Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2025.30586

Improving the Assessment of Post-Earthquake Building Damage in Underdeveloped Regions with Vision Transformers

William Lu[1*], Zhanming Yang[2]

[1]*St. George's Senior School, Vancouver, Canada*
[2]*Semiahmoo Secondary School, Surrey, Canada*
**Corresponding Author. Email: luw9072@gmail.com*

Abstract. Assessing building damage following an earthquake is vital for first responders to effectively target their efforts in disaster-stricken areas. Satellite imagery is a powerful tool for visualizing such damage, particularly in underdeveloped regions where infrastructure is limited and access to disaster sites is challenging. This paper focuses on enhancing the process of classifying building damage following an earthquake by utilizing high-resolution satellite imagery and the state-of-the-art Vision Transformers (ViT) model. Experiments are carried out using two real-world datasets from the Ludian and Yushu earthquakes, contrasting the effectiveness of ViTs with sophisticated CNNs like ResNet50, Inception-V3, and EfficientNet-B0. The results show that ViT can more effectively assess rural buildings compared to other models. It also demonstrated better generalization across different earthquake scenarios and stays robust when trained and tested on smaller datasets. Furthermore, we proposed a new architecture that incorporates a CNN-based Inception module so the local features on the damaged buildings can be better extracted. improved the model's ability. The results show that ViTs have the potential to be a reliable and powerful tool for constructing damage assessments in disaster-affected areas, providing a more precise and effective solution than conventional CNN-based techniques. Furthermore, the proposed Inception-enhanced ViT architecture presents a viable path for further study, with the potential to be refined and validated on larger datasets and applied to a broader range of disaster scenarios

Keywords: Vision Transformers, Graph recognition, Earthquake

1. Introduction

One of nature's most unpredictable and destructive calamities, earthquakes pose serious risks to infrastructure and human life. The destruction caused by these events often results in considerable casualties and economic losses, particularly in underdeveloped and rural areas. These regions frequently lack the robust infrastructure and resources necessary to withstand seismic activity, making them especially vulnerable during earthquakes. Moreover, their remoteness often complicates the efforts of first responders, delaying critical aid and exacerbating the impact of the

DOI: 10.54254/2755-2721/2025.30586

disaster. Quick evaluation of building damage in the wake of an earthquake is essential to preventing additional casualties and property damage.

Traditional field surveys, while accurate, are often time-consuming, unsafe, and labor-intensive in nature [1]. These limitations make them inadequate for providing the timely and actionable information required for an effective emergency response. The development of remote sensing technologies, which utilize sensors and satellite imagery, offers a promising alternative. These technologies allow quicker and more accurate assessments of building damage over large areas, providing a more efficient and cost-effective solution, especially in regions that are difficult to access and lack resources.

A number of techniques have been developed recently to use satellite imagery to evaluate building damage following an earthquake. Based on the data that they employ, these techniques typically fall into two categories: single-temporal and multi-temporal evaluations. While single-temporal data is more readily available, multi-temporal data compares images from before and after the earthquake, which is often more challenging to acquire and less practical in real-world disaster scenarios [2].

Traditional machine-learning methods, including Decision Trees, Random Forests, and Support Vector Machines, were used in the early research in this area [3]. These methods often employ object-oriented classification algorithms, which have limitations in feature selection and segmentation [4,5].

The advent of deep learning has led to the development of Convolutional Neural Networks (CNNs) that significantly outperform traditional machine learning models in post-disaster damage classification [6-9]. Despite these advancements, the application of ViT for this task remains largely unexplored. While CNNs have proven effective in many scenarios, they are inherently limited by their reliance on localized feature extraction through convolutional layers, which can restrict their ability to capture global context and long-range dependencies within an image. This limitation is particularly problematic when analyzing complex post-disaster scenes, where understanding the broader spatial relationships between different structures is crucial for accurate damage assessment.

Moreover, CNNs often require extensive fine-tuning and complex architectures to handle multi-scale information, which can increase computational costs and reduce efficiency. Vision Transformers (ViTs), with their self-attention mechanisms, offer a promising alternative by inherently capturing both local and global features within the data. However, ViTs also present challenges, particularly in their data-hungry nature and the need for large datasets to achieve optimal performance. This can be a significant barrier in disaster scenarios where data may be limited or difficult to obtain.

This paper makes two key contributions to the field of post-earthquake building damage assessment. First, it applies the ViT model to classify building damage in underdeveloped regions, offering insights into its performance in a task where ViTs have not been extensively tested. Second, it introduces a novel framework that enhances ViT's performance by integrating an Inception CNN module, improving the model's ability to extract local features critical for accurate damage assessment.

The rest of the paper is organized as follows. Section 2 analyzes related works, their insights, and limitations. Section 3 introduces the experiment methodology, including the dataset and model. Section 4 discusses the results and their implications. Lastly, we conclude the study and ways this area for future research.

