

CMPT 742 Visual Computing Lab I

Building Damage Extent Detection and Classification Using Satellite Imagery for Large Natural Disasters

Yuan Chang

Jiachen Huang

Maaheen Yasin

Submitted to:

Professor Ali Mahdavi-Amiri

Simon Fraser University

Burnaby, British Columbia

December 15, 2024

Introduction

Background

Natural disasters, such as hurricanes, earthquakes, and wildfires, cause significant damage to buildings and infrastructure, resulting in human casualties and economic losses. The ability to quickly and accurately assess the extent of damage is critical for effective emergency response and resource allocation. Traditional methods, including manual inspections and ground-based sensors, often suffer from limitations such as time consumption, limited coverage, and inconsistent results. Satellite imagery, when combined with deep learning models, offers an innovative solution by enabling large-scale and real-time damage assessment, even in inaccessible areas.

Objectives and goal

This project sought to address the pressing need for rapid and accurate damage assessment following natural disasters. By utilizing satellite imagery and deep learning techniques, the primary objectives included:

1. Identifying efficient methods for disaster damage assessment by automating the segmentation and classification of building damage.
2. Enabling real-time damage detection through a streamlined processing pipeline, from pre- and post-disaster satellite imagery to pixel-wise damage classification.
3. Providing actionable insights into the extent of disaster impact to support emergency response efforts, with models capable of achieving high F1 scores and accurate damage categorization, even in challenging conditions like cloud-obstructed images.

Literature Review

The first paper we mainly focused on is “*Large-Scale Building Damage Assessment Using A Novel Hierarchical Transformer Architecture On Satellite Images*”. (Kaura et al., 2023) This study introduces a novel methodology for automated disaster damage assessment using satellite imagery, addressing key limitations of traditional approaches. The proposed model, DAHiTrA, combines the strengths of hierarchical feature extraction from UNet with the global attention mechanisms of transformers. A central innovation is the "difference block," which encodes pre- and post-disaster images into a shared feature space to calculate temporal changes, ensuring a more focused and precise analysis of structural damage. Unlike conventional convolutional methods that rely on simple concatenation of temporal features, DAHiTrA's approach allows it to distinguish changes more effectively. Its hierarchical design, leveraging multi-resolution features, enables the model to balance detailed local predictions with a broad spatial context. Additionally, the use of focal and dice loss functions addresses the challenge of class imbalance, improving performance on underrepresented categories like "minor damage" and "destroyed."

The results are compelling. On the xBD dataset, DAHiTrA achieved an overall F1-score of 0.819, significantly outperforming existing models, particularly in the classification of minor damage. Furthermore, the model demonstrated adaptability to new datasets, such as Ida-BD, where fine-tuning a pre-trained version yielded an F1-score of 0.585. This showcases its potential for real-world scenarios, especially where labeled data is scarce. However, some limitations remain. The model struggles with heavily underrepresented damage classes and exhibits challenges in delineating building boundaries, particularly in high-resolution images where noise from environmental factors like clouds or vegetation is prevalent. These shortcomings suggest future research directions, including enhanced augmentation strategies, better loss functions for edge precision, and the incorporation of dynamic learning frameworks such as GANs or ensemble methods.

CNNs, U-Nets, and FCNs have been demonstrated as effective baselines for building damage assessment in satellite imagery through various studies. The use of CNNs, as detailed in Xu et al.'s work, focuses on extracting features from pre- and post-disaster images to identify building damage. (Xu et al., 2019) Their study employed a twin-tower subtract (TTS) CNN architecture, achieving an AUC of 0.8302 on the Haiti earthquake dataset, emphasizing CNNs' capacity to detect structural changes in disasters.

Similarly, U-Nets have been adapted for building segmentation and damage classification, as demonstrated by Deng and Wang. (Deng & Wang, 2022) Their two-stage U-Net framework combines segmentation and Siamese network-based damage classification, achieving F1 scores of 0.8741 for building localization and 0.7536 for damage classification on the xBD dataset. This study highlights U-Nets' strength in handling multi-scale features and integrating pre- and post-disaster imagery for damage analysis.

Additionally, FCNs have shown potential for building extraction in multisource imagery, as outlined in Ji et al.'s work. (Ji et al., 2019) The team proposed a new variant of FCN which they called SiU-Net. The study improved segmentation accuracy for buildings of various scales, particularly in large-scale cases. Their results demonstrated the scalability and adaptability of FCN-based methods in segmenting buildings, a key precursor for damage assessment applications.

These studies collectively establish CNNs, U-Nets, and FCNs as foundational approaches for automated building damage assessment in satellite imagery, highlighting their respective roles in segmentation, feature extraction, and scale handling for disaster response tasks.

Methodology

Project Architecture

The architecture of this project is designed to address the challenge of efficiently assessing the damage caused by different natural disasters using pre- and post-disaster satellite imagery. The objective of this architecture is to systematically process the acquitted data, train the models, and then evaluate their performance and have a comparison of which model is more accurate in detecting and assessing damages. This architecture comprises four core stages: data processing, model selection, training, and evaluation.

