Leveraging Pretrained Geospatial Foundation Models for Post-diasterous Building Damage Assessment

Shufan Li
Department of Computer Science
University of California Los Angeles
Los Angeles, CA 90095
jacklishufan@cs.ucla.edu

Chun-Yao Chang
Department of Computer Science
University of California Los Angeles
Los Angeles, CA 90095
cyc1028@g.ucla.edu

Rishab Dalai
Department of Computer Science
University of California Los Angeles
Los Angeles, CA 90095
dalair@g.ucla.edu

Abstract

As climate-induced natural disasters intensify, precise damage estimation is crucial for effective response and recovery. Existing solutions, mainly lightweight and domain-specific, face limitations in addressing diverse forms of damage. This study leverages ClimaX, a geospatial foundation model pretrained on weather data, and applies it to post-disaster building damage assessment using satellite imagery. We made several important architecture changes to address the domain shift from weather data to satellite images. In particular, we increased the patch size of ClimaX to handle high-resolution images and introduced an iterative upsampling process to refine the output of ClimaX. Results reveal that our modified version of ClimaX outperforms state-of-the-art methods by (+0.9 F1) on the xView2 building damage classification benchmark under comparable settings. This study highlights the potential of geospatial foundation models in accurate natural disaster damage assessment, offering a valuable contribution to disaster management strategies. The code is available at this link. A Colab notebook is available at this link

1 Introduction

With the increasing frequency and intensity of natural disasters due to climate change, it is critical for local governments and related stakeholders to accurately estimate the extent of damage caused by natural disasters. Inaccurate estimates can have far-reaching consequences, affecting resource allocation for disaster response and recovery efforts. More specifically, overestimating damage could lead to unnecessary expenses, while underestimating it may result in insufficient resources, delayed responses, and prolonged recovery periods. On the other hand, accurate predictions can help governments, emergency responders, insurance agencies, and disaster management organizations develop more effective mitigation strategies, allocate resources efficiently, and enhance disaster resilience.

While there have been a few existing ML-based solutions for this task, they have several key limitations. Specifically, most existing solutions such as FloodNet [16] use lightweight models that are domain-specific (for example, predicting flood-related damage). This is a major limitation since damage caused by natural disasters can manifest in multiple forms (a hurricane can cause damage

Preprint. Under review.

through floods, high wind speeds, etc.). Furthermore, these lightweight models do not benefit greatly from the wealth of satellite data that is currently available for training.

With the recent advent of geospatial foundation models, particularly in the field of climate prediction, there is an opportunity to fine-tune these models for damage assessment. Cutting-edge geospatial models can provide more resilient and accurate information regarding natural disasters, which is essential for estimating damage in a region. As a result, using geospatial foundation models as a baseline for damage and cost estimation models is vital for a data-driven approach to disaster management.

To address this need, we propose to fine-tune ClimaX [15], a geospatial foundation model, for damage assessment using the xBD [9] dataset of xView2 Challenge. To address the discrepancy in input resolution and variables between the satellite images in the xBD dataset and weather data, we implemented multiple architectural changes. In particular, features such as temperature, humidity, and other climate-related features are removed, and additional input layers are incorporated for the R, G, and B channels of satellite images. In addition, we increased the patch size for memory efficiency and introduced iterative upsampling to progressively refine the low-resolution ClimaX output to a high-resolution damage map. Additionally, high-resolution convolution features are concatenated during the upsampling process.

We compared our model against the first-place solution first-place winner of the xView2 AI Challenge. Results show that ClimaX outperforms the first-place model under comparable settings. In particular, it achieves a gain of (+0.1) on localization F1 and (+0.9) on classification F1.