DOI: 10.54254/2755-2721/2025.30586

2. Related works

CNNs have served as the backbone of deep learning approaches for image classification, including post-disaster building damage assessment. Their capability to automatically learn both high-level and specific features from images has made them highly effective for this purpose [10]. Low-level CNN layers extract basic features such as color, edges, and corners, while deeper layers capture high-level, task-specific features. For instance, Ci et al. [6] developed a model for earthquake building damage classification that combined a CNN with ordered regression, demonstrating significantly higher accuracy compared to traditional machine learning methods. Similarly, Berezina et al. [8] utilized CNNs to assess hurricane damage, emphasizing CNNs' ability to capture subtle, high-level features critical for distinguishing between varying degrees of damage in disaster scenarios.

However, post-disaster images often exhibit significant changes in the shape and texture of buildings, which challenges CNN's feature extraction capabilities. Such variations can blur the distinction between different levels of damage, leading to misclassifications, particularly between slight and moderate damage [4]. Consequently, several studies have explored the integration of additional texture, spatial, or temporal information to improve classification accuracy. For example, Ji et al. [11] combined CNN-derived features with GLCM texture features (including contrast, entropy, and homogeneity) extracted from pre- and post-event images, enhancing the model's ability to assess damage in the 2010 Haiti Earthquake. Qing et al. [12] introduced a complex workflow incorporating extra feature enhancement bands (EFEBs) and multi-spectral information for more accurate damage analysis. Similarly, Zheng et al. [13] employed a Siamese architecture, using a two-step process that involved different networks to analyze pre- and post-disaster images, enhancing feature and change detection.

While these methods improve CNN performance, they also introduce challenges, such as increased model complexity and computational demands, which can lead to overfitting and reduce effectiveness in real-time or resource-constrained scenarios. Moreover, the reliance on bitemporal data—requiring both pre-and post-disaster images—complicates data access and delays analysis in real-world situations. These challenges underscore the need for new approaches that can effectively capture contextual global features and understand building conditions using only post-event satellite images.

Another significant limitation of existing post-disaster models is their dependence on large amounts of training data, which is often difficult to obtain quickly during an emergency response. Additionally, these models tend to focus on single datasets, limiting their ability to generalize to new earthquake scenarios. To address these limitations, Lin et al. [9] explored the use of transfer learning with CNNs for seismic damage assessment, demonstrating that pre-trained CNNs could be fine-tuned for specific disaster scenarios, thereby enhancing performance through the transfer of learned features from related tasks. Qing et al. [12] similarly found that transfer learning improved CNN performance in domain-specific tasks like landslide detection.

ViTs, which utilize the self-attention mechanism, have emerged as a promising alternative to CNNs, particularly for tasks that require capturing global context. The self-attention mechanism enables ViTs to consider relationships between all parts of an image. Dosovitskiy et al. [14] showed that ViTs could outperform traditional CNNs in image recognition tasks when sufficient data is available. ViTs have also shown potential in building damage classification tasks. Bazi et al. [15] applied ViTs to land use classification with satellite imagery, showing strong performance for remote sensing data. Kaur et al. [16] proposed a hierarchical vision transformer, demonstrating ViT's

DOI: 10.54254/2755-2721/2025.30586

flexibility and scalability in assessing building damage on a large scale. Ramesh et al. [17] applied ViTs to post-hurricane building classification, successfully outperforming several CNN models.

Despite these promising developments, the application of ViTs for assessing buildings in post-earthquake settings remains limited. Moreover, studies have yet to investigate the effectiveness of ViTs in underdeveloped regions, where geographical conditions and building types might affect model performance. Most research to date has utilized the xBD dataset, which provides general post-disaster imagery from diverse global scenarios [1]. However, this dataset predominantly includes urban and semi-urban regions with well-documented infrastructure, potentially limiting the generalizability of the findings to underdeveloped areas.

3. Methodology

3.1. Dataset

The 2010 Yushu earthquake in Qinghai Province and the 2014 Ludian earthquake in Yunnan Province are two major earthquakes that occurred in impoverished parts of China that are the subject of this study's high-resolution remote sensing datasets. The Yushu dataset was obtained on April 16, 2010, two days after a 7.1-magnitude earthquake struck. These UAV images featured RGB spectral bands with a 0.1-meter spatial resolution, providing a clear and detailed visual record of the widespread building collapse in Jiegu Town. The Ludian dataset consists of aerial images captured on August 4, 7, and 14, 2014, following a 6.5-magnitude earthquake. The dataset similarly contained high-resolution RGB image with a 0.2-meter spatial resolution, sampling collapsed buildings near Longtoushan town.