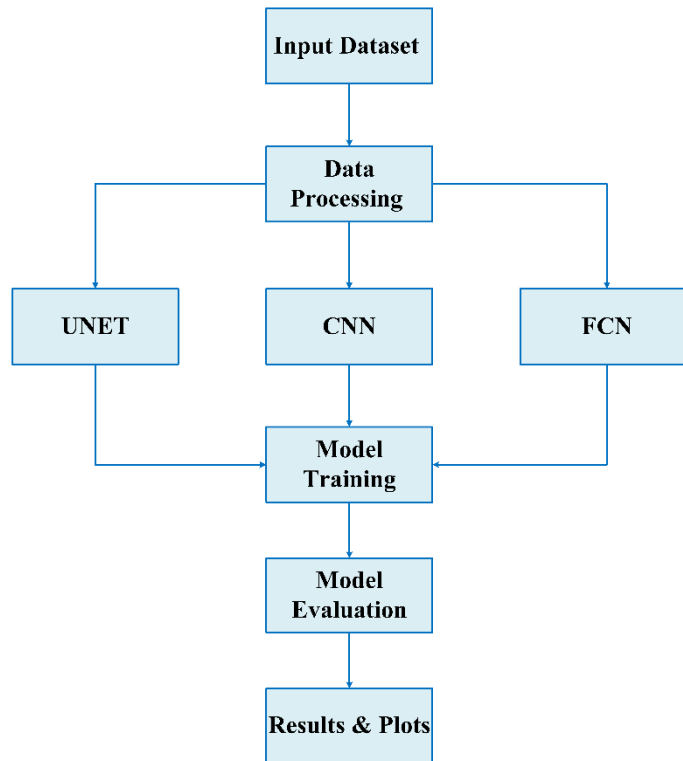


Figure 1 Overall Architecture

Dataset

The project architecture begins with the input dataset stage which focuses on selecting and preparing high-quality data for damage assessment by the models. While there are multiple datasets available, the xBD dataset has been primarily used for model training and evaluation. This dataset provides a large-scale and diverse collection of high resolution pre- and post-disaster satellite images, paired with corresponding damage annotations from various natural disasters such as earthquakes, floods, hurricanes, and wildfires. The annotations categorize the building damage into multiple severity levels, including no damage, minor damage, major damage, and destroyed. These labeled ground truths are essential for supervised image segmentation and classification.

The diversity of the xBD dataset makes it a strong foundation for the models, as it enables them to learn robust features for detecting and classifying damaged areas and generalize effectively across multiple disaster scenarios. Moving forward, to further enhance the models' detection capabilities on unseen data and ensure greater generalization, there are plans to incorporate the Ida dataset. The Ida dataset contains 87 before-and-after satellite images collected following Hurricane Ida in 2021, specifically in Louisiana, USA. By integrating this additional dataset, the models can be trained and tested on more recent and localized disaster data, improving their adaptability and performance for real-world applications.

Data Processing

The data processing stage is crucial for preparing the input data for efficient model training and evaluation. The xBD dataset contains annotations in JSON format and requires preliminarily preprocessing workflow to create accurate ground truth masks. Meanwhile, the Ida dataset already contains ground truth masks and would not require this preprocessing step.

For the xBD dataset, the process begins by extracting and parsing the data from the JSON files, which contain detailed coordinate information about the pre- and post-disaster building outlines, along with the associated damage severity levels. These coordinates are then used to generate polygons that form precise building footprints for each image, suitable for segmentation tasks. The generated masks are subsequently saved in a designated folder called "masks," which the model will access as the ground truth during training.



Figure 2 Data Processing Steps

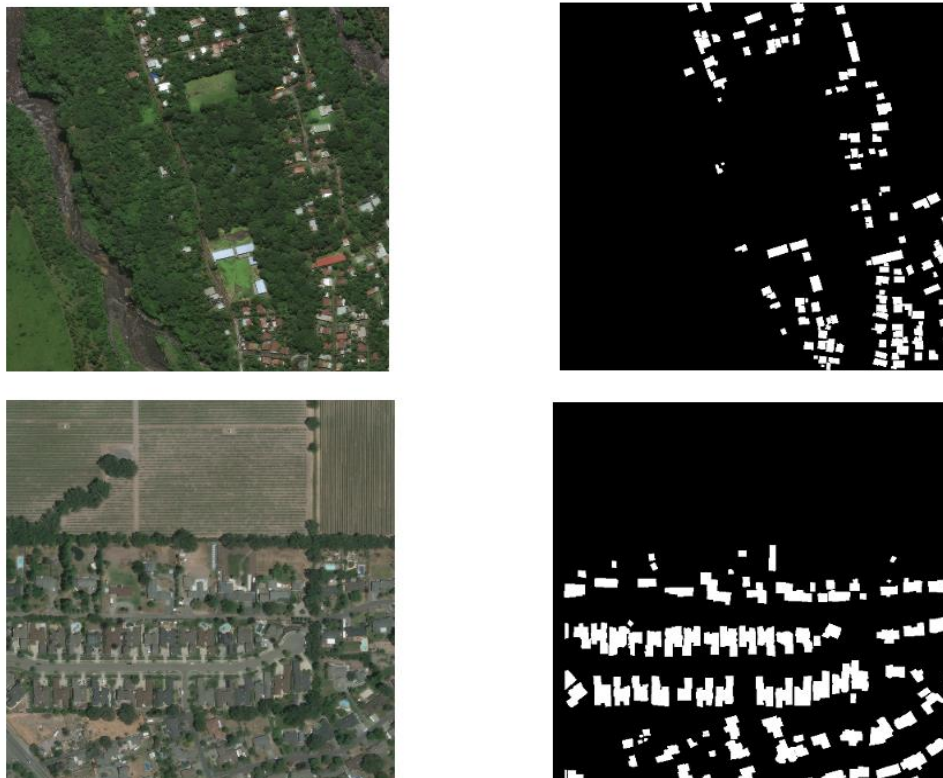


Figure 3 Ground Truth Masks

Above are two ground truth masks generated for the satellite imagery on the left, highlighting the outlines of the buildings in the image.