2 Related Work

2.1 Building Damage Assessment

Automatic damage assessment has drawn wide interest from the machine learning community. Gupta et al. [9] first proposed xBD, a large-scale dataset for building segmentation and damage assessment. They also established a deep learning baseline using the ResNet-50 [10] model. Rahnemoonfar et al. [16] proposed FloodNet, a dataset that focused specifically on post-flood scene understanding. Cheng et al. [1] improved the CNN baseline using a stacked CNN architecture. MV-CNN [11] further improved CNN baselines by combining information from different views of a damaged building through 3D aggregation. ChangeOS [20] proposed an object-based semantic change detection framework as an alternative to CNN baselines. It combined an object localization network and a deep damage classification network to achieve optimal performance in both tasks. Wu et al. [19] proposed a Siamese architecture consisting of U-Net [18] with attention blocks. It outperformed previous works in the xBD benchmark. However, all of these works use relatively lightweight networks and did not explore some of the state-of-the-art architectures in computer vision, such as ViT [6] and Swin-Transformer [12].

2.2 Geospatial Foundation Models

The remote sensing community has shown strong interest in large-scale pretraining on satellite imagery. Neumann et al. [14] first provided a systematic review of visual representation learning methods when applied in the remote sensing field. Gao et al. [8] discovered the effectiveness of MAE pretraining in remote sensing. SatMAE [2] extended MAE to large-scale datasets of multispectral satellite images and established new state-of-the-art results in various downstream tasks, such as aerial image classification and building segmentation. ClimaX [15] was another geospatial foundation model that focused specifically on climate and weather data, and it operated at a much larger scale than other remote-sensing-oriented methods. Many of these works utilized Transformer-based architectures and outperformed previous convolution benchmarks. However, their work largely focused on static image classification and segmentation tasks and was not directly applicable to building damage assessment, which required aggregating information from pre-damage and post-damage images.

3 Methods

The current damage assessment models employed in the aftermath of natural disasters are domain-specific and limited in complexity, thus impeding the accuracy of damage assessments. To address this issue, our objective is to refine a geospatial foundation model to enhance the precision of damage estimation resulting from natural disasters.

3.1 Evaluation Metric

The problem of damage assessment can be formalized as two sub-tasks: building localization and damage classification. Given an input aerial image X, the localization task demands a binary mask output X_{seg} indicating the binary pixel-level classification of "background" and "building". Given a pair of pre-and-post disaster aerial images X_1, X_2 and predicted X_{seg}, the classification task demands a class label amongst "no damage", "minor damage", "major damage" and "destroyed". The F1 score is used to evaluate the performance on each of these tasks. For classification, class-specific F1 is also employed. The F1-score is calculated using precision (P) and recall (R) with Equation 1.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

3.2 ClimaX

In order to adapt Climax to damage classification task, two key architectural changes were implemented.

In the original Climax architecture, variable embedding layers were employed to process weather variables such as wind speed and Geopotential from the ERA5 dataset. However, these layers were not suitable for handling satellite image data. To overcome this, we removed the variable embedding layers related to the weather variables and introduced three new variables representing the Red, Green, and Blue (RGB) channels from the satellite images. These new variables were integrated into the network architecture and shared for both pre and post-disaster images, ensuring consistency and efficiency in the model's processing.

Another challenge is the resolution disparity between the ERA5 weather data and the satellite images in Xview dataset. ERA5 data is available at a coarser resolution of 0.5 degrees, while our satellite images are captured at a much finer resolution of 0.8km. This amounts to approximately 0.000005 degree resolution. In the original Climax model, a smaller patch size of 2 was utilized for feature extraction. However, to effectively process the high-resolution satellite images, we increased the patch size to 16 to reduce memory footprints. This change effectively reduced the number of patches needed in an image by a factor of 64 and makes training of satellite images computationally feasible.

3.3 Iterative Upsampling

Since Climax generates embeddings for each patch, the raw output from ClimaX is reshaped to produce a 4D tensor with the channel dimension representing the embedding. More concretely, the shape of the output is transformed from $N \times (H * W) \times D$ to $N \times D \times H \times W$, where N is batch size, D is embed depth, and H and W are height and width, respectively. This approach of reshaping the output of ClimaX into a 4D tensor from a 3D tensor is performed to maintain spatial information between nearby patches.