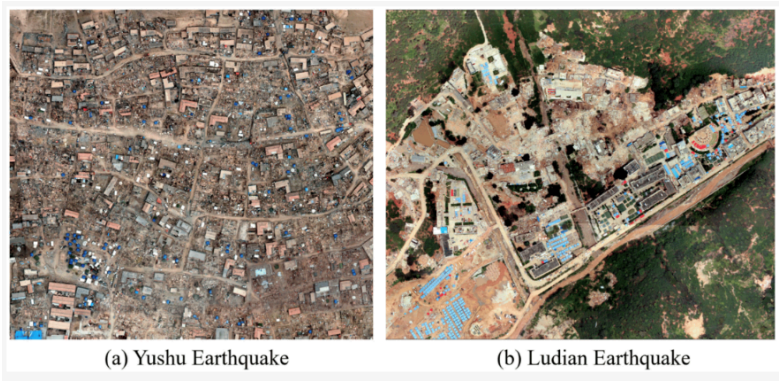

Figure 1. Satellite image from the two sites assessed (a) Yushu Earthquake (b) Ludian Earthquake

As shown in Figure 1, traits of the two reigns, Yushu and Ludian, could be assessed through satellite images. The unique geographical and architectural features of the Yushu and Ludian regions make building classification more challenging but significant. The Ludian region is surrounded by mountains, populated by low-lying masonry-timber and soil-timber structures in villages. These structures are harder to distinguish from the surrounding environment, while green trees and vegetation have also covered the tops of some houses. Meanwhile, the Yushu dataset is characterized by a rugged and mountainous terrain. The dataset similarly comprises low-rise structures, likely traditional Tibetan architecture constructed with locally sourced materials. While these features add complexity and challenge for computer vision models, they simultaneously reveal the lack of infrastructure and difficulty of access.

DOI: 10.54254/2755-2721/2025.30586

The dataset of different damage degrees was built by Ci et al [6], centered and framed to 88 × 88 pixels to capture the essential building features. Building damage was divided into four categories: completely intact (no visible damage; D0 damage grade); lightly damaged (some breakage and cracking; D1 damage grade); heavily damaged (significant deformations across cracks in load-bearing elements; D2 damage); and collapsed (full collapse of the structure or part of a floor; D3 damage). Building characteristics, including outline, geometry, texture, and their interaction with the surrounding environment, were taken into consideration when classifying these structures. While both datasets share the same damage classes, criteria, and economic situation, the two datasets have major differences in geographical region, local conditions, and image sampling methods. It is important for models to be effective across these disparities in real-world scenarios.

The Ludian dataset is significantly more abundant, with 13,780 samples, while the Yushu dataset contains 3501 samples. For D0 damage grade, Ludain Dataset has a number of 2680 while Yushu Dataset has a number of 778; For D1 damage grade, Ludain Dataset has a number of 5013 while Yushu Dataset has a number of 918; For D2 damage grade, Ludain Dataset has a number of 2807 while Yushu Dataset has a number of 665; and For D3 damage grade, Ludain Dataset has a number of 3280 while Yushu Dataset has a number of 1140; Visualization of damage grade in both datasets can be observed through Figure 2 below (Sample of damaged buildings from each class in both earthquakes). The size shows equal distribution across four categories.

Our study uses the larger Ludian dataset to maximize the performance of pre-trained ViTs, also allowing us to more thoroughly analyze its efficacy in comparison to CNN models. The performance of transfer learning models will also be tested on unseen Yushu samples to evaluate its ability to be applied to different eathquake scenarios. Meanwhile, the performance of our improved ViT model incorporating a CNN module will be compared and tested on the Yushu dataset.

Figure 2. Sample of damaged buildings from each class in both earthquakes (Lin et al, 2019)

3.2. Data augmentation

In this study, we applied data augmentation techniques to increase the variability of the dataset. This approach effectively reduces overfitting and helps the model generalize to unseen data. Techniques that have been used include: (1) random horizontal flip, (2) random vertical flip: These techniques involve flipping the images horizontally or vertically with a certain probability. It helps the model learn that the orientation of the image may vary, especially in satellite imagery where buildings and structures can appear in different orientations. (3) random rotation: This technique randomly rotates the images within a specified range of degrees. By introducing different angles, the model becomes more adept at recognizing buildings and damage from various perspectives. and (4) random affine: Affine transformations involve a combination of scaling, rotation, translation, and shearing, altering

DOI: 10.54254/2755-2721/2025.30586

the geometry of the images. This method introduces slight distortions, such as tilting or skewing the images, making the model more resilient to changes in shape.

3.3. Implementation details

Stochastic Gradient Descent, with a batch size of 16, was the optimizer utilized in the transfer learning experiment because it accelerated the model's convergence and enhanced generalization. The learning rate was set as 0.001 (It's high enough to ensure that the pre-trained model quickly adapts to the new task but not too high to cause large, unstable updates to the weights). For training and validating our proposed model and ViT with no pretraining, the same optimizers were used for this experiment. The batch size was set as 32 because it can stabilize the gradient estimates and typically results in faster convergence. The learning rate was set as 0.0003 for the hybrid model, which is lower than the rate used for transfer learning. A lower learning rate is more beneficial when training more complex models from scratch or in hybrid architectures because it allows for more gradual updates to the model parameter. For both datasets, the images are split in 60/20/20 training/validation/testing dataset. The images are transformed into [224, 224] and normalized.