Moreover, as a part of the data preprocessing, all images are resized to a uniform dimension to ensure consistency across the dataset before being used for model training and testing. Additionally, resizing and normalization transformation are applied to enhance model performance as resizing standardizes the input size, while normalization adjusts the pixel values, ensuring that the data is on a consistent scale, which helps the model learn more effectively and reduces potential biases during training.

Models

The objective of this project is to evaluate the performance and accuracy of five different deep learning models for damage assessment. The models include Convolutional Neural Networks (CNN), Fully Convolutional Networks (FCN), U-Net, Recurrent Neural Networks (RNN), and You Only Look Once (YOLO). In the initial phase, the CNN, FCN, and U-Net models were implemented, trained, and tested using the xBD dataset to assess their effectiveness in identifying and segmenting damage in building structures.

Convolutional Neural Network (CNN)

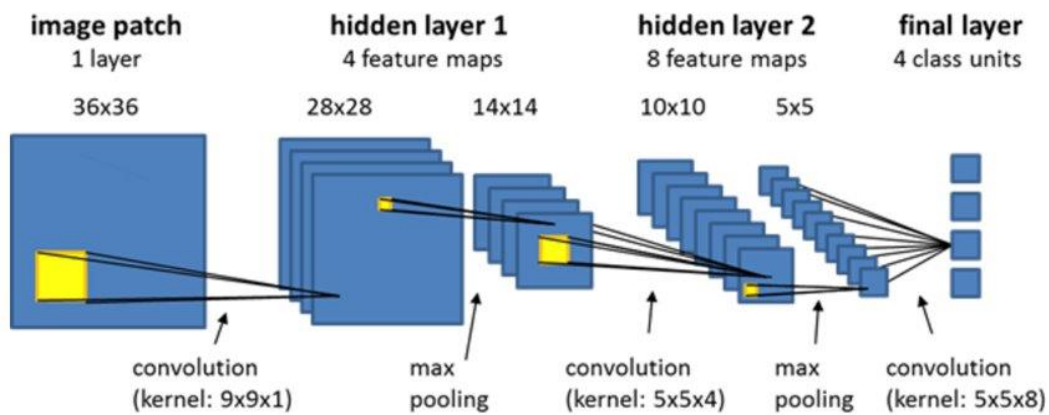


Figure 3 CNN Model Architecture

The Convolutional Neural Network (CNN) is a deep learning architecture that excels in image classification due to its ability to learn spatial hierarchies in images through convolutional and pooling layers. This model is particularly effective for feature extraction, as it learns to identify patterns related to building damage in pre- and post-disaster images. In this project, the CNN architecture was trained using the xBD dataset to classify images based on damage severity. The CNN model uses multiple convolutional layers, pooling layers, and fully connected layers to progressively extract high-level features and classify images into distinct damage categories.

Fully Convolutional Network (FCN)

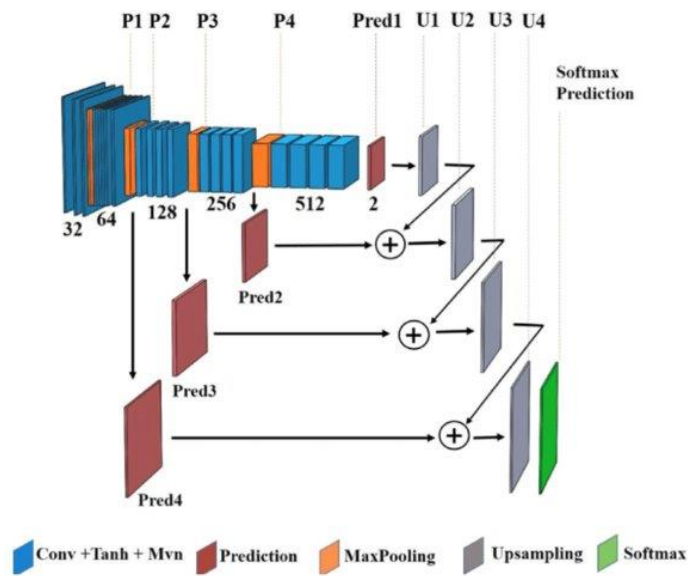


Figure 4 FCN Model Architecture

The Fully Convolutional Network (FCN) is a variation of the Convolutional Neural Network (CNN) that replaces fully connected layers with convolutional layers, enabling it to perform pixel-wise predictions for image segmentation. This makes FCNs particularly suited for generating semantic segmentation maps, where each pixel is classified as either belonging to a damaged or undamaged region. Such pixel-level classification is crucial in post-disaster scenarios, where precise localization of damage is needed for detailed analysis and decision-making. The FCN model was trained on the xBD dataset, leveraging its ability to learn spatial relationships across the image and produce detailed damage maps, highlighting the extent and specific locations of the damage.