The resulting representation is characterized by low resolution and high channel depth. Thus, the reshaped output needs to be upsampled to increase the spatial resolution while reducing the channel dimension. To achieve this, transposed convolution, a common approach for upsampling 4D tensors, is utilized. By using transposed convolution over the reshaped output, the resulting upsample benefits from the spatial information preserved by the 4D representation. To smoothen upsampling, this process is performed over multiple passes instead of at once. In particular, each upsample block upsamples the feature map by a factor of two. Hence, a total of four upsample blocks is adopted to achieve a total of 16 times upsample, bringing the final spatial resolution in line with the expected output.

Regarding CNNs, recent research in applying transformer-like blocks (named ConvNeXt blocks) between downsampling steps allows CNNs to perform competitively with state-of-the-art vision transformers [13]. Since downsampling is similar to upsampling in the sense that spatial resolution changes, ConvNeXt blocks, as shown in 1, are applied right after each upsampling step. The ConvNeXt blocks process the upsampled output through several convolutions. This processed output

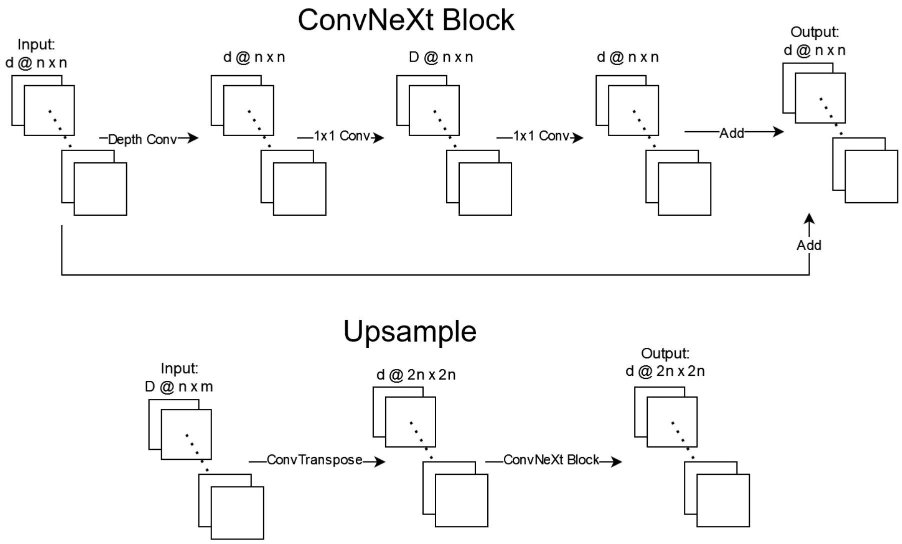


Figure 1: Architecture of upsampling step.