3.4. Metrics

The four following metrics were used to measure the model's performance, including total accuracy, precision, recall, F1 and confusion matrix.

(1) Total Accuracy

Description: Accuracy measures the proportion of correctly classified instances out of the total instances. It gives an overall indication of how well the model is performing across all classes.

Equation: $\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$ where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

(2) Precision:

Description: Precision is the ratio of correctly predicted positive observations to the total predicted positives. It indicates how many of the instances predicted as positive are actually positive.

Equation: $\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$

Precision is especially useful when the cost of false positives is high.

(3) Recall:

Description: Recall (or sensitivity) is the ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to capture all relevant instances of the positive class.

Equation: $\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$

High recall is crucial when the cost of false negatives is high.

(4) F1 Score:

The F1 score is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful in situations where there is a class imbalance.

Equation: $\text{F1}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

The F1 score gives a single metric that balances the trade-offs between precision and recall.

DOI: 10.54254/2755-2721/2025.30586

3.5. CNN models

CNNs are a class of deep learning models specifically designed for processing data that has a grid-like topology, such as images. The core idea behind CNNs is the use of convolutional layers that apply a set of filters (kernels) across the input data to extract features. Each filter convolves over the input image, producing feature maps that capture local patterns such as edges, textures, and other low-level features. As the data passes through successive layers of convolutions, pooling, and non-linear activation functions, the network builds increasingly abstract representations, capturing complex patterns and hierarchical features of the image.

ResNet, Inception V3, and EfficientNet are three prominent CNN architectures commonly used in image classification tasks. ResNet, short for Residual Network, introduced by He et al. [18], addresses the vanishing gradient problem by using residual connections, allowing very deep networks to be trained effectively. Szegedy et al. [19] created Inception V3, which uses a complicated architecture with numerous filter sizes inside the same layer to capture features at different scales and enhance the model's processing capability for a variety of image structures. Tan and Le's [20] EfficientNet uses a compound scaling method to scale the network's depth, width, and resolution in a balanced way. This allows for great accuracy with a reduced number of parameters and processing resources.

3.6. Vision transformers

While CNNs have traditionally excelled in computer vision by capturing local features through layers of convolutions, their ability to grasp global context is constrained by the fixed receptive fields of convolutional filters. Transformers, with their self-attention mechanisms, can model global dependencies effectively. ViTs exploit this by treating images as sequences of patches, analogous to how text is processed in NLP tasks, allowing the model to capture both local and global features with greater efficiency.

The architecture of ViT consists of a number of essential parts. The input image is first divided into fixed-size, non-overlapping patches by the ViT model; these patches are usually 16 by 16 pixels in size. After that, each patch is made into an embedding vector by flattening it and passing it through a linear projection layer. This procedure is similar to how NLP tasks embed sentence words into vectors. The end product is a series of patch embeddings, each of which represents a distinct area of the picture. Positional encodings are added to the patch embeddings because Transformers do not naturally comprehend the order or position of the patches. The model is able to preserve its comprehension of the image's structure because to these encodings, which give it information on the spatial arrangement of the patches inside the original image.

3.6.1. Transformer encoder blocks

The core of the Vision Transformer architecture consists of a series of Transformer encoder blocks. Each block includes (1) Multi-Head Self-Attention: This mechanism allows the model to weigh the importance of each patch relative to others, capturing dependencies across the entire image. Multi-head attention further refines this process by allowing the model to focus on different aspects of the relationships simultaneously. (2) Feedforward Neural Network: Following the attention mechanism, each Transformer block contains a fully connected feedforward network that processes the attention outputs, introducing non-linearity and further transforming the features. (3) Layer Normalization and Residual Connections: To stabilize and improve the learning process, each block includes layer

DOI: 10.54254/2755-2721/2025.30586

normalization and residual connections. These components ensure that the model can train efficiently and that infomation can flow through the network without degradation. (4) The attention mechanism plays a pivotal role in capturing relationships between different patches of an image, enabling the model to build a comprehensive global representation. The attention mechanism in ViT is implemented through multi-head self-attention layers (MSA), where multiple independent attention heads operate in parallel. Each head processes the input independently, and the results are then concatenated and projected back into the original feature space through a linear transformation.