Siamese U-NET

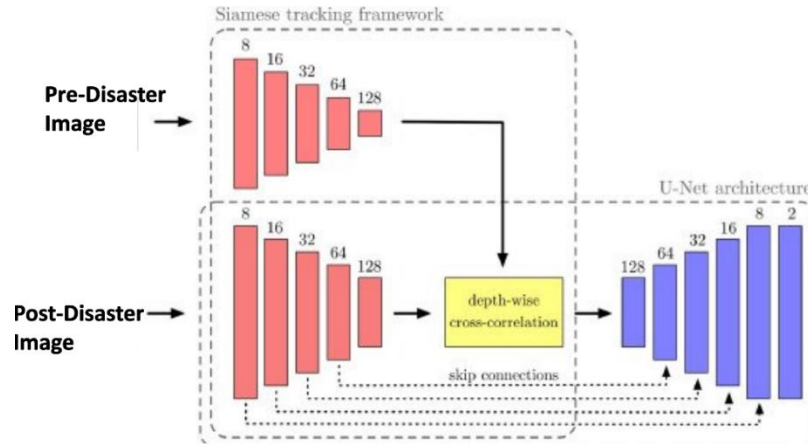


Figure 5 Siamese U-NET Model Architecture

The Siamese U-Net model is designed to handle image comparison tasks, which makes it particularly useful for damage assessment using pre- and post-disaster images. This model is an extension of the traditional U-Net architecture, incorporating a Siamese network structure in which two identical U-Net models share weights and are trained simultaneously to assess the changes between the two images. By comparing the pre- and post-disaster images in this manner, the model can focus on identifying the localized changes in the images and accurately assessing damage in post-disaster scenarios. The Siamese U-Net is trained to output a damage map that highlights the exact regions of damage in the post-disaster image.

Model Training

The training phase involves teaching the models to identify and assess damage severity using the prepared dataset. The main objective is to minimize the difference between the predicted outputs and the actual ground truths. To achieve this, the cross-entropy loss function is used to measure the training loss while the Adam optimizer is used to adjust the model parameters and reduce the loss. For each model, hyperparameters such as learning rate, batch size, and the number of epochs are carefully configured before training begins, ensuring optimal learning.

Model Evaluation

After training, the models are evaluated on the test dataset to assess their ability to generalize to unseen data. The evaluation process involves comparing the model's predictions against the ground truth using standard performance metrics, including the F1-score and Intersection over Union (IoU). The F1-score evaluates the model's precision and recall, providing a balanced measure of accuracy while the IoU assesses the overlap between predicted and actual segmented regions.

Results & Plots

The results of the damage assessment are visualized to demonstrate the effectiveness of the trained models in segmenting and classifying post-disaster images. The pre- and post-disaster images along with the segmented and classified output are visualized. In addition, the training and validation loss, as well as the F1-score and IoU are plotted to track model performance throughout the training and validation phases.

Project Management

The roadmap for this project is divided into three phases. The first phase spanned the initial two months and focused on conducting a detailed literature review and implementation of three baseline models. This foundational work set the stage for subsequent phases dedicated to expanding model capabilities, refining their performance, and integrating advanced techniques to improve classification accuracy. During the first phase, the team reviewed the existing literature to understand the current methodologies and techniques used in disaster damage assessment, particularly with deep learning models. According to the literature review, we designed the pipeline to accomplish the tasks and then we mainly focused on implementing the three baseline models, CNN, FCN, and U-Net.

From the second semester, we will start to expand the model capabilities. To achieve a comprehensive assessment of disaster impacts, the model will be extended to classify additional categories such as roads, trees, and land use. We will integrate GANs to address challenges posed by cloud obstructions in satellite imagery. GANs will be trained to remove clouds while preserving critical image details, which could enhance the quality of input data for downstream analysis.

Through the final phase, we will focus on optimizing the model performance. In addition, we will try to integrate meta-attributes such as elevation maps or LiDAR data, to improve classification outcomes. These attributes provided more detailed contextual information, enabling the models to differentiate more effectively between various types of damage and environmental features.

The following is a brief-summarized roadmap of our project.

Phase I (Month 1 – 2): Initial Research and Model Design

Step 1: Literature Review

- Conduct a comprehensive review of existing literature related to image segmentation, deep learning models (UNet, CNNs, etc.), and disaster assessment using satellite images.
- Focus on techniques for classifying damaged buildings and other affected areas using pre- and post-disaster satellite images.

Step 2: Model Design

- Design the architecture of the model, initially focusing on building segmentation.
- Choose deep learning models, such as UNet or CNN-based architectures, to implement for classification of disaster impacts.

Step 3: Training the Model with Pre- and Post-Disaster Satellite Images

- Collect and preprocess the pre- and post-disaster satellite images for the building classification task.
- Train the model to classify building structures and assess damage based on segmented areas.
- Perform initial testing to ensure the model can accurately distinguish different classes (e.g., damaged, undamaged buildings).

Phase II (Month 3 – 4): Expanding Model Capabilities and Enhancements

Step 4: Classify Other Areas for Disaster Assessment

- Extend the model to classify non-building areas such as forests, land, and roads, to assess the wider environmental impact of the disaster.
- Incorporate these classifications into the damage assessment framework.

