2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR) | 979-8-3503-5142-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/MIPR62202.2024.00072

Parameter-Efficient Adaptation of Foundation Models for Damaged Building Assessment

Fei Zhao and Chengcui Zhang
Department of Computer Science
The University of Alabama at Birmingham
Birmingham, USA
{larry5, czhang02}@uab.edu

***Abstract*—The accurate and timely assessment of building damage is critical for effective post-disaster response efforts. However, traditional methods, reliant on manual inspection, are time-consuming and impractical in the face of large affected areas. This work introduces a novel vision foundation model-based framework (SAM-RS) that leverages the knowledge embedded in pre-trained Segment Anything Model (SAM) with the remote sensing imagery to enhance building damage assessment performance. The proposed SAM-RS exploits pairs of high-resolution satellite images captured pre- and post-disaster, processing them through the SAM to boost the downstream damaged building segmentation task. Meanwhile, the parameter-efficient fine-tuning paradigm (PEFT) is adopted to accelerate the training process. We propose a multi-stage fusion adapter module (MSFA), injected at the end of each Transformer block to merge visual information from different temporal states of the observed area, thereby enhancing the model's ability to discern subtle differences indicative of damage levels. Comparative evaluations demonstrate that SAM-RS achieves a F1 score of 0.68 surpassing state-of-the-art (SOTA) models by 18.63% on building damage assessment, marking this approach as a pioneering application of foundation models in the domain of disaster management technology. This advancement not only sets a new benchmark for rapid and precise damage evaluation but also significantly reduces the computational and storage demands typically associated with such tasks, paving the way for its adoption in operational disaster response workflows.**

***Index Terms*—Deep Learning, Vision Foundation Model, Adaptation, PEFT, Damage Assessment**

I. Introduction

Natural disasters such as earthquakes, floods, and hurricanes can cause extensive economic and human losses, with damages often running into billions of dollars and posing severe risks to life, as shown in Fig. 1. Rapid and precise assessment of building damage in the aftermath of such events is essential for minimizing adverse outcomes by facilitating quick emergency service deployment, targeted resource allocation, and informed decision-making regarding evacuations. However, traditional assessment methods, which rely heavily on manual inspections, are time-consuming and subject to human errors, particularly in large-scale disaster scenarios. This inefficiency highlights the urgent need for scalable and reliable damage assessment methods.

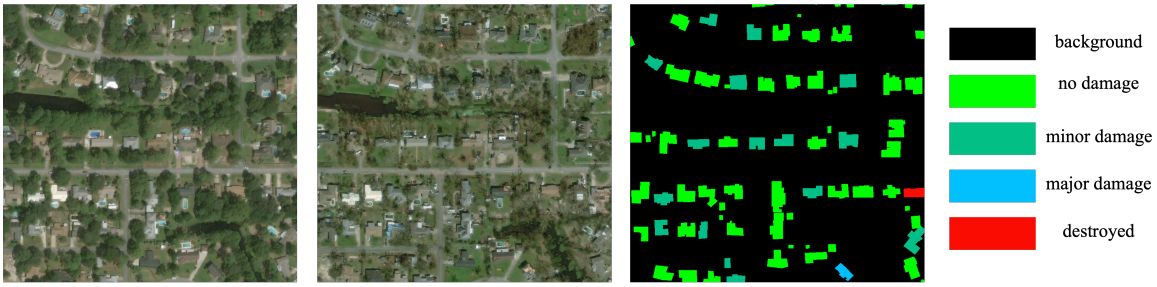
Recent advancements in remote sensing technologies have revolutionized the ability to monitor building conditions swiftly by capturing high-resolution satellite imagery after disaster events. This capability allows for rapid access to comprehensive data across affected areas, sidestepping the time-consuming and often subjective on-site manual inspection processes traditionally used for damage assessment. Concurrently, deep learning technologies have gained prominence, automating the prediction of building damage conditions effectively. However, deep learning-based methods typically encounter several limitations: 1. Generalization limitation: the state-of-the-art (SOTA) deep learning models are often trained on specific types of disasters and geographical locations, which severely restricts their applicability to new, unseen scenarios. 2. Computational and time constraints: the training of these models requires substantial computational resources and extensive datasets, making the process both time-consuming and costly. Adapting these models to handle new types of disasters involves a retraining process, further exacerbating the resource demands. Therefore, there is a critical need for a method that not only handles various types of disasters across different locations but also adapts to new data with minimal computational and time cost.