The inputs to an attention head include the key K, query Q, and value V matrices, which are derived from the input embeddings $\mathrm{Z} \in \mathrm{R}^{\mathrm{L} \times \mathrm{D}}$ where L is the sequence length, and D is the embedding dimension) as follows:

$$\mathrm{K} = \mathrm{ZW_k}$$

$$\mathrm{Q} = \mathrm{ZW_q}$$

$$\mathrm{V} = \mathrm{ZW_v}$$

where , $\mathrm{W_k}, \mathrm{W_q}$, and $\mathrm{W_v} \in \mathrm{R}^{\mathrm{D} \times \mathrm{d_k}}$ are learnable weight matrices, and $\mathrm{d_k}$ is the dimension of the key and query vectors. The self-attention mechanism for a single attention head is computed as:

$$\text{Attention}\left(\mathrm{Q}, \mathrm{K}, \mathrm{V}\right) = \text{Softmax}\left(\frac{\mathrm{QK}^{\mathrm{T}}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V}$$

This operation allows the model to compute the weighted sum of the value vectors, where the weights are determined by the similarity between the query and key vectors.

The multi-head self-attention mechanism combines the outputs from multiple attention heads:

$$\text{MSA}\left(\mathrm{Z}\right) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)\mathrm{W_o}$$

where each $\text{head}_\mathrm{j} = \text{Attention}\left(\mathrm{Q_j}, \mathrm{K_j}, \mathrm{V_j}\right)$, h is the number of attention heads, and $\mathrm{W_o} \in \mathrm{R}^{h \cdot \mathrm{d_k} \times \mathrm{D}}$ is the output projection matrix. This multi-head approach allows the model to attend to different parts of the image simultaneously, capturing various aspects of the relationships between image patches. These self-attention layers are crucial for learning complex dependencies and interactions across different regions of buildings in our study, allowing the model to develop a rich and holistic understanding of the data.

3.6.2. MLP classification head

The MLP classification head (the whole classifying process is shown in Figure 3, with an arrow pointing down each stage) in our model consists of two fully connected layers, along with a ReLU activation function and dropout for regularization. The first dense layer compresses the transformer output into a more compact feature representation, which is then passed through a ReLU activation to introduce non-linearity. Dropout is applied to the inputs and outputs of this layer to prevent overfitting. A second dense layer is used for progressive refinement, capturing more specific and detailed attributes, outputting logits for a SoftMax function and multi-class prediction.

DOI: 10.54254/2755-2721/2025.30586

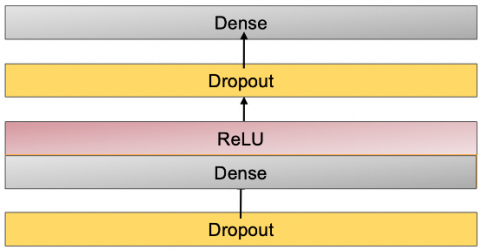

Figure 3. MLP classifier for building- damage classification

3.6.3. Proposed model pipeline

This paper proposes a model that integrates the Vision Transformer (ViT) architecture with a parallel Inception module to enhance feature extraction, leveraging the strengths of both Transformers and CNN. This approach is particularly effective for complex image classification tasks, such as post-disaster building damage assessment.

(1) Patch Embedding and Positional Encoding

The model begins by dividing the input image I of size $H \times W \times C$ (where H, W, and C represent the height, width, and number of channels, respectively) into a grid of non-overlapping patches, each of size $P \times P$. The image is thus divided into $N = \frac{H \times W}{P^2}$ patches. Each patch P_i is then flattened into a 1D vector and passed through a linear embedding layer to produce an embedding vector e_i of dimension D The sequence of patch embeddings E is represented as:

$$E = [e_1, e_2, \ldots, e_N]$$

A classification token (CLS) is appended to the beginning of this sequence, and positional encodings are added to retain spatial information, resulting in the final input sequence for the Transformer block:

$$E = [CLS; e_1 + p_1; e_2 + p_2; \ldots; e_N + p_N]$$

where p_i is the positional encoding corresponding to patch P_i

(2) Inception Module

In our model, we integrate an inception module to capture local features. To maintain coherence for the overall model and consistent feature processing across the network, we reshape the sequence of patch embeddings into a 2D spatial format and apply the CNN. This approach allows the CNN to operate within the same high-dimensional feature space as the Transformer, facilitating seamless integration of local and global features.

Specifically, the sequence E of shape (B, N, D) where B is the batch size, N is the number of patches, and D is the embedding dimension, is reshaped into a 3D tensor E_{img} of shape $\left(B, D, \frac{H}{P}, \frac{W}{P}\right)$, which corresponds to a feature map with D channels and spatial dimensions $\frac{H}{P} \times \frac{W}{P}$.