Step 5: Apply GANs for Cloud Removal

- Implement Generative Adversarial Networks (GANs) to remove clouds from satellite images to enhance the visibility of affected regions.
- Fine-tune the GAN model for optimal cloud removal performance without losing critical image details needed for disaster assessment.

Phase III (Month 5 – 6): Model Optimization and Final Adjustments

Step 6: Compare Different Model Performances

- Evaluate the performance of different deep learning models based on accuracy, recall, and processing time.
- Use test datasets to validate model robustness, focusing on how well each model handles varying damage extension.

Step 7: Add Other Meta Attributes

- Integrate additional meta-attributes like elevation maps, land cover, or LiDAR data to improve the model's prediction accuracy and enhance its understanding of terrain variations.
- Fine-tune the model to ensure the additional data sources improve classification results without significantly increasing computation time.

Through the project, each team member will take responsibility for one type of baseline models to ensure the tasks can be completed efficiently. The table below shows the work allocation.

Group Member	Maaheen Yasin	Jiachen Huang	Yuan Chang
Model	U-Net	CNN	FCN

At the end of the first term, we implemented the selected baseline models. But the models are not completed yet, we still need to work on these models in the next term. In addition, we will keep following the roadmap to enhance the model performance and integrate the models with GAN and meta-attributes.

Preliminary Results

CNN

Design: The CNN model is based on the ResNet architecture, utilizing pre-trained weights from ImageNet to accelerate convergence. The model incorporates residual connections to overcome the vanishing gradient problem and employs a customized classification head designed for the specific damage categories: No Damage, Minor Damage, Major Damage, and Destroyed. These adjustments ensure robust performance across a variety of scenarios.

Parameters: Training was conducted with a batch size of 32, a learning rate of 0.001, and for 10 epochs. The model uses Cross Entropy Loss to evaluate classification performance and the Adam optimizer to adjust weights effectively.



Figure 6 Train and Validation Loss of CNN model

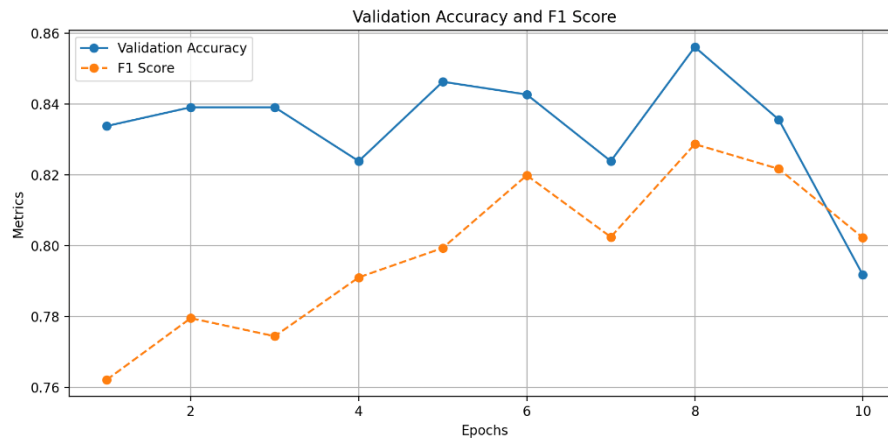


Figure 7 Validation Accuracy and F1 Score of CNN model

Results: The CNN demonstrated strong results, achieving an F1 Score of 0.82 and an IoU of 0.75. These metrics highlight its ability to accurately classify damage levels.

Visual results from the final presentation are included in Figure 7 and Figure 8. Figure 7 depicts the training and validation loss curves across 10 epochs. The training loss curve shows a steady decline, reflecting consistent learning and optimization. The validation loss curve exhibits minor fluctuations but stabilizes towards the latter epochs, confirming the model's ability to generalize effectively to unseen data. Both curves emphasize the model's robust performance and highlight its reliability in disaster damage classification tasks. These figures underscore the model's capacity to handle complex disaster scenarios effectively and demonstrate consistent performance across training and validation phases.

Visual outputs from the final presentation slides showcase the model's segmentation masks, where clear delineations between damage levels are observed. Additionally, the training and validation loss trends, as seen in Figure 7, validate the effectiveness of the

learning process. The curves demonstrate the model's capacity to learn complex patterns while maintaining robustness in validation, ensuring high reliability for real-world applications. These results underscore the model's applicability to real-world disaster scenarios.

FCN

The Fully Convolutional Network architecture was utilized for building segmentation tasks. The model consists of an encoder-decoder structure designed to efficiently extract features and reconstruct segmentation maps from high-resolution input images. The input dimensions are $256 \times 256 \times 3$, corresponding to RGB satellite images. The encoder comprises sequential convolutional layers, each followed by ReLU activation and max pooling operations. Specifically, the first convolutional layer (Layer1) includes 64 filters with dimensions $256 \times 256 \times 64$, followed by a max pooling layer that reduces spatial dimensions to $128 \times 128 \times 64$. The second convolutional layer (Layer2) doubles the filter count to 128 while retaining a similar design, and its corresponding max pooling layer further reduces the dimensions to $64 \times 64 \times 128$. This pattern continues with Layer3, which increases the filter count to 256, followed by max pooling to $32 \times 32 \times 256$.