Vision foundation models (VFMs), pre-trained on diverse and extensive datasets, offer a profound base of knowledge that excels in various computer vision tasks. Their comprehensive pre-training provides VFMs with robust and versatile high-level feature representations, crucial for specialized tasks such as damage assessment. The parameter-efficient fine-tuning (PEFT) paradigm further enhances the adaptability of these models to new tasks or datasets with minimal modifications, significantly reducing the necessity for extensive retraining. By fine-tuning only a select subset of parameters, PEFT enables the rapid deployment of these models in new disaster scenarios, thereby drastically cutting down both computational costs and deployment time. To address the limitations of SOTA methods in building damage evaluation, this work introduces the SAM-RS framework, which utilizes the **S**egment **A**nything **M**odel (SAM) [1] integrated with **r**emote-**s**ensing imagery. As the pioneering application of foundation models to this task, SAM-RS not only enhances the precision and efficiency of damage assessments but also sets a new benchmark in disaster management technology.

Our contribution can be summarized as follows:

1. **Foundation Model-based Damage Segmentation and Assessment**: We introduce a novel adaptation framework that

2770-4319/24/$31.00 ©2024 IEEE
DOI 10.1109/MIPR62202.2024.00072



(a) Pre-disaster (b) Post-disaster (c) Ground Truth

Fig. 1. Satellite imagery of an area affected by a hurricane, showing conditions before and after the disaster. Panel (a) displays the pre-disaster satellite image, while panel (b) shows the post-disaster satellite image of the same area. Panel (c) visualizes the ground truth for building damage assessments with polygons representing individual buildings. Each color corresponds to one of the four assessed damage levels, ranging from "no damage," "minor damage," "major damage," to "destroyed."

utilizes the Segment Anything Model (SAM), a pre-trained vision foundation model, tailored specifically for the task of damaged building segmentation. This framework, SAM-RS, leverages high-level features extracted from pre- and post-disaster satellite imagery to enhance segmentation accuracy. Additionally, we incorporate a parameter-efficient fine-tuning (PEFT) strategy, injecting a minimal set of trainable parameters to rapidly align the model's knowledge for the specific task of damage assessment.

2. **Multi-Stage Fusion Adapter Module**: Integral to the SAM-RS framework, this module is designed to refine the interaction between feature sets derived from temporally distinct satellite images for the disaster-impacted areas. Positioned at multiple stages within the model's image encoder, the MSFA effectively synthesizes pre- and post-disaster data, enabling detailed and nuanced damage level predictions (i.e., no-damage, minor-damage, major-damage, and destroyed building). This method, achieving disaster-agnostic via analyzing&fusing the pre- and post-disaster difference focused on building objects, not only improves the precision of the segmentation but also enhances the model's overall robustness to varying disaster scenarios. More broadly, this could create a new practical foundation architecture for contrastive learning focused on object of interest.

This work introduces a highly efficient and scalable approach for assessing building damage following natural disasters. By analyzing pairs of pre- and post-disaster satellite images, our SAM-RS framework utilizes a disaster-agnostic adaptation strategy to expedite and refine damage evaluations. This method significantly boosts the responsiveness and effectiveness of disaster management initiatives, offering timely and accurate assessments critical for effective response planning.

The structure of this paper is organized as follows: Section II reviews SOTA methods in building damage evaluation and advancements in the use of foundation models. Section III details the innovative adaptation strategy and the architectural nuances of our SAM-RS framework. Section IV describes the dataset employed in this study, including the experimental protocols. Section V discusses the experimental outcomes and insights from an ablation study that underscores the efficacy of our proposed modules. Section VI provides concluding remarks and explores potential avenues for future research to further enhance the utility and applicability of foundation models in disaster-related assessments.

II. Background and Related Work

A. Damage Evaluation Methods

The evaluation of building damage using remote sensing imagery has intensified with the advancement of deep learning technologies, which have substantially improved the accuracy and automation of assessments over traditional manual methods. Nonetheless, SOTA approaches face several significant challenges: 1. Limited generalizability: many existing models are meticulously designed for specific disaster types, such as hurricanes or earthquakes, and are often restricted to particular geographical regions. For example, the authors in [2] developed a building damage classification model using Inception networks [3] specifically for earthquake-affected areas in Qinghai Province, China. Similarly, the authors in [4] tailored their neural networks exclusively to hurricane damage assessment. This specialization significantly restricts their utility across varied disaster scenarios. 2. Two-stage processing: Approaches such as the one proposed in [5] involve a two-stage process that first detects buildings using a U-Net model and then classifies the damage levels of the detected buildings by using another convolutional neural network. This method increases computational complexity and requires separate models for each task. 3. Pre-disaster image underutilization: a significant number of studies do not adequately leverage the available pre-disaster imagery, which is critical for establishing a baseline against which changes can be measured. The failure of exploiting the comparative information from pre- and post-disaster imagery can severely impact the accuracy of change detection and damage assessment, as demonstrated in the limited approach of [5]. 4. Binary damage assessment: many models simplify damage evaluation to a binary outcome—damaged or undamaged—which fails to capture the nuanced spectrum of damage severity. For example, the work in [4] explored several convolutional neural network-based binary classification models on post-hurricane satellite images for building damage assessment. In real scenarios, fine-grained damage level segmentation providing more

detailed and valuable information is essential for optimizing rescue efforts. These challenges underscore the necessity for models that not only enhance generalizability across various types of disasters and regions but also optimize computational efficiency and expand the granularity of damage assessment.

B. Foundation Models and Parameter-Efficient Adaptation

Foundation models have revolutionized various domains of natural language processing and computer vision due to their training on vast, heterogeneous datasets that endow them with extensive generalization capabilities. Notable examples include BERT [6] for natural language tasks, GPT [7] for generative applications, Vision Transformer [8] for image analyses, and the Segment Anything Model (SAM) [9], known for its exceptional task-agnostic segmentation abilities. Specifically, SAM's robustness in segmenting diverse objects makes it an ideal candidate for adaptation to specialized tasks such as segmenting damaged buildings from satellite imagery.

Despite their strengths, foundation models such as SAM are not directly tailored for specific tasks such as end-to-end building damage assessment due to their generalist pre-training. This limitation can be effectively addressed by parameter-efficient fine-tuning (PEFT) techniques, which adapt these large models to specific tasks by tuning a minimal subset of parameters, thus avoiding the high costs associated with traditional full-model fine-tuning. Among PEFT methods, the use of adapters, which are trainable subnetworks, is especially beneficial. These adapters are inserted into the foundation model architecture, allowing for precise modifications to the model's internal representations without overhauling its foundational weights.

This research capitalizes on the PEFT framework through an innovative application to SAM, significantly enhancing its utility for post-disaster building damage assessment. We introduce a Multi-Stage Fusion Adapter Module (MSFA) integrated with SAM's architecture, enabling sophisticated feature fusion between pre- and post-disaster images. This integration not only captures the nuanced changes in the disaster-stricken landscapes but also outputs detailed and accurate damage assessments. Our approach markedly reduces both the computational load and the training time compared to traditional methods, ensuring that the model is not only effective but also scalable and quick to deploy in disaster-affected areas. By embedding the MSFA within SAM, we set a new benchmark for applying vision foundation models in disaster management, paving the way for rapid, reliable, and efficient damage evaluation solutions.

III. Methodology

This study introduces the SAM-RS framework, an innovative adaptation tailored for assessing building damage in disaster-impacted areas using high-resolution remote sensing imagery. SAM-RS enhances the Segment Anything Model (SAM) with adaptive insights specific to disaster scenarios through a novel Multi-Stage Fusion Adapter Module (MSFA), facilitating precise semantic segmentation of damaged structures.

A. Adaptation Strategy and Overall Architecture

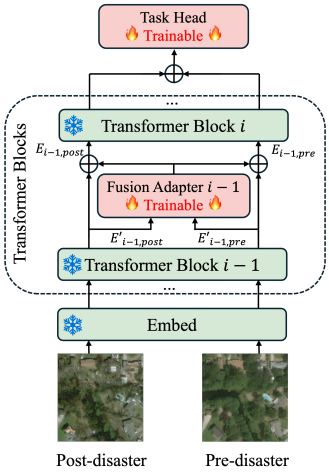


Fig. 2. Schematic representation of the dual-branch architecture incorporating MSFA for enhanced damage evaluation. The injected mutli-stage fusion adapters and the task head are trainable. Other layers in the model are frozen.

SAM-RS treats building damage evaluation as a semantic segmentation task. The overall architecture is shown in Fig. 2. It processes pre-disaster (I_{pre}) and post-disaster (I_{post}) RGB images through SAM's image encoder to obtain high-level features of observed areas. The workflow begins by processing these images through the patch embedding layer ($Embed(\cdot)$) to obtain image embeddings (E):

$$E_{0,pre} = Embed(I_{pre}) \tag{1}$$

$$E_{0,post} = Embed(I_{post}) \tag{2}$$

These embeddings are propagated through N Transformer blocks ($Block_i(\cdot)$) within the image encoder, where $i = 1, 2, \ldots, N$:

$$E'_{i,pre} = Block_i(E_{i-1,pre}) \tag{3}$$

$$E'_{i,post} = Block_i(E_{i-1,post}) \tag{4}$$

At each stage i, the MSFA ($Adapter_i(\cdot)$) adjusts the embeddings by integrating adaptive features from both pre- and post-disaster branches to enhance the model's sensitivity to changes between the pre- and post-disaster states:

$$E_{i,pre} = E'_{i,pre} + Adapter_i(E'_{i,pre}, E'_{i,post}) \tag{5}$$

$$E_{i,post} = E'_{i,post} + Adapter_i(E'_{i,pre}, E'_{i,post}) \tag{6}$$

The adapter-enhanced features: $E_{N,pre}$ and $E_{N,post}$ are finally used to generate the segmentation output through a specialized task head ($Head(\cdot)$):

$$Y_{seg} = Head(E_{N,pre}, E_{N,post}) \tag{7}$$

The integration of MSFA enhances SAM's capability to interpret complex remote-sensing visual data, capturing detailed damage indicators across pre- and post-disaster imagery. This structured approach ensures efficient, context-aware processing and accurate damage segmentation.

B. Multi-Stage Fusion Adapter Module (MSFA)

The MSFA is pivotal in enhancing the feature processing capabilities of SAM. Positioned at strategic intervals within the image encoder, MSFA refines feature integration from dual-temporal image inputs (pre- and post-disaster images), essential for assessing varying (subtle) damage levels. The detailed design of MSFA is shown in Fig. 3:

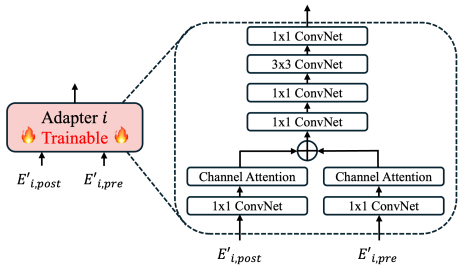


Fig. 3. The architecture of MSFA

MSFA processes the embeddings extracted from pre-disaster ($E'_{i,pre}$) and post-disaster ($E'_{i,post}$) images through a series of convolutional layers (ConvNet) and attention layers, structured as follows: 1. dimensionality reduction: features are compressed to a lower-dimensional space (from 768 to 192 channels) using a 1x1 ConvNets, enhancing computational efficiency. This reduction not only decreases computational load but also prepares the features for more efficient subsequent processing. 2. channel attention: a Channel Attention Mechanism [10] is adopted to refine the feature focus, reducing redundancy and aligning feature extraction more closely with damage indicators. This mechanism filters out irrelevant information, ensuring that the model concentrates on the most salient features needed for accurate damage assessment. 3. feature fusion and expansion: following attention, the features are expanded and synthesized through a sequence of convolutions. This includes a 3x3 depthwise convolutions (3x3 ConvNet) designed to capture local spatial patterns effectively. Then, the features are projected back to the original dimension (768) using 1x1 ConvNet, preparing them for subsequent processing.

The adapter-enhanced features $E_{N,pre}$ and $E_{N,post}$ are combined and then upsampled using two sets of 2x2 transpose ConvNet ($Upsample(\cdot)$) to align with the input image's resolution scale. Following this, a multi-scale convolutional strategy ($MSConv(\cdot)$), proposed in [11], integrates these features across various scales. The segmentation map is then derived using a linear layer ($Linear(\cdot)$) , which classifies the extent of damage for each pixel into categories ranging from "no damage" to "destroyed".

$$E_N = E_{N,pre} + E_{N,post} \tag{8}$$

$$Y_{seg} = Linear(MSConv(Upsample(E_N))) \tag{9}$$

By embedding the MSFA within the SAM architecture, we harness enhanced contextual and spatial information from pre- and post-disaster satellite imagery, which significantly improves the segmentation accuracy and robustness of the damage assessment model. In this integration, all parameters in the SAM layers are frozen to preserve their pre-trained generalist capabilities, with only the parameters in the MSFA and the task head being trainable. This selective training not only refines the model's predictive performance but also ensures computational efficiency in training process.

C. Loss Function

Our dataset, detailed in Section IV, exhibits significant class imbalance, predominantly favoring "no damage" instances, which can skew model training towards the most common class. To address this, we built a customized loss function that integrates weighted Cross-Entropy Loss, Dice Loss, and Focal Loss [12].

$$Loss_{total} = w_{ce}L_{ce} + w_f L_{focal} + w_d L_{dice} \tag{10}$$

In Equation (10), w_{ce}, w_f, and w_d represent the weights assigned to Cross-Entropy, Focal, and Dice Losses, respectively. For details on the method used to tune these weights, see [12]. This customized loss is designed to enhance the model's performance across variably represented categories by specifically addressing the challenges of imbalanced data and difficult-to-classify samples.

IV. Dataset and Experiment

A. Dataset Overview

The xBD dataset [5], central to this study, is instrumental in advancing building damage assessment research, particularly within the contexts of humanitarian assistance and disaster recovery. It comprises 850,736 annotated buildings affected by 19 different types of natural disasters, including floods, hurricanes, and earthquakes [5]. Each annotation spans a broad spectrum of damage levels, from "no damage" to "destroyed", shown in Fig. 1. Unique for its extensive collection, the dataset includes 22,068 high-resolution satellite images (1024x1024 pixels, 3 channels) captured before and after the disasters. These features support the development of robust models capable of detailed damage analysis. A challenge in the xBD dataset is class imbalance: "no damage" instances far outnumber the "minor" and "major" damage cases. To address this, we implemented a data augmentation strategy, proposed in [13], to enhance the representation of less frequent categories, thereby improving the model's training and performance on these critical but under-represented classes. The distribution of the dataset across different splits is detailed in Table I.

TABLE I
Distribution of Images Across Different Dataset Splits

Split	Image Number
Train	12,030
Validation	640
Test	1,866

B. Experiments and Metrics

The experiments were conducted on the dataset shown in Table I, using the pre-trained base version of SAM as provided in [1]. Training was performed on a NVIDIA A100 GPU, using a batch size of 4. The AdamW optimizer was utilized, with a learning rate set at 1×10^{-5}, and the model was trained over 100 epochs on the training dataset to ensure adequate convergence for the complex task of damage assessment.

To evaluate model performance, $F1$ score was calculated for each of the four defined damage levels: no-damage, minor-damage, major-damage, and destroyed. Addressing the dataset's class imbalance, the Harmonic Mean of these $F1$ scores was used as the aggregate metric, providing a balanced measure of precision and recall across all categories. The Harmonic $F1$ score is defined by the following equation, where $F1_i$ represents the $F1$ score for the i^{th} damage level ranging from 1 to 4 corresponding to the 4 damage levels (with 1 being "no damage" and 4 being "destroyed"), and ϵ is a small constant to prevent division by zero:

$$F1_{harmonic} = \frac{4}{\sum_{i=1}^{4} (F1_i + \epsilon)^{-1}} \tag{11}$$

V. Result and Discussion

As Table II illustrates, SAM-RS model achieves the highest Harmonic Mean $F1$ score of 0.6800, indicating its superior balance of precision and recall, particularly in handling the class imbalances inherent in the xBD dataset. This score is a substantial improvement over the baseline model's 0.0342 [5] and the Siamese model's 0.5732 [13], marking improvement of nearly 20 times and 18.63%, respectively. Furthermore, SAM-RS consistently outperforms the other models across all individual damage levels, demonstrating its robustness and effectiveness in distinguishing between varying extents of damage.

TABLE II
THE $F1$ SCORE OF EACH MODEL

Damage Level	Baseline	Siamese	SAM-RS
No	0.6631	0.7140	**0.8620**
Minor	0.1435	0.3955	**0.4840**
Major	0.0094	0.6037	**0.7129**
Destroyed	0.4657	0.7181	**0.7981**
Mean	0.320	0.6078	**0.7142**
Harmonic Mean	0.0342	0.5732	**0.6800**

These results underscore SAM-RS's qualitative advancements in applying foundation models to the field of disaster management. By integrating SAM with MSFA, we have significantly enhanced the model's capability to analyze and interpret complex disaster-affected imagery for more accurate damage assessment. This achievement marks a pioneering application of foundation models to building damage evaluation and paves the way for further innovations in the utilization of deep learning technologies for disaster response and recovery.

A. Ablation Study of the Fusion Adapter Module

We conducted an ablation study to quantitatively assess the contribution of the Multi-Stage Fusion Adapter Module (MSFA) to the SAM-RS model's performance. This study compared the performance of the model with and without MSFA: 1. SAM-RS with MSFA: this is a complete implementation of our proposed SAM-RS framework, in which the MSFA and the task head are trainable. 2. SAM-RS without MSFA: the model used the pre-trained SAM's image encoder to process pre- and post-disaster images, relying on foundational features without enhancement through MSFA. Only the task head is trainable.

TABLE III
$F1$ SCORES COMPARISON UNDERSCORING THE MSFA'S EFFECTIVENESS

Damage Level	Without MSFA	With MSFA
No	0.4216	**0.8620**
Minor	0.0607	**0.4840**
Major	0.0742	**0.7129**
Destroyed	0.0123	**0.7981**
Mean	0.1422	**0.7143**
Harmonic Mean	0.0352	**0.6800**

The results in Table III elucidate the critical role of MSFA in the SAM-RS model. The model without MSFA performs notably well only in detecting "no damage" scenarios, suggesting that while the pre-trained SAM image encoder can effectively identify undamaged buildings, it struggles to utilize contrastive pre- and post-disaster information essential for assessing actual damage. In contrast, the inclusion of MSFA allows the model to effectively discern and classify varying damage levels, as evidenced by substantial gains in $F1$ scores for "minor damage", "major damage", and "destroyed" categories.

These findings underscore MSFA's effectiveness in bridging the gap between foundational model capabilities and the specific needs of damage assessment tasks. By facilitating better integration and contextualization of satellite image data, MSFA significantly enhances the model's predictive performance and robustness, making it a valuable component for deploying foundation models in complex, real-world disaster response applications.

B. Ablation Study of Pre-disaster Branch

We conducted an ablation study to examine the value of incorporating pre-disaster imagery into our model. 1. Single-branch model (Single): this model utilizes only post-disaster images for the damage evaluation task. 2. Dual-branch model (Dual): as our full model, it leverages both pre- and post-disaster images, allowing for comparative analysis and enhanced feature extraction across different temporal states, facilitated by our comprehensive adaptation strategy.

The results from Table IV clearly show significant performance gains when pre-disaster imagery is used. Specifically, the dual-branch model markedly outperforms the single-branch configuration across all damage levels, especially for minor and destroyed categories. As shown in Table IV, the minor damage class has an improvement from 0.2577 to 0.4840

TABLE IV
COMPARISON OF $F1$ SCORES AND PERCENTAGE IMPROVEMENTS HIGHLIGHTING THE EFFECTIVENESS OF THE DUAL-BRANCH MODEL.

Damage Level	Single	Dual	Improved (%)
No Damage	0.7711	0.8620	+11.79%
Minor Damage	0.2577	0.4840	+87.82%
Major Damage	0.6139	0.7129	+16.13%
Destroyed	0.5986	0.7981	+33.33%
Mean	0.5603	0.7143	+27.49%
Harmonic Mean	0.4719	0.6800	+44.10%

(**+87.82%**). This category shows the most significant improvement, underscoring the dual-branch model's enhanced ability to detect subtle damages that are typically overlooked by less sophisticated models. The substantial increase in $F1$ score for minor damages suggests that incorporating pre-disaster imagery allows the model to better contextualize slight alterations to the infrastructure, which are indicative of minor damages.

In real scenarios, the notable improvement in detecting "minor damage" could be particularly critical for disaster response efforts. Buildings with minor damage can often be overlooked in rapid assessments, leading to delayed or inadequate repairs that could escalate the damage severity in subsequent events. Our dual-branch model's proficiency in this area could greatly enhance operational responses, ensuring more accurate prioritization and resource allocation. Furthermore, the increased accuracy across all categories, especially for "destroyed" buildings, ensures that emergency measures can be directed promptly and effectively, minimizing potential hazards to rescue and recovery teams.

These results affirm the substantial benefits of integrating comprehensive pre- and post-disaster imagery analyses. Our dual-branch SAM-RS approach sets a new benchmark in disaster response analytics by providing detailed and actionable insights that are vital for optimizing rescue operations and rebuilding strategies.

VI. CONCLUSION AND FUTURE WORK

This study has developed and implemented a novel adaptation framework for building damage evaluation, employing the Segment Anything Model (SAM) in conjunction with a Multi-Stage Fusion Adapter Module (MSFA). Central to our methodology is the strategic use of pre- and post-disaster imagery, enabling comprehensive assessments of damage ranging from "no damage" to "destroyed". This integration of foundation models and PEFT significantly reduces computational demands and lessens the need for extensive training datasets, making the approach both efficient and scalable. Meanwhile, our model can handle various types of disasters across different locations since its disaster-agnostic adaptation design and training on the xBD dataset, which consists of diverse disaster scenarios. These improvements mark a significant advancement over traditional methods and underscore the potential of our framework in real-world applications.

Future efforts will aim to broaden the scope of this work to include a wider array of disaster types, thereby enhancing its applicability and robustness in diverse conditions. Furthermore, while the current work has concentrated on leveraging parameter-efficient adaptation techniques that adjust internal model representations, subsequent research will explore the integration of prompt learning strategies. Prompt tuning, by altering how models interpret input data through soft prompts, provides a nuanced means of task-specific fine-tuning without the need for modifying the model's internal representations.

This foundation model-based framework sets a new benchmark in disaster management, significantly enhancing disaster resilience and response capabilities. It introduces a practical foundation architecture designed to that guides the model to discern subtle differences in complex visual data, pivotal for contrastive learning focused on specific objects. This innovation not only boosts model performance but also broadens the scope of foundation models in precision tasks beyond traditional applications.

REFERENCES

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[2] H. Ma, Y. Liu, Y. Ren, D. Wang, L. Yu, and J. Yu, "Improved cnn classification method for groups of buildings damaged by earthquake, based on high resolution remote sensing images," *Remote Sensing*, vol. 12, no. 2, p. 260, 2020.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[4] B. Liu, C. Lyu, and S. Wang, "Convolutional neural networks' efficiency on binary classification of damaged building on post hurricane satellite imagery," in *International Conference on Cloud Computing, Performance Computing, and Deep Learning (CCPCDL 2022)*, vol. 12287, pp. 238–246, SPIE, 2022.

[5] R. Gupta, B. Goodman, N. N. Patel, R. Hosfelt, S. Sajeev, E. T. Heim, J. Doshi, K. Lucas, H. Choset, and M. E. Gaston, "xbd: A dataset for assessing building damage from satellite imagery," *ArXiv*, vol. abs/1911.09296, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[8] 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.

[9] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[11] Z. Zhang, X. Wang, S. Chen, X. Liu, Q. Chen, Y. Zhang, Y. Cao, and B. Liu, "Mlmsa: Multi-level and multi-scale attention for lesion detection in endoscopy," in *2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom)*, pp. 144–150, IEEE, 2023.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[13] F. Zhao and C. Zhang, "Building damage evaluation from satellite imagery using deep learning," in *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 82–89, IEEE, 2020.