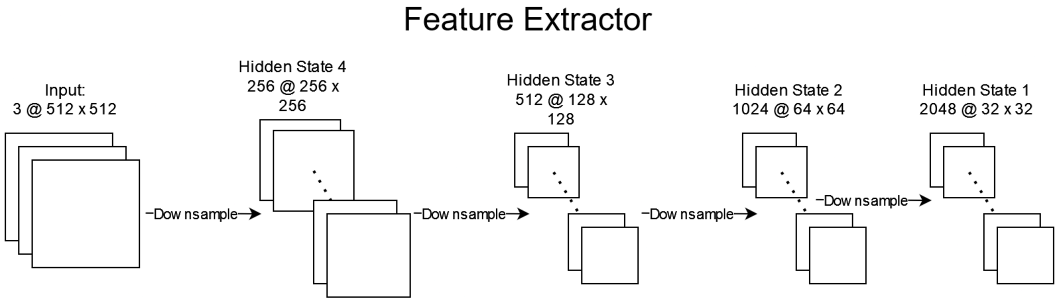


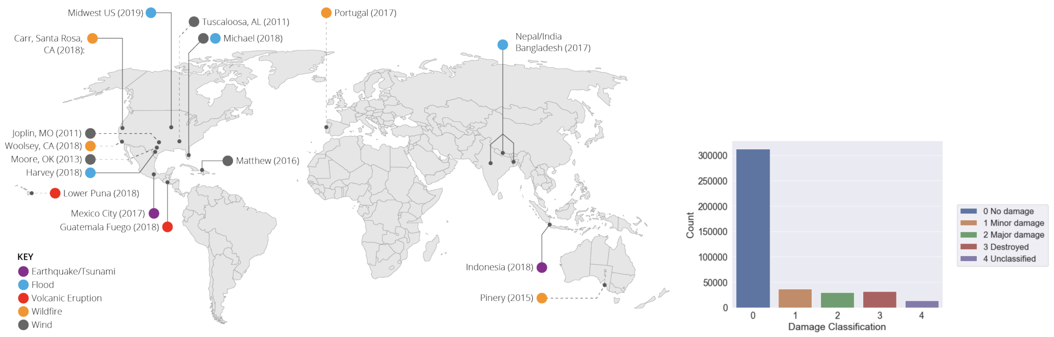
Figure 2: Architecture of feature extractor.

is then weighted by a learned parameter and added to the original upsampled output, which is then passed to the next upsample step.

3.4 Feature Map Extraction

In the context of U-Net [17], the integration of skip connections connecting corresponding encoder and decoder layers has proven effective in leveraging information from early layers during the generation of segmentation maps. This is especially valuable in image segmentation tasks, where precise localization and a nuanced understanding of both global and local features are pivotal for achieving accurate predictions.

Motivated by this insight, our approach involves the extraction of feature maps at each downsampling stage. Subsequently, we concatenate these feature maps into their corresponding layers at each upsampling stage. For this project, we explore two distinct feature extractor architectures. The first employs a conventional convolutional feature extractor with varying dimensions of convolutional layers. The second adopts a residual connection in Convolutional Neural Networks (ConvNets). Residual connections in ConvNets enhance training of deep architectures by addressing the vanishing gradient problem. In a residual block, the input is added to the output of convolutional layers, forming a skip connection. Mathematically, the output y is defined as $y = F(x) + x$, where x is the input and $F(x)$ is the transformed output. This design enables the model to learn residuals, simplifying optimization and facilitating the training of deeper networks. Residual connections promote gradient flow, mitigating challenges associated with deep architectures and enhancing the overall performance of ConvNets in computer vision tasks. This modification enhances the model's



(a) Disaster types and disasters represented around the world.

(b) Distribution of damage class labels.

Figure 3: Information about xBD dataset.

ability to capture and integrate both low-level and high-level features, thereby contributing to the refinement of segmentation map generation. Figure 2 shows the architecture of the feature extractor.

4 Experiments

4.1 Dataset

We leverage the xBD dataset [9]. A new, large-scale dataset for the advancement of change detection and building damage assessment for humanitarian assistance and disaster recovery research. xBD provides pre- and post-event satellite imagery across a variety of disaster events with building polygons, ordinal labels of damage level, and corresponding satellite metadata. Furthermore, the dataset contains bounding boxes and labels for environmental factors such as fire, water, and smoke. xBD is the largest building damage assessment dataset to date, containing 850,736 building annotations across 45,362 km^2 of imagery.

Note that the distribution of damage classifications is highly skewed towards "no damage," which had more than eight times the representation of the other classes. "No damage," "minor damage," "major damage," and "destroyed" are composed of 313,033, 36,860, 29,904, and 31,560 polygons respectively. There are an additional 14,011 polygons labeled as "unclassified." Figure 3b breaks this difference down visually.

4.2 Experiment Settings

4.2.1 Baseline

In establishing our baseline models, we adopt the first-place solution from the challenge as our reference.

