Visual Prompt Learning of Foundation Models for Post-Disaster Damage Evaluation

Fei Zhao[1*], Chengcui Zhang[1], and Tianyang Wang[1]

[1]Department of Computer Science, The University of Alabama at Birmingham, Birmingham, USA.

Abstract

In response to the urgent need for rapid and precise post-disaster damage evaluation, this study introduces the **Vi**sual **P**rompt **D**amage **E**valuation (ViPDE) framework, a novel contrastive learning-based approach that leverages the embedded knowledge within the Segment Anything Model (SAM) and pairs of remote sensing images to enhance building damage assessment. In this framework, we propose a learnable cascaded **V**isual **P**rompt **G**enerator (VPG) that provides semantic visual prompts, guiding SAM to effectively analyze pre- and post-disaster image pairs and construct a nuanced representation of the affected areas at different stages. By keeping the foundation model's parameters frozen, ViPDE significantly enhances training efficiency compared to traditional full-model fine-tuning methods. This parameter-efficient approach reduces computational costs and accelerates deployment in emergency scenarios. Moreover, our model demonstrates robustness across diverse disaster types and geographic locations. Beyond mere binary assessments, our model distinguishes damage levels with a finer granularity, categorizing them on a scale from 1 (no damage) to 4 (destroyed). Extensive experiments validate the effectiveness of ViPDE, showcasing its superior performance over existing methods. Comparative evaluations demonstrate that ViPDE achieves an F1 score of 0.7014, surpassing state-of-the-art (SOTA) models by 22.37%. This foundation model-based approach sets a new benchmark in disaster management. It also pioneers a new practical architectural paradigm for foundation model-based contrastive learning focused on specific objects of interest.

1 Introduction

Throughout history, natural disasters have persistently challenged the resilience of our societies, often leading to tragic losses of life and widespread destruction. From 1998 to 2017, over 35,000 such events have resulted in more than one million deaths and 4.4 billion injuries [1]. For instance, the 2011 Fukushima tsunami resulted in significant human and structural losses [2]. In recent years, technological advancements in forecasting natural hazards have contributed to reducing fatalities from disasters. However, forecasting systems show varied reliability across different disaster types and are insufficient as standalone solutions for comprehensive emergency management. Critically,

the majority of disaster-related casualties occur within the initial hours following a disaster event, often due to delayed rescue operations. This reality positions effective disaster response as a crucial complement to forecasting efforts. Therefore, the ability to rapidly and accurately assess building damage is important to effective disaster response, guiding the deployment of rescue and aid efforts where they are needed the most. Meanwhile, beyond binary evaluation of damage, e.g., no-damage or damaged building, a nuanced determination of damage severity from minor to complete destruction is needed and instrumental in optimizing rescue team deployment and enhancing the overall response efficacy.

(a) The pre-disaster images (b) The post-disaster images

Figure 1: The pre- and post-disaster images of Fukushima. In the first row of images, by visual inspection, we can tell that lots of residential buildings are destroyed by the disaster. In the bottom row images, most buildings, including residential houses and commercial buildings, are destroyed by the disaster.

As the need for accurate damage evaluation becomes increasingly recognized, experts across diverse fields such as geology, meteorology, and oceanography are enlisted to provide real-time data analysis and expert guidance during rescue missions. However, the sheer scale of devastation a single disaster can inflict, impacting an entire city or country, renders manual assessment methods impractical and unsustainable. While expert input is invaluable, the extensive scope of major disasters can compromise the effectiveness of such evaluations, introducing a high likelihood of error due to fatigue and the overwhelming volume of assessments required. Additionally, these assessments are subject to biases and subjectivity inherent to individual evaluators, which could skew the critical objectivity needed in these urgent situations. Therefore, there is a pressing need for a reliable and

objective system capable of assessing damage across extensive areas, enhancing the efficiency and effectiveness of emergency response efforts by accurately determining the need and urgency of the situation.

As technology advances in remote sensing, satellites are capable of capturing high-resolution imagery for an area before and after a disaster event. As referenced in Figure 1, the discrepancy between pre- and post-disaster images illustrates the transformative impact caused by natural disasters on the built environment. Those completely destroyed buildings can be evident and easily recognized from images. High quality satellite images provide emergency management teams with a complementary approach to rapidly evaluate the extent of building damage for a disaster-affected area. However, accurately assessing the extent of damage with fine granularity, such as damage level 2 (minor damage) or 3 (major damage), of the buildings is hard but much needed. Such detailed and precise damage assessment can be crucial for directing emergency response efforts and allocating resources effectively to the areas in the greatest need.