The reshaped embeddings E_{img} are then passed through the Inception module. The Inception module consists of multiple parallel convolutional branches, each designed to capture features at different scales:

DOI: 10.54254/2755-2721/2025.30586

$$\text{Feature Map}=f_{\text{inception}}(E_{\text{img}})$$

The output feature map retains the same spatial dimensions $\frac{H}{P} \times \frac{W}{P}$. and has a depth of D channels. This feature map is then flattened and reshaped back into the sequence format to match the original embedding dimensions, resulting in a new sequence E_{CNN} of shape (B, N, D)

(3) Transformer Block and Fusion

In parallel, the original sequence of embeddings E is passed through the standard Transformer block, where multi-head self-attention mechanisms allow the model to exchange contextual information across all patches. The output of this process is a refined sequence of embeddings E_{ViT} of shape $(B, N+1, D)$

Next, the output from the Inception module E_{cnn} is added to the corresponding patch embeddings from the Transformer block, excluding the classification token:

$$E_{\text{combined}} = E_{ViT}[:, 1:] + E_{\text{cnn}}$$

This combined sequence E_{combined} now enriched with both global context and fine-grained features and passed through an additional Transformer block to further refine and fuse the model's understanding:

$$E_{\text{refined}} = f_{\text{transformer}}(E_{\text{combined}})$$

(4) Final Normalization and Classification Head

The refined output E_{refined} is normalized using layer normalization and passed through the classification head. The classification token CLS which aggregates information from all patches, is extracted and processed by fully connected layers to produce the final class predictions.

The proposed model pipeline, shown in Figure 4, aims to combine the strengths of ViT and CNNs by using the Inception module to extract local fine-grained features of the satellite images. These features are then fused with the global context captured by the transformers, allowing the model to better understand both large-scale patterns and intricate details for the damaged buildings.

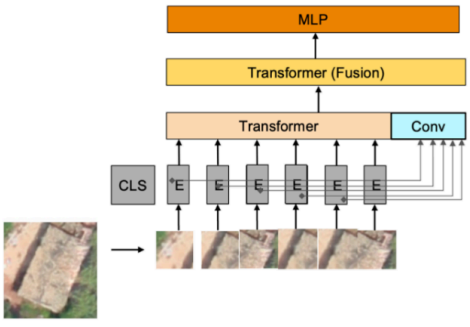

Figure 4. Proposed model pipeline

4. Results and discussion

The Vision Transformer b-16 model with pre-trained weights outperformed the compared CNN architectures across all metrics. Among the CNN models being compared, Inception-V3

DOI: 10.54254/2755-2721/2025.30586

demonstrated strongest performance, followed by EfficientNet-B0 and ResNeXT-D. Table 1 for comparing VIT and CNN models on Ludian Dataset illustrates the models' performance based on plotted numbers: The results underscore the efficacy of the Vision Transformer in modeling complex, high-resolution satellite images and suggest its superiority over traditional CNN architectures for this particular application.

Table 1. Comparison of ViT and CNN models on the Ludian dataset

Model Name	Accuracy	Precision	Recall	F1
ResNet50	0.7202	0.7137	0.7198	0.7164
Inception-V3	0.7334	0.7308	0.7265	0.7282
EfficientNet-B0	0.7319	0.7271	0.7329	0.7298
ViT-B16	0.7592	0.7675	0.7498	0.7576

The observed performance differences among the models can be attributed to their underlying designs. he Vision Transformer's use of self-attention mechanisms allows it to capture long-range dependencies and global context more effectively than CNNs, which rely on localized feature extraction through convolutional layers. This capability is particularly advantageous in the analysis of satellite imagery, where capturing spatial relationships over large areas is crucial. ViTs excel in their ability to process entire images holistically, preserving the contextual integrity that is often lost in traditional CNNs due to their limited receptive fields. Furthermore, the inherent flexibility of ViTs in handling various scales and rotations within the data makes them more adaptable to the diverse and complex patterns found in real-world disaster scenarios. This adaptability is particularly valuable when assessing damage in heterogeneous environments, such as those found in post-earthquake settings. In contrast, Inception-V3's relatively strong performance can be linked to its architecture, which processes multiple scales of information in parallel, allowing it to capture both fine and coarse details within the images. EfficientNet's competitive performance is due to its efficient scaling strategy, which balances network depth, width, and resolution, enabling it to model complex features without overfitting. Finally, the ResNet model, despite its modular design and increased cardinality, may have been less effective due to its emphasis on localized feature learning, which might not fully exploit the global patterns present in satellite imagery.

Table 2. Comparison of ViT and CNN models trained on Ludian dataset, tested with Yushu dataset

Model Name	Accuracy	Precision	Recall	F1
ResNet50	0.5510	0.4974	0.5015	0.4820
Inception-V3	0.5875	0.5392	0.5406	0.5356
EfficientNet-B0	0.5119	0.4615	0.4591	0.4397
ViT-B16	0.6307	0.5967	0.59	0.5763

When the models trained on the Ludian dataset were applied to the Yushu dataset for testing, the Vision Transformer (ViT-B16) continued to outperform the CNN-based models across all metrics. The pattern can be seen in Table 2, which varies the performance of ViT and CNN models trained on the Ludian Dataset, tested with the Yushu dataset. The ViT-B16 model achieved the highest scores across all metrics. Among the CNN models, Inception-V3 exhibited the best performance with an accuracy. ResNet50 and EfficientNet-B0 followed with lower performance, particularly in precision and recall.