4.2.2 Training Details

We use 8 Nvidia V100 GPU for training. We employ a learning rate of 0.00015 and a weight decay of 0.000001. The learning rate is dropped at a factor of 10 at 70th and 90th percentile of the training schedule. We train one model for building localization and one model for damage classification. These hyperparameters remain the same for both the localization and classification model. For both models, we train the model for a total of 200 epochs. Because of quadratic memory footprint of Attention mechanism, we only train models on 512 resolution. For baseline methods, we train both on 512 and 1024 resolution.

4.3 Results

We qualitatively evaluate our methods against the top permforming model on the xView2[9] benchmark. We provide evaluation results of localization F1$_{loc}$, which evaluate the performance of binary

Model	Res	F1_{loc}	F1	Undamaged	Minor	Major	Destroyed
SeResNext50	1024	89.1	61.8	91.5	45.9	52.5	77.0
SeResNext50	512	86.7	58.4	91.2	44.8	48.2	69.3
ClimaX	512	**86.8**	**59.3**	**91.4**	**46.6**	**51.0**	**70.2**

Table 1: Single model performance of first place leaderboard solution on the test set. ClimaX outperforms SeResNext50, the first place on public leaderboard, under comparable resolution. Gray color denotes the SeResNext50 results trained and evaluated on a higher resolution.