(a) hurricane (b) flooding (c) tsunami

Figure 2: Sample images for hurricane, flooding, and tsunami. The images in the top row are satellite images captured before the disasters. The images in the bottom row are obtained after the disaster strikes these areas. In (a), most of the buildings are destroyed by the hurricane except the buildings in the right corner. In (b), the building is major damaged. It is surrounded by waters. In (c), the two commercial buildings within the red rectangle are destroyed by the tsunami.

Machine learning techniques have drastically transformed the analysis of satellite imagery, automating processes and enhancing prediction accuracy. However, a critical limitation of these state-of-the-art models is their specialized design, which is often tailored for a specific type of disaster or distinct geographical region. This specialization significantly restricts their generalizability and applicability across varying scenarios. When confronted with disasters beyond their training scope,

these models often exhibit diminished performance, underscoring their limited versatility across different disaster types and regions. Furthermore, these approaches typically necessitate training the entire model from scratch, which demands significant computational resources, extensive datasets, and substantial training time. This requirement not only increases the training burden but also limits the models' ability to be promptly applied to new or unforeseen disaster scenarios. This specialization and inefficiency highlights a critical research gap: the absence of a unified model that can seamlessly handle various disaster types and geo-locations while circumventing the computational and data-intensive demand of traditional deep learning's full-model training process.

Therefore, our research proposes the Visual Prompt Damage Evaluation (ViPDE) framework, a novel contrastive learning-based approach that leverages the embedded knowledge within foundation models and the discrepancy within pairs of remote sensing images to enhance building damage assessment. Our contribution can be summarized as follows:

- **Visual Prompt Damage Evaluation Framework(ViPDE):** We propose a contrastive learning-based dual-branch architecture that enables the Segment Anything Model (SAM) to dynamically utilize high-level features from paired pre- and post-disaster satellite images to enhance building damage evaluation. ViPDE integrates cascaded lightweight learnable Visual Prompt Generators while keeping the foundation model's pre-trained weights frozen. This approach effectively utilizes the contrasts between the image pair to enhance accuracy and efficiently fine-tunes the model's performance for damage evaluation while avoiding the extensive retraining that traditional methods typically require.

- **Visual Prompt Generator (VPG):** We introduce a learnable, lightweight, cascaded Visual Prompt Generator that provides tailored visual prompts, enriched with semantic information from pre- and post-disaster images. These prompts act as navigational cues, focusing the pre-trained vision foundation models on essential damage indicators. The generator's design, with its minimal trainable parameters, strategically amplifies the model's pre-trained knowledge, enabling a more precise and expedient evaluation of disaster damage.

- **Disaster-Agnostic Damage Evaluation:** By leveraging the power of contrastive learning, our framework utilizes discrepancies between pre- and post-disaster image pair to gain insights into damage levels. This approach enables the model to handle a wide range of natural disasters and geographic locations, allowing for direct application to varied disaster scenarios without additional training. By analyzing contrasts in image data, our model segments buildings into categories of damage: "no damage," "minor damage," "major damage," and "destroyed." These detailed damage assessments are vital for the strategic deployment of rescue teams and improving the overall effectiveness of disaster response efforts.

The remainder of this paper is structured as follows: Section 2 provides a review of the relevant literature. Section 3 elaborates on our methodology, detailing the innovative techniques utilized in our approach. Section 4 describes the dataset used in this study. Section 5 outlines the experimental setup and metrics. Section 6 analyzes the results, discussing the implications and significance of our

findings. Finally, Section 7 summarizes the key outcomes of our research and proposes directions for future work, highlighting potential advancements of our work.