DOI: 10.54254/2755-2721/2025.30586

When tested across different earthquake scenarios, the ViT model demonstrated a greater ability to generalize than the CNN models, which struggled more with the Yushu dataset. This limitation in CNNs likely stems from their reliance on localized feature extraction, which may not fully capture the complex variations in building structures and damage patterns across different earthquakes. Inception-V3's relatively better performance can be attributed to its multi-scale processing capability, but it still falls short of ViT's adaptability. Overall, the results suggest that ViTs may offer a more versatile and robust solution for post-earthquake damage assessment.

Table 3. Vision Transformer's individual performance on Yushu and Ludian datasets

Dataset	Accuracy	Precision	Recall	F1
Yushu	0.7514	0.7326	0.7258	0.7283
Ludian	0.7592	0.7675	0.7498	0.7576

As we can see, the data in Table 3, which indicates ViT's individual performance on Yushu and Ludian datasets, is assessed quite similarly when it goes through transfer learning, trained, and tested on the Yushu dataset. Despite the Yushu dataset being much smaller, the ViT model maintained strong performance, suggesting its effectiveness in scenarios with limited data availability. This demonstrates the model's robustness and highlights the importance of developing models that can perform well even when large amounts of new data are not accessible, as often encountered in sudden disaster situations.

Table 4. Comparison of ViT and proposed model on the Yushu dataset

Model	Accuracy	Precision	Recall	F1
ViT-b16	0.5781	0.5428	0.5458	0.5326
Proposed Model	0.5866	0.5600	0.5431	0.5450

According to Table 4, a comparison between the regular ViT model and the proposed model we created, the proposed model showed a stable improvement in accuracy, precision, and F1 compared to the baseline ViT model on the Yushu dataset; both were trained with no pre-trained weights. By integrating the Inception module into the ViT architecture, the model effectively overcomes the limitations of ViTs in local feature extraction. This hybrid approach not only captures global context but also fine-grained details of damaged structures, making it particularly valuable in scenarios where precise damage classification is critical, yet data is scarce.

Furthermore, the model's ability to achieve stable improvements even without pre-trained weights underscores its robustness. This capability is essential for applications in disaster-stricken regions where large, labeled datasets might not be readily available. The research provides a compelling direction for future work in disaster image classification models, particularly in exploring ways to enhance local feature extraction in vision transformers.

However, the study also indicates that while the proposed hybrid model improves performance, it still lags behind ViTs fine-tuned through transfer learning. This indicates that the vision transformer model requires substantial amounts of data for training.

5. Conclusion

This paper analyzed and improved the efficacy of ViTs for post-disaster building damage classification, particularly in underdeveloped areas where timely and accurate assessments are

DOI: 10.54254/2755-2721/2025.30586

critical. By leveraging transfer learning, ViTs demonstrated superior performance compared to traditional CNN models, such as ResNet50, Inception-V3, and EfficientNet-B0. ViTs have not provided a more effective assessment of rural buildings, even when trained on smaller datasets. It also exhibited more robust generalization across different earthquake scenarios.

A key contribution of this research is the proposed Inception-enhanced ViT architecture, which addresses the inherent limitations of ViTs in local feature extraction. By integrating a CNN-based Inception module, the model's ability to capture and process fine-grained details of building damage was significantly improved, resulting in enhanced classification accuracy and generalization. This hybrid approach represents a promising direction for developing more reliable and efficient damage assessment systems.

The findings underscore the potential of ViTs as a robust tool for building damage assessment in disaster-stricken areas, offering a more accurate, adaptable, and efficient alternative to traditional CNN-based methods. This improvement not only enhances the reliability of damage classification but also potentially accelerates the response and recovery processes, thereby mitigating the impact on affected communities. Furthermore, the integration of Inception with ViT offers a promising framework and direction for future research, providing a more comprehensive approach to combining local and global feature extraction in similar tasks. For future work, the proposed model could be modified and improved on larger datasets. Additionally, exploring advanced techniques for combining local and global features could further enhance model performance, such as XgBoost or concatenation. The insights gained from this research pave the way for broader applications of ViTs in disaster response and management, with the ultimate goal of better supporting those affected by natural disasters. By improving the accuracy and speed of damage assessments, these advancements could lead to more efficient allocation of resources, quicker mobilization of aid, and more informed decision-making, ultimately saving lives and reducing the socioeconomic impact on vulnerable communities.

References

[1] Gupta, Ritwik, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. 2019. “xBD: A Dataset for Assessing Building Damage from Satellite Imagery.” arXiv. http: //arxiv.org/abs/1911.09296.

[2] Dong, Laigen, and Jie Shan. 2013. “A Comprehensive Review of Earthquake-Induced Building Damage Detection with Remote Sensing Techniques.” ISPRS Journal of Photogrammetry and Remote Sensing 84 (October): 85–99. https: //doi.org/10.1016/j.isprsjprs.2013.06.011.