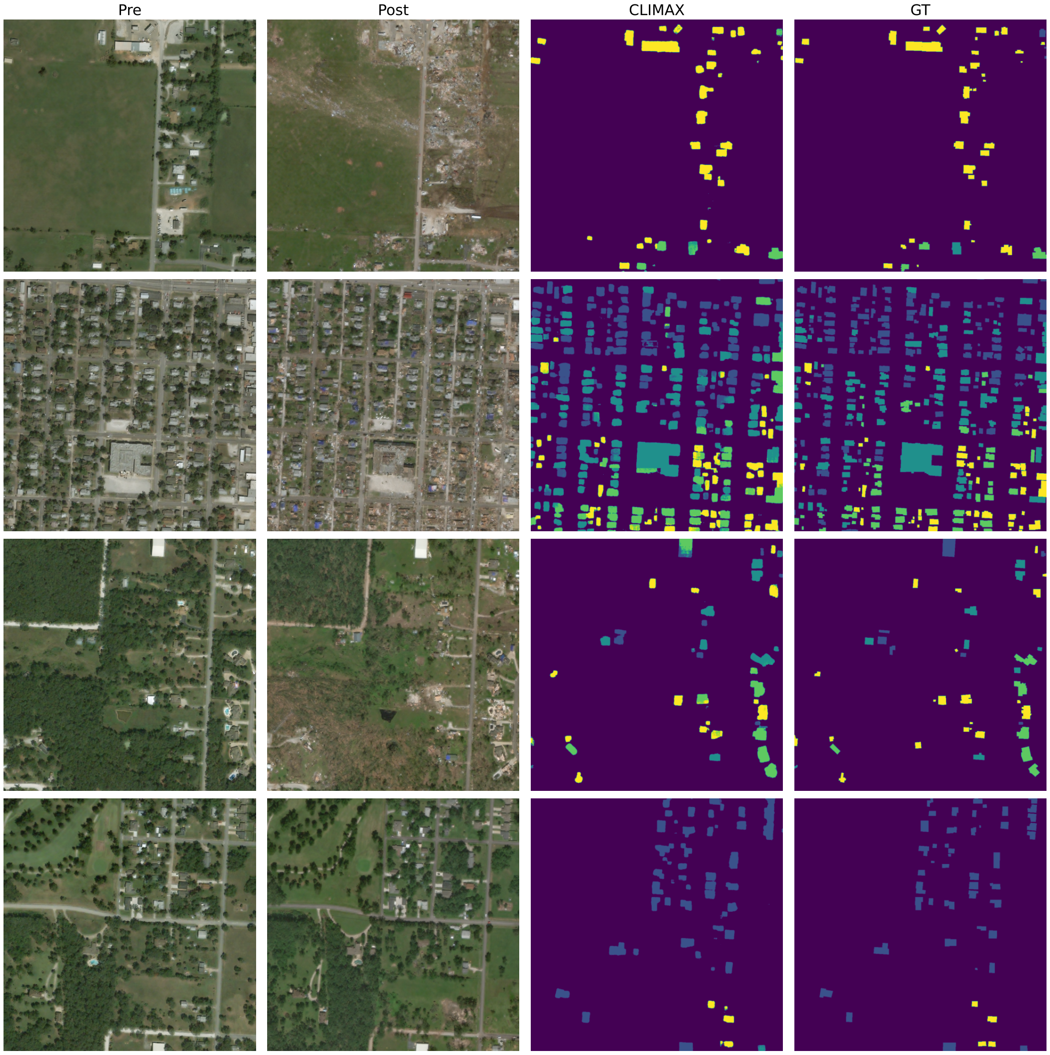


Figure 4: Visualization of Model Output. ClimaX is able to correctly identify varying levels of damage. In the figures, ligther color indicates higher level of damage. With yellow being the highest.

building segmentation on pre-diaster images, as well as classification F1, which evaluate the performance of damage classification. Results are shown in Table 1. Compared with baseline method on the same resolution, our model outperforms the baseline on all evaluation metrics. We achieve a gain of +0.1 on localization and +0.9 on classification. Most significantly, we outperform the baseline method by +2.8 on classifying buildings suffering major damages. This is particularly important in our use case.

Figure 4 presents visualizations of the model output. Notably, our model adeptly identifies diverse levels of damage within satellite images. This observation underscores the effectiveness of our proposed approach in elevating the accuracy and precision of damage assessment across a range of scenarios.

4.4 Ablation Studies

Model	F1_{loc}	Delta
Zeros Mask	38.6	-
ClimaX (Scratch)	62.3	+23.7
+ CMIP Pretrain	67.2	+4.9
+ 4X Batch Size	70.5	+3.3
+ Better Lr[1]	75.1	+4.6
+ Iterative Upsample (ConvTranspose2D)	78.9	+3.8
+ Group Norm	80.5	+1.6
+ ConvNeXt Block after Upsample	83.6	+3.1
+ Conv Feature Encoder	82.1	-1.5
+ Larger Conv Feature Encoder [2]	85.1	+3.0
+ Residual Connection in ConvNet	**86.8**	+1.7
+ IN 1k Pretraining [3]	86.3	-0.5

Table 2: Ablation Study on Design Choices. [1] We increased the learning rate to 0.00015. It is dropped by a factor of 10 at 140th epoch and 180th epoch. [2] We increase the dimension of convolution features from 128 across all scales to [128,258,768,1024] at different scales. [3] We use a pretrained checkpoint of ImageNet-1k classification task. [5].

Table 2 presents the performance differences resulting from various design choices. In the subsequent section, we delve into a detailed discussion on the impact of these different methods.

4.4.1 ClimaX

While Climax is pretrained on a significantly different dataset at a coarser resolution, loading pretrained weights on CMIP-6 [7] weather data still leads to a $+4.9$ improvement. While the data follows a completely different distribution, we hypothesize that the model can still inherit some general capability of spatial reasoning from pretrained weather models.

4.4.2 Training receipes

Changing the training receipe leads to a signifcant leap of $+7.9$. In particular, changing from a batch size of 1 per GPU to a batch size of 4 per GPU increased the performance by $+3.3$. Furthermore, sweeping across learning rates leads to an $+4.6$ increase in performance at optimal settings. These improvements follow the general rule of machine learning where scaling up the training process and tuning the hyperparameters can result in considerable improvements in performance.