2 Related Work

2.1 Damage Evaluation Methods

As the field of damage evaluation evolves, it has produced a range of sophisticated models and algorithms capable of assessing damage severity from satellite imagery. However, despite these technological advances, the field still faces significant challenges that hinder further progress: **(1) Image Classification and Damage Detection:** Earlier approaches utilized image classification models to distinguish between damaged and undamaged buildings within satellite imagery of disaster-affected areas. The authors of [3] propose such a model specifically for hurricane damage detection, although its application was confined to hurricanes, highlighting a gap in generalizability across different disaster types. Similarly, the authors from [4] develop three image classification models for damage prediction that lacked building localization capabilities, suggesting the necessity of integrating these models with localization techniques for practical use. **(2) Utilization of Pre- and Post-Disaster Imagery:** The difference between pre- and post-disaster imagery has been leveraged for assessing building damage. The authors of [5] utilize this contrast but faced challenges with varying building types. In contrast, [6] employs a method that identifies building locations in pre-disaster images, using those same locations to extract features from the corresponding post-disaster images. However, this approach is limited by its reliance on spatial consistency across the image pairs. Any misalignment between pre- and post-disaster images can compromise its reliability. Our unified model addresses this limitation by processing the pre- and post-disaster image pair simultaneously, allowing the model to learn to manage misalignments within the data. With sufficient training data, the deep learning model can automatically adapt to spatial inconsistencies, enhancing robustness in damage detection. **(3) Ensemble Methods:** The adoption of two-stage methods for damage detection, as proposed in [7, 8], increases computational cost because it utilizes separate models for building detection and damage level classification. The common challenges faced by all of these approaches are: the specificity of models to particular disasters, the binary approach to damage assessment, the reliance on two-stage processing pipelines, the underutilization of valuable pre-disaster imagery, and the dependency on hard-coded distance calculations for feature comparison. These gaps highlight the need for a more adaptable, universally applicable model that can handle various disaster types and geographical locations efficiently.

2.2 Prompt Learning

The advent of foundation models such as Chat-GPT [9] and GPT-4 [10] has markedly advanced AI, showcasing their broad applicability, particularly in natural language processing (NLP). This success has spurred adaptations in computer vision with models such as BEiT [11], ViT [12], and the Segment Anything Model (SAM) [13], which apply transformer-based processing to visual tasks. However, these models are not inherently suited for tasks requiring contrastive image analysis, such

as damage assessment in remote sensing. The evolution from "pre-training and fine-tuning" to "pre-training and prompting" represents a significant trend in the deployment of foundation models [14, 15]. Innovations such as VPT [15], which appends a set of learnable parameters to transformer encoders, significantly outperform full fine-tuning across multiple downstream recognition tasks. AdaptFormer [16] incorporates lightweight modules into ViT, achieving superior performance over fully fine-tuned models on action recognition benchmarks.

Overall, addressing this gap, our work introduces a unified model capable of handling diverse disasters and locations with greater efficiency. By leveraging a pre-trained vision foundation model with frozen backbone parameters, we significantly reduce computational costs and expedite the training process. Central to our approach is the innovative use of a learnable visual prompt generator that requires training on only a minimal number of parameters. This methodology not only enhances training efficiency but also extends the model's applicability, offering a scalable and comprehensive solution to disaster damage assessment across various environments.

3 Methodology

In this section, we introduce the Visual Prompt Damage Evaluation (ViPDE), a novel contrastive learning-based approach that utilizes a pre-trained vision foundation model, specifically SAM, for the semantic damage segmentation of satellite imagery. The cornerstone of ViPDE is the learnable Visual Prompt Generator (VPG) module, which can provide visual prompts guiding SAM to effectively analyze pre- and post-disaster image pairs and form a nuanced representation of the affected. As illustrated in Figure 3, VPG is meticulously designed to generate multi-stage visual prompts enriched with semantic information extracted from both pre- and post-disaster imagery. These prompts serve as navigational semantic cues, steering the foundation model's focus toward critical features indicative of damage. The VPG automates contrastive learning, enabling the model to effectively differentiate between damaged and undamaged areas, thereby significantly boosting accuracy in damage assessment.