The decoder begins with Layer4, which applies a convolution operation with 512 filters and dimensions $32 \times 32 \times 512$, followed by a series of transposed convolutional layers. Layer5 applies a transposed convolution, reducing the filter count to 256 while upsampling the spatial dimensions to $64 \times 64 \times 256$. Layer6 performs another transposed convolution to further reduce the filter count to 128 and upsample the dimensions to $128 \times 128 \times 128$. Similarly, Layer7 applies transposed convolution to reduce the filter count to 64 and upsample the spatial dimensions to $256 \times 256 \times 64$. Finally, Layer8 employs a single convolutional layer to produce the final segmentation output with dimensions $256 \times 256 \times 2$, corresponding to two segmentation classes: buildings and background.

The model was trained on the xBD dataset, which includes annotated satellite images before and after various disasters. The model was trained using the Adam optimizer, with a learning rate of 1×10^{-6} . The training process was conducted over 15 epochs with a batch size of 16, ensuring stable convergence and effective learning. During training, data augmentation techniques such as flipping, scaling, and rotation were applied to enhance the variability of training data, improving the model's ability to generalize across diverse building layouts and image conditions. Cross-Entropy Loss was used as the loss function because it effectively handles pixel-wise classification tasks by penalizing incorrect predictions and providing well-behaved gradients for stable optimization.

The performance of the FCN model was evaluated using metrics such as F1 score, IoU, and loss. As shown in Figure 9, the training and validation F1 scores increased consistently over the epochs, diagram we can see that the training and validation F1 scores steadily improved during the training process, and the validation F1 score reached 0.968 by the final epoch. In Figure 10, the IoU metric displayed steady growth throughout training, achieving a final validation IoU of 0.947, indicating good overlap between predicted and ground truth segmentation maps. The

training and validation loss curves in Figure 11 demonstrate a sharp decline in the early epochs, stabilizing towards the end of training, with a final validation loss of 0.031.

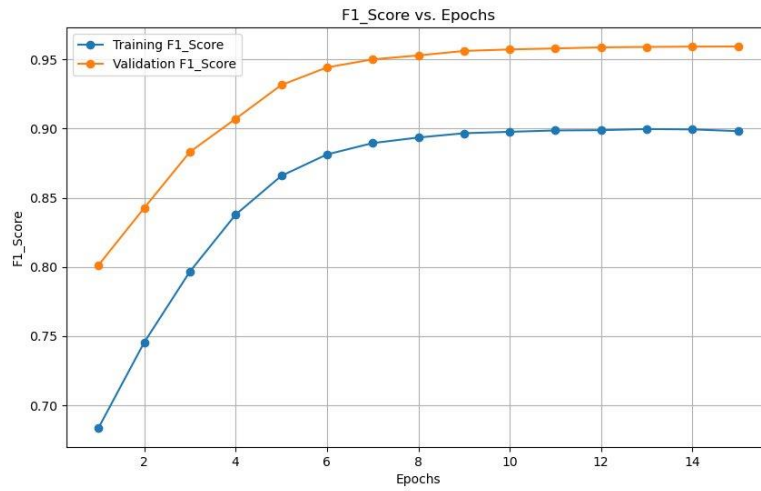


Figure 8 Training and Validation F1 Scores of FCN model

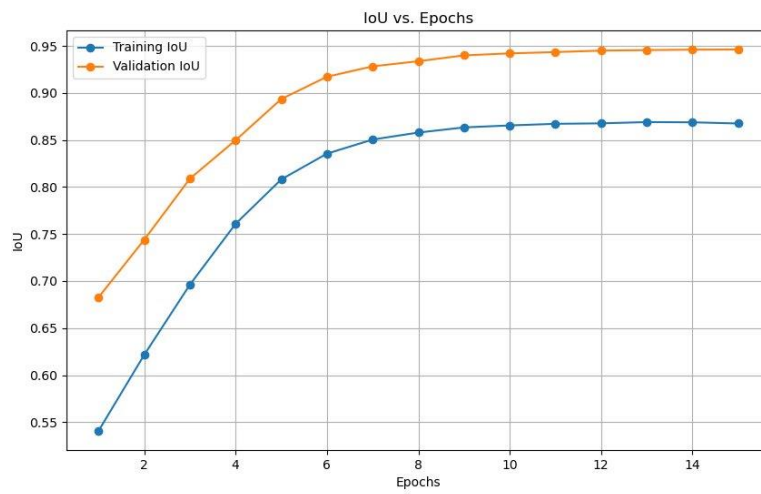


Figure 9 Training and Validation IOU of FCN model

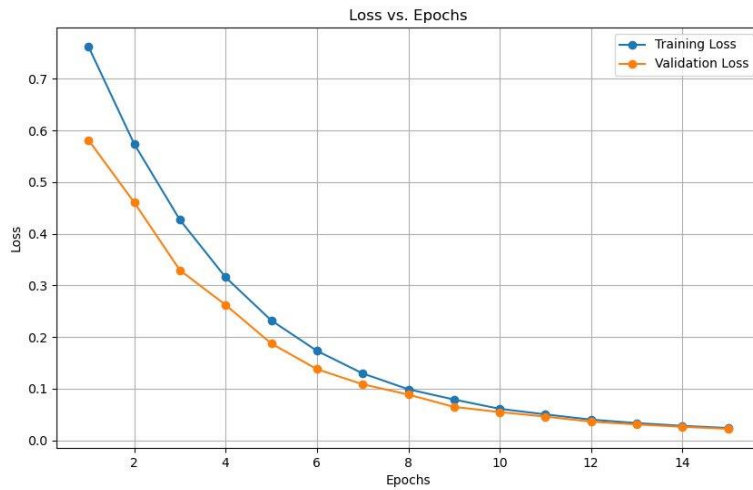


Figure 10 Training and Validation Loss of FCN model

The results highlight the model's ability to achieve precise segmentation, as evidenced by the high F1 score and IoU values. The convergence patterns in the loss curves further validate the model's robustness and its capability to effectively distinguish buildings from the background. These metrics underscore the effectiveness of the FCN architecture for building segmentation tasks, leveraging its encoder-decoder design and well-optimized hyperparameters to deliver high-performance results.