4.4.3 Iterative Upsampling

The original method for retrieving the output for each pixel for a patch was to process the corresponding embedding through a linear layer that mapped the embeddings to the per-pixel values in that patch. Reshaping the ClimaX output and performing iterative upsampling solely using transposed convolution yielded a significant +3.8 improvement. This demonstrates how critical spatial information between the patches is for an accurate reconstruction. Additionally, adding group normalization between upsample steps and increasing the upsampling steps from two 4x upsamples to four 2x upsamples further improves the performance by +1.6, showing how gradually upsampling over many steps yields better results than a few large upsamples. Lastly, adding the ConvNeXt block after each upsampling step improves performance by a significant +3.1. Considering that the performance improvement nearly matches the jump caused by the initial implementation of iterative upsampling, this demonstrates that ConvNeXt blocks perform well post downsampling and upsampling.

4.4.4 Feature Map Extraction

The performance improves by $+3$ after switching to a larger convolution feature encoder. This is because the higher-dimensional feature space allows for the capture of more intricate patterns and nuanced details present in satellite images, crucial for accurate damage assessment. Additionally, the larger feature space helps address information loss during downsampling, providing a richer representation that enhances the overall segmentation accuracy.

The performance improves by $+1.7$ after switching to a residual connection encoder. This is attributed to the advantages of preserving information throughout the network. Residual connections facilitate the flow of information from the input to the output, aiding the model in learning and retaining crucial features. This architectural choice eases the training process, particularly in capturing complex patterns and preserving fine details essential for accurate damage assessment.

Notably, the performance improvement gets worse by -0.5 upon adding pretrained weights from ImageNet. This raises the possibility of a domain shift between ImageNet and satellite images. The pretraining on ImageNet might not be directly transferable to the satellite image domain due to differences in image characteristics, context, and features. This underscores the importance of domain-specific pretrained weights or the need for fine-tuning on a dataset more closely aligned with the target task, such as damage assessment in satellite images. Adjusting the pretrained weights to align with the specific nuances of the satellite image domain could potentially enhance model performance in this context.

5 Limitations

During experimentation, we ran into a couple key limitations of our methodology. Since ClimaX is a vision transformer, it incurs a quadratic memory footprint due to attention. This impacted our choices for hyper-parameters and input sizes. For example, we had to reduce the input resolution for the images and increase patch size to ensure the models fit in memory. This also forced the use of exotic hardware to ensure there is enough VRAM to train the model. Additionally, the dataset is biased to particular regions around the world. Most disasters stem from North America, India, and Australia. Thus, it is possible that the trained model may perform significantly worse in areas that have significantly different terrain than the mentioned regions.

6 Future Work

Regarding future work, there are several promising paths to move forward with. For example, leveraging pre-trained satellite image models, such as SatMAE [2], in geospatial models can mitigate domain shift issues, thus enhancing the model's adaptability to diverse terrains. To address ClimaX's high memory usage, exploring cutting-edge techniques, such as flash attention[4, 3], can reduce the memory footprint of ClimaX while maintaining its accuracy. Taking this further, we could explore other non-transformer-based architectures to evaluate the memory and accuracy tradeoffs. Regarding damage assessment as a whole, it is valuable to further research methods of predicting damage before a natural disaster, as this would allow local agencies to create evacuation plans and other mitigation strategies. Also, the notion of general damage assessment beyond natural disasters warrants further research.

References

[1] C.-S. Cheng, A. H. Behzadan, and A. Noshadravan. Deep learning for post-hurricane aerial damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, 36(6):695–710, 2021.

[2] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[3] T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023.

[4] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[8] Y. Gao, X. Sun, and C. Liu. A general self-supervised framework for remote sensing image classification. *Remote Sensing*, 14(19):4824, 2022.

[9] R. Gupta, R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] A. B. Khajwal, C.-S. Cheng, and A. Noshadravan. Post-disaster damage classification based on deep multi-view image fusion. *Computer-Aided Civil and Infrastructure Engineering*, 38(4):528–544, 2023.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[13] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, June 2022.

[14] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019.

