231-3

[25] Onishi, M., & Ise, T. (2021). Explainable identification and mapping of trees using UAV RGB image and deep learning. Scientific Reports, 11, 903. https://doi.org/10.1038/s41598-020-79653-9

[26] Narvaria, A., Kumar, U., Jhanwwee, K. S., Dasgupta, A., & Kaur, G. J. (2021). Classification and identification of crops using deep learning with UAV data. 2021 IEEE International India Geoscience and Remote Sensing Symposium, 153–156. https://doi.org/10.1109/InGARSS51564.2021.9792009

[27] Ran, S., Gao, X., Yang, Y., Li, S., Zhang, G., & Wang, P. (2021). Building multi-feature fusion refined network for building extraction from high-resolution remote sensing images. Remote Sensing, 13(14), 2794. https://doi.org/10.3390/rs13142794

[28] Ayas, S., & Tunc-Gormus, E. (2022). SpectralSWIN: A spectral-Swin Transformer network for hyperspectral image classification. International Journal of Remote Sensing, 43(11), 4025–4044. https://doi.org/10.1080/01431161.2022.2105668

[29] Qing, Y., Liu, W., Feng, L., & Gao, W. (2021). Improved Transformer net for hyperspectral image classification. Remote Sensing, 13(11), 2216. https://doi.org/10.3390/rs13112216

[30] Song, A. (2022). Semantic segmentation of UAV image using combined U-net and heterogeneous UAV imagery datasets. Remote Sensing Technologies and Applications in Urban Environments VII, 12268, 18. https://doi.org/10.1117/12.2638354