[3] Chen, M., X. Wang, A. Dou, and X. Wu. 2018. “The Extraction of Post-Earthquake Building Damage Informatiom based on Convolutional Neural Network.” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII–3 (April): 161–65. https: //doi.org/10.5194/isprs-archives-XLII-3-161-2018.

[4] Gong, Lixia, Qiang Li, and Jingfa Zhang. 2013. “Earthquake Building Damage Detection with Object-Oriented Change Detection.” In 2013 IEEE International Geoscience and Remote Sensing Symposium - IGARSS, 3674–77. Melbourne, Australia: IEEE. https: //doi.org/10.1109/IGARSS.2013.6723627.

[5] Wang, Yanping, Yanbin Wang, Yong Da, Xiaoyan Liu, Jiebo Li, and Jingyi Huang. 2011. “An Object-Oriented Method for Road Damage Detection from High Resolution Remote Sensing Images.” In 2011 19th International Conference on Geoinformatics, 1–5. Shanghai, China: IEEE. https: //doi.org/10.1109/GeoInformatics.2011.5981141.

[6] Ci, Tianyu, Zhen Liu, and Ying Wang. 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] Ma, Haojie, Yalan Liu, Yuhuan Ren, Dacheng Wang, Linjun Yu, and Jingxian Yu. 2020. “Improved CNN Classification Method for Groups of Buildings Damaged by Earthquake, Based on High Resolution Remote Sensing Images.” Remote Sensing 12 (2): 260. https: //doi.org/10.3390/rs12020260.

DOI: 10.54254/2755-2721/2025.30586

[8] Berezina, Polina, and Desheng Liu. 2022. “Hurricane Damage Assessment Using Coupled Convolutional Neural Networks: A Case Study of Hurricane Michael.” Geomatics, Natural Hazards and Risk 13 (1): 414–31. https: //doi.org/10.1080/19475705.2022.2030414.

[9] Lin, Qigen, Tianyu Ci, Leibin Wang, Sanjit Kumar Mondal, Huaxiang Yin, and Ying Wang. 2022. “Transfer Learning for Improving Seismic Building Damage Assessment.” Remote Sensing 14 (1): 201. https: //doi.org/10.3390/rs14010201.

[10] Khan, Asifullah, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. 2020. “A Survey of the Recent Architectures of Deep Convolutional Neural Networks.” Artificial Intelligence Review 53 (8): 5455–5516. https: //doi.org/10.1007/s10462-020-09825-6.

[11] Ji, Min, Lanfa Liu, Runlin Du, and Manfred F. Buchroithner. 2019. “A Comparative Study of Texture and Convolutional Neural Network Features for Detecting Collapsed Buildings After Earthquakes Using Pre- and Post-Event Satellite Imagery.” Remote Sensing 11 (10): 1202. https: //doi.org/10.3390/rs11101202.

[12] Qing, Yuanzhao, Dongping Ming, Qi Wen, Qihao Weng, Lu Xu, Yangyang Chen, Yi Zhang, and Beichen Zeng. 2022. “Operational Earthquake-Induced Building Damage Assessment Using CNN-Based Direct Remote Sensing Change Detection on Superpixel Level.” International Journal of Applied Earth Observation and Geoinformation 112 (August): 102899. https: //doi.org/10.1016/j.jag.2022.102899.

[13] Zheng, Zhuo, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. 2021. “Building Damage Assessment for Rapid Disaster Response with a Deep Object-Based Semantic Change Detection Framework: From Natural Disasters to Man-Made Disasters.” Remote Sensing of Environment 265 (November): 112636. https: //doi.org/10.1016/j.rse.2021.112636.

[14] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv. http: //arxiv.org/abs/2010.11929.

[15] Bazi, Yakoub, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. 2021. “Vision Transformers for Remote Sensing Image Classification.” Remote Sensing 13 (3): 516. https: //doi.org/10.3390/rs13030516.

[16] Kaur, Navjot, Cheng-Chun Lee, Ali Mostafavi, and Ali Mahdavi-Amiri. 2023. “Large-Scale Building Damage Assessment Using a Novel Hierarchical Transformer Architecture on Satellite Images.” arXiv. http: //arxiv.org/abs/2208.02205.

[17] Ramesh, Amrita, Sanjana K R Prasad, Siddhanth Srikanth, and Shikha Tripathi. 2023. “Hurricane Damage Detection Using Computer Vision.” In Proceedings of the 2023 5th International Conference on Image, Video and Signal Processing, 126–32. Singapore Singapore: ACM. https: //doi.org/10.1145/3591156.3591174.

[18] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. “Deep Residual Learning for Image Recognition.” arXiv. http: //arxiv.org/abs/1512.03385.

[19] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. “Rethinking the Inception Architecture for Computer Vision.” arXiv. https: //doi.org/10.48550/ARXIV.1512.00567.

[20] Tan, Mingxing, and Quoc V. Le. 2019. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” https: //doi.org/10.48550/ARXIV.1905.11946.