Siamese U-NET

Design:

The Siamese UNet architecture is specially designed to distinguish features by comparing two images. This architecture combines the UNet's capability for feature extraction and semantic segmentation with a Siamese framework for capturing changes across the pre- and post-disaster images. The architecture contains two main components:

1. Twin UNets for processing pre- and post-disaster images
2. Siamese difference module for change detection

In the first component, each pair of satellite images is passed through identical UNet structures to extract high-level features. The encoder-decoder structure of UNet ensures effective spatial feature learning while skip connections preserve details across multiple levels. The encoder consists of four stages of convolutional blocks followed by max-pooling layers, progressively reducing the spatial dimensions while capturing high level features. The bottleneck layer bridges the encoder and decoder while also acting as the deepest feature representation layer. The decoder systematically upsamples features to reconstruct the spatial resolution, concatenating them with corresponding encoder features via skip connections to retain the fine details.

In the second component, after processing the pre and post disaster images through their respective UNets, the differences of the features are computed between corresponding bottleneck and encoder outputs. This captures critical changes that reflect damage caused by the natural disaster. The computed difference feature maps are then passed through a classifier network that generates a damage assessment map with multiple classes: 0 – background, 1 - no damage, 2- minor damage, 3 – major damage, and 4 - destroyed.

Parameters

The training of the Siamese UNet model is conducted with the following hyperparameters:

- **Learning Rate:** 0.001
- **Batch Size:** 16
- **Total Epochs:** 15

The model utilizes the cross-entropy loss function to calculate the training loss and in order to minimize this loss during the training phase and optimize the model, the Adam optimizer is used.

Results

To evaluate the accuracy and generalization of the model on unseen data, the following standard performance metrics were calculated and plotted, i.e. F1-score, and Intersection over Union (IoU). Moreover, the training and validation loss is also calculated and plotted to have further insights into the performance of the model.