3.1 The Overall Architecture of ViPDE

Our proposed approach treats the building damage evaluation task as a semantic segmentation task. It utilizes a pair of pre-disaster (X_{pre}) and post-disaster (X_{post}) RGB images as the input to enhance building damage evaluation accuracy. The frozen SAM's image encoder, structured in a sequence of Transformer blocks, is adopted to extract nuanced features from those images. For each image pair, X_{pre} and X_{post}, the process begins by projecting them into initial token embeddings $E^0_{pre} = Embed(X_{pre})$ and $E^0_{post} = Embed(X_{post})$, respectively. These embeddings are then combined with the visual prompt P^1 (Figure 3) to create new merged embeddings, $E^0_{merged,pre}$ and $E^0_{merged,post}$. Subsequently, these refined embeddings are processed through N encoder Transformer blocks $Block_i(\cdot)$, where $i = 1, \ldots, N$:

$$E^i_{pre} = Block_i(E^{i-1}_{merged,pre}) \tag{1}$$

$$E^i_{post} = Block_i(E^{i-1}_{merged,post}) \tag{2}$$

Here, E^i_{pre} and $E^i_{post}(1 \leq i \leq N)$ are outputs from the i^{th} encoder block for pre- and post-disaster images, respectively. Our Visual Prompt Generator (VPG) module augments the original RGB data flow with context-specific semantic visual prompts tailored to damage assessment:

$$E^{i-1}_{merged,pre} = E^{i-1}_{pre} + P^i \tag{3}$$

$$E^{i-1}_{merged,post} = E^{i-1}_{post} + P^i \tag{4}$$

$E^{i-1}_{merged,pre}$ and $E^{i-1}_{merged,post}$ $(1 \leq i \leq N)$ denote the input token sequences for pre- and post-disaster images, respectively, at the i-th stage, each enhanced by the addition of prompt P^i provided by the VPG module. This inclusion of disaster-specific prompts at multiple stages effectively enriches SAM's semantic analysis across different levels of feature abstraction.

These processed features are combined to generate the final segmentation output Y_{seg} through a specialized decoder $Decoder(\cdot)$ that accounts for the nuances of disaster impact:

$$Y_{seg} = Decoder(E^N_{pre}, E^N_{post}) \tag{5}$$

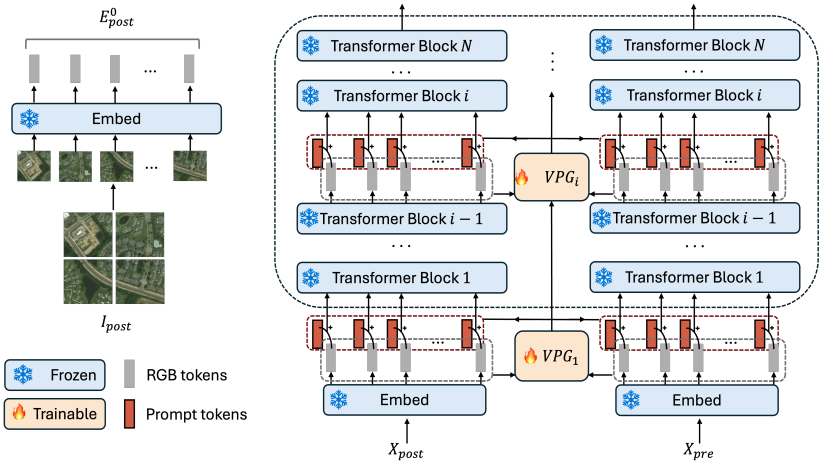


Figure 3: The overall architecture of ViPDE, illustrating the dual-branch input of pre- and post-disaster images processed through the Segment Anything Model (SAM) encoder. The Visual Prompt Generator (VPG) introduces stage-wise visual prompts that guide the encoder's focus on disaster-relevant features, enhancing damage segmentation precision. Outputs are decoded to a damage-level segmentation map, categorizing buildings into varying damage levels.

ViPDE maintains all SAM parameters in a frozen state, including those for patch embedding and feature extraction, except for the learnable VPG module. The VPG only introduces a minimal number of trainable parameters for generating and integrating visual prompts into the pre-disaster and post-disaster input sequences. By doing so, ViPDE ensures the pre-trained model's architecture is preserved while enabling efficient, task-specific adaptations through prompt insertion. This approach allows for precise post-disaster damage assessment without necessitating extensive retraining

or requiring an explicit distance function for alignment between pre- and post-disaster features.

3.2 Visual Prompt Generator (VPG)

As depicted in Figure 3, our learnable VPG module is innovatively integrated at multiple stages within the foundational network to inject task-specific enhancements. The architecture of VPG is visualized in Figure 4 .

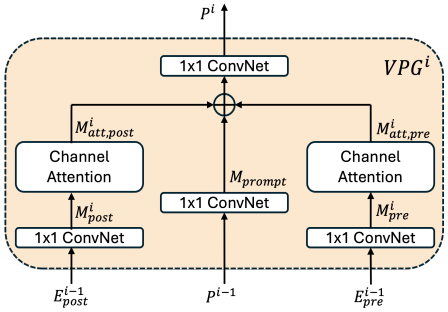


Figure 4: The visualization of VPG generator

As shown in Equation 6, visual prompt P^i is generated by $VPG^i(\cdot)$ module based on the prompt P^{i-1} and embedding sequences E^{i-1}_{pre} and E^{i-1}_{post} generated in the previous stage. The VPG module is designed to learn and apply visual prompts that effectively capture the difference between the pre- and post-disaster states. The process proceeds as follows:

$$P^i = \begin{cases} VPG_i(E^{i-1}_{pre}, E^{i-1}_{post}), & \text{if } i = 1 \\ VPG_i(E^{i-1}_{pre}, E^{i-1}_{post}, P^{i-1}), & \text{if } 1 < i \leq N \end{cases} \tag{6}$$

where P^i denotes the input visual prompts at the i-th stage of the Transformer blocks, seeded by three flows: E^{i-1}_{pre}, E^{i-1}_{post}, and P^{i-1}. This strategy enables stage-wise refinement, enhancing the model's sensitivity to visual changes induced by disasters. The VPG module undergoes three key phases: (i) projecting each flow to a lower-dimensional space for streamlined processing, (ii) enriching feature representations to highlight areas of change or damage, and (iii) merging these enhanced embeddings to form comprehensive visual prompts:

$$M^i_{pre} = g_1(E^{i-1}_{pre}) \tag{7}$$

$$M^i_{post} = g_2(E^{i-1}_{post}) \tag{8}$$

$$M^i_{prompt} = g_3(P^{i-1}) \tag{9}$$

The channel size is reduced to a lower-dimensional space (from 768 to 192 channels) using 1x1 convolutional networks, denoted as $g_1(\cdot)$, $g_2(\cdot)$, and $g_3(\cdot)$, for the pre-disaster, post-disaster, and previously generated prompt flows, respectively.

Following this, Channel Attention [17] is applied to both M^i_{pre} and M^i_{post}, focusing the model's attention on areas of interest within the images. This module dynamically recalibrates channel-wise features using global spatial information from two pipelines: average and max pooling. Specifically, each pipeline's importance is computed by processing the pooled features through a shared Multi-Layer Perceptron (MLP), followed by a sigmoid activation to obtain attention weights:

$$M^i_{att,pre} = M^i_{pre} \cdot \sigma(MLP(AvgPool(M^i_{pre})) + MLP(MaxPool(M^i_{pre}))) \tag{10}$$

$$M^i_{att,post} = M^i_{post} \cdot \sigma(MLP(AvgPool(M^i_{post})) + MLP(MaxPool(M^i_{post}))) \tag{11}$$

Where σ denotes the sigmoid function, ensuring the resulting attention weights range between 0 and 1, AvgPool and MaxPool represent the global average and max pooling operations, and MLP is the shared Multi-Layer Perceptron that models channel-wise dependencies. This attention mechanism significantly enhances the model's ability to focus on regions undergoing changes between the pre- and post-disaster states, effectively highlighting areas of damage while suppressing irrelevant background noise. The feature maps, $M^i_{att,pre}$ and $M^i_{att,post}$, are thus refined representations that emphasize critical damage indicators for subsequent processing steps.

The subsequent phase merges the processed embeddings: $M^i_{att,pre}$, $M^i_{att,post}$ and M^i_{prompt} to construct the final prompt for the next stage:

$$P^i = \begin{cases} g_4(M^i_{att,pre} + M^i_{att,post}), & \text{if } i = 1 \\ g_4(M^i_{att,pre} + M^i_{att,post} + M^i_{prompt}), & \text{if } 1 < i \leq N \end{cases} \tag{12}$$

Here, $g_4(\cdot)$ is 1x1 ConvNet to project the features back to the original dimension.

Our decoder, as outlined in Equation 5, processes only the final features, E^N_{pre} and E^N_{post}, derived from the last set of Transformer blocks. Those features are merged and subsequently upsampled through two 2x2 transpose convolution layers ($Upsample(\cdot)$) to match the resolution of the input image. Following the upsampling, a multi-scale convolutional strategy ($MSConv(\cdot)$) as proposed in [18] effectively integrates these features across different scales. The resulting segmentation map is generated by a linear layer ($Linear(\cdot)$), which classifies the level of damage for each pixel, ranging from "no damage" to "destroyed".

$$Y_{seg} = Linear(MSConv(Upsample(E^N_{pre} + E^N_{post}))) \tag{13}$$

Through the integration of multi-scale convolution and resolution reconstruction, the ViPDE decoder effectively synthesizes the processed features into a precise segmentation of building damage levels. This approach ensures that ViPDE not only leverages the power of pre-trained foundation models through the VPG but also applies advanced decoding techniques to achieve unparalleled accuracy in post-disaster image analysis.

(a) The pre-disaster image (b) The post-disaster image

Figure 5: The ground truth of pre- and post-disaster images. In (a), each polygon represents a location for a building. In (b), the green, blue, orange, and red colors represent the four different damage levels: "no damage", "minor damage", "major damage", and "destroyed", respectively.

4 Dataset and Loss Functions

In this study, we utilize the xBD satellite image dataset [8], consisting of 22,068 high-resolution images (1024x1024x3) that cover a comprehensive range of pre- and post-disaster scenarios. Several samples are displayed in Figure 2, with the top row showing pre-disaster images and the bottom row presenting the corresponding post-disaster images. These three pairs of images are sampled from three different types of disasters, i.e., tornado, flooding, and tsunami. In (a), most buildings are destroyed by the hurricane except the buildings at the bottom right corner. In (b), the building suffered major damage because it is completely surrounded by water. In (c), the two commercial buildings within the red rectangle are destroyed by the tsunami. Even though the xBD dataset owns lots of merits mentioned above, there are two main issues with this dataset, which are an imbalanced class issue and a misalignment issue between pre- and post-disaster images.

xBD is uniquely advantageous for offering paired pre- and post-disaster images for each affected location. It encompasses 11 types of natural disasters, including floods, hurricanes, and earthquakes, etc., and spans 19 cities and countries with a mix of residential and commercial buildings. Notably, it is the only dataset that categorizes damage levels ranging from "no damage" to "destroyed". These characteristics make xBD ideally suited for training models to reliably perform across diverse disaster types and building architectures. The fine-grained damage levels are exemplified in Figure 5.

4.1 Imbalanced Classes

In the xBD dataset, there are 5 classes, including "no damage", "minor damage", "major damage", "destroyed", and "unclassified". A building labeled with "no damage" class means that there is no sign of waters, structural or shingle damage, or burn marks. Buildings with "minor damage" means that these buildings are partially burnt, or in water surrounding structures, etc. A building

(a) no damage (b) minor damage

(c) major damage (d) destroyed

Figure 6: The fine-grained damage levels. (a),(b),(c), and (d) represent "no damage", "minor damage", "major damage", and "destroyed", respectively. Even by visual inspection, sometimes it is hard to differentiate buildings with "minor damage" from buildings with "major damage".

annotated as "major damage" means that there exists partial wall or roof collapse, or the building is surrounded by water or mud, etc. The "destroyed" class is for the building which is completely collapsed, covered by water or mud, etc. The "unclassified" class is the data without ground truth, which are removed from our study. Several samples are shown in Figure 6.

The distributions of these classes are shown in Figure 7. In this figure, it is evident that the "no damage" class is at least an order of magnitude larger than any other class. In this context, models trained on this original dataset will have low predictive accuracy for the infrequent classes, especially "minor damage" class and "major damage" class since they are not only minority classes, but also harder to distinguish from each other due to similar damage levels and patterns..

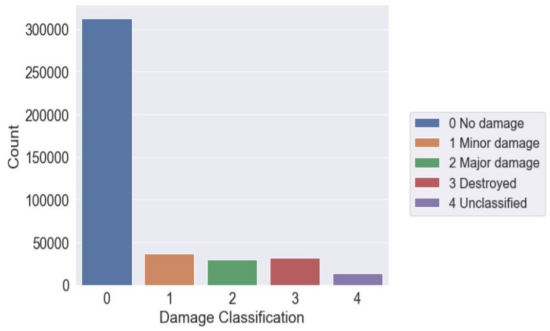