[15] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[16] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[18] O. Ronneberger, P. Fischer, and T. Brox. Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015 Conference Proceedings*, 2022.

[19] C. Wu, F. Zhang, J. Xia, Y. Xu, G. Li, J. Xie, Z. Du, and R. Liu. Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets. *Remote Sensing*, 13(5):905, 2021.

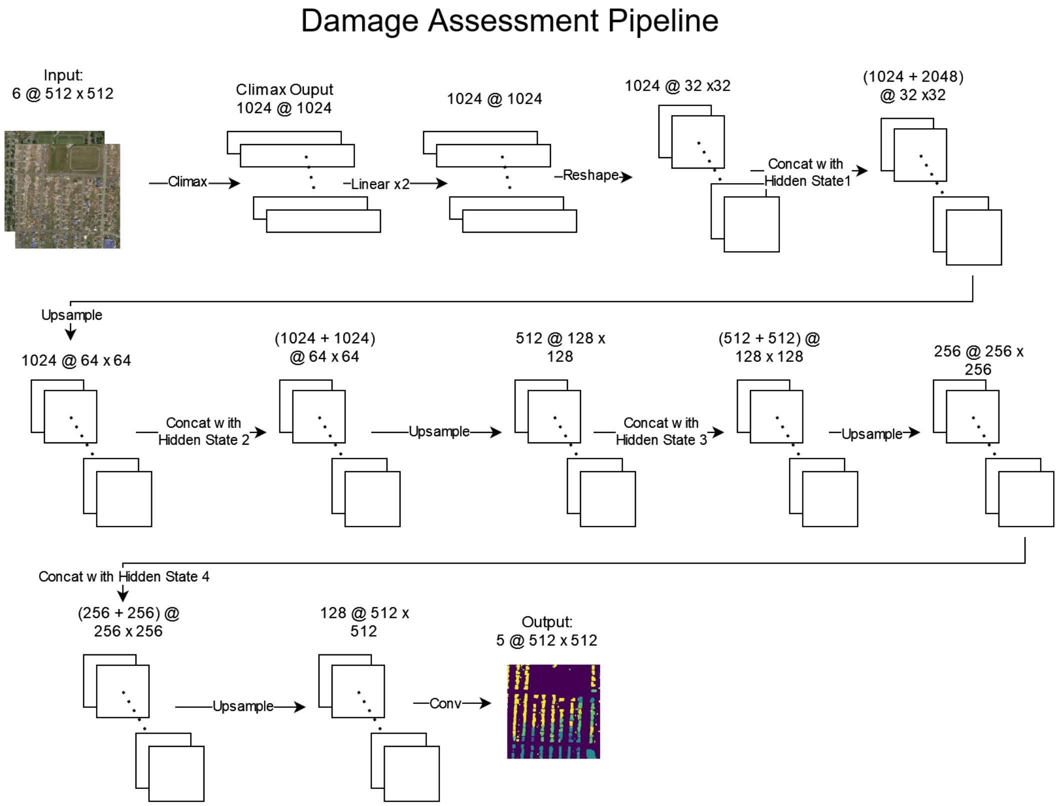
[20] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021.

Appendix

A.1 Work Split

Shufan Li worked on extending the base code to distributed training, adapting ClimaX to satellite imagery, and running experiments. Rishab Dalai worked on iterative upsampling and integrating architecture changes into the head. Chun-Yao Chang worked on implementing different kinds of feature extractors.

A.2 Implementation Details



In addition to what was mentioned in the main paper, we provide further implementations details.

Handling Pre-and-Post Images We concatenate the pre-disaster and post-disaster images along the channel dimension. This leads to a final input dimension of $N \times C \times H \times W$, where N is batch size, C is the number of channels, H, W are height and widths respectively.

Sliced Forward Due to memory constraints, we used a sliced encoding strategy. The 1024 sized input is chipped to 4 chips of size 512. The output is re-concatenated to the final output.

Normalization We adopted GroupNorm with 4 groups for all normalization layers.