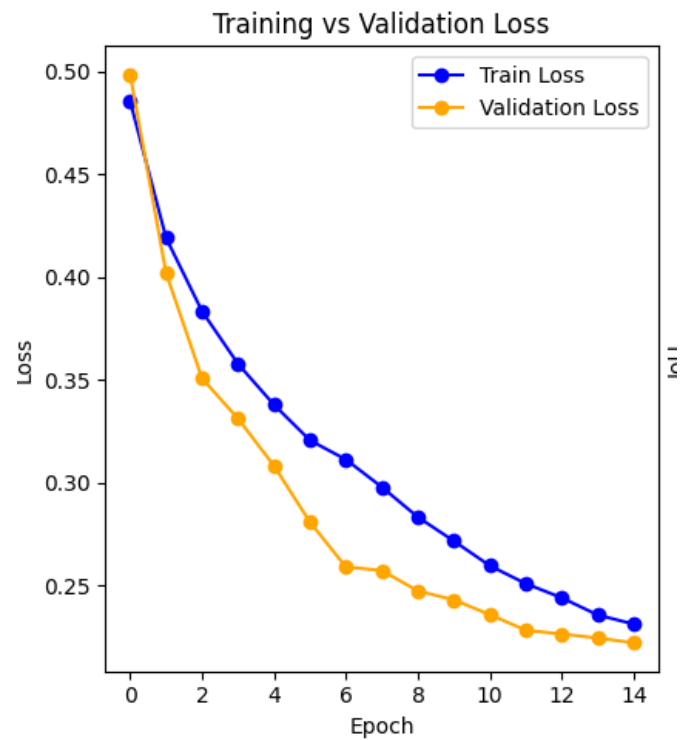


Figure 11 Training and Validation Loss

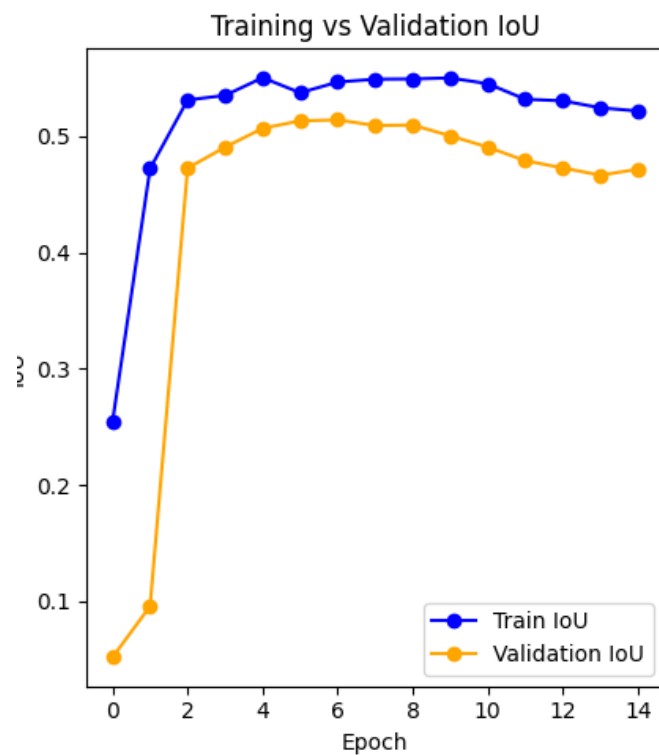


Figure 12 Training and Validation IoU

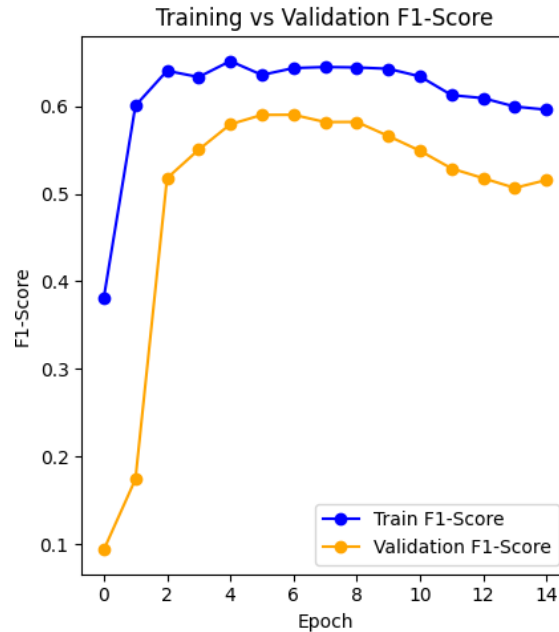


Figure 13 Training and Validation F1 Score

Figure 12 is the plot of the training and validation loss calculated for each epoch. At the beginning of the training phase, the training loss is relatively high starting from approximately 0.49 and gradually reduces after each epoch to 0.25. Similarly, the validation loss is relatively high at the beginning of the validation phase starting from 0.50 and gradually reduces to 0.24 after the 15 epoch.

Moreover, the training and validation IoU is plotted as illustrated in Figure 13. The training IoU started from 0.25 and reached 0.50 which suggests that the model needs fine tuning to achieve better accuracy. The validation IoU started from 0.02 and reached 0.49 which is less than the training IoU which implies that the model has overfitted to the training data and steps should be taken to overcome this.

Similarly, the training and validation F1-Score is plotted in Figure 15 with the training F1-score starting from 0.38 and reaching till 0.59 while the validation F1-Score starts from 0.02 and reaches till 0.50 which also implies that the model has overfitted and needs improvement.

Comparison with DAHiTra

Model	F1-Score	IoU
FCN	0.96	0.94
CNN	0.80	0.75
UNet	0.60	0.51
DAHiTra	0.796	0.872

The table compares the performance of four models—FCN, CNN, UNet, and DAHiTra—using F1-Score and IoU (Intersection over Union) as evaluation metrics. Among the models, FCN achieves F1-Score of 0.96 and an IoU of 0.94. The CNN model follows with an F1-Score of 0.80 and an IoU of 0.75.

Conclusion

Damage assessment in regions affected by natural disasters is a critical task that influences the allocation of resources and the reconstruction process. Moreover, by understanding the extent and severity of the damage, the decision makers can estimate budgets for reconstruction and can prioritize recovery efforts. Traditional methods are often time-consuming and labor intensive and can be overcome with the help of deep learning models by performing the damage assessment using the satellite imagery of the affected areas.

This project focuses on implementing and evaluating various deep learning models for the damage assessment using the xBD dataset. The initial phase of the project has been completed in which CNN, FCN, and Siamese UNET models have been trained to segment and classify building damage based on pre- and post-disaster images. The training phase involved optimizing the models with the Adam optimizer and cross-entropy loss, leading to significant improvements in model accuracy to detect damage across varying severity levels: no damage, minor damage, major damage, and destroyed. The evaluation phase was conducted using performance metrics, including F1-score and Intersection over Union (IoU), providing valuable insights into the models' ability to generalize and accurately assess damage.

Moving forward, future work will focus on enhancing the performance of the currently implemented models while also incorporating the Ida dataset to evaluate how well the models perform on unseen data. Additionally, the remaining two models — RNN and YOLO — would be implemented for damage assessment and a comprehensive comparison of all five models based on multiple performance metrics would be conducted to determine the most effective architecture for damage assessment.

References

- Deng, L., & Wang, Y. (2022). Post-disaster building damage assessment based on improved U-Net. *Scientific Reports*, 12(1), Article 15862. <https://doi.org/10.1038/s41598-022-20114-w>
- Ji, S., Wei, S., & Lu, M. (2019). Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1), 574–586. <https://doi.org/10.1109/TGRS.2018.2858817>
- Kaura, N., Lee, C.-C., Mostafavi, A., & Mahdavi-Amiri, A. (2023). Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. *arXiv*. <https://arxiv.org/abs/2208.02205>
- Xu, J. Z., Lu, W., Li, Z., Khaitan, P., & Zaytseva, V. (2019). Building damage detection in satellite imagery using convolutional neural networks. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. arXiv. <https://arxiv.org/abs/1910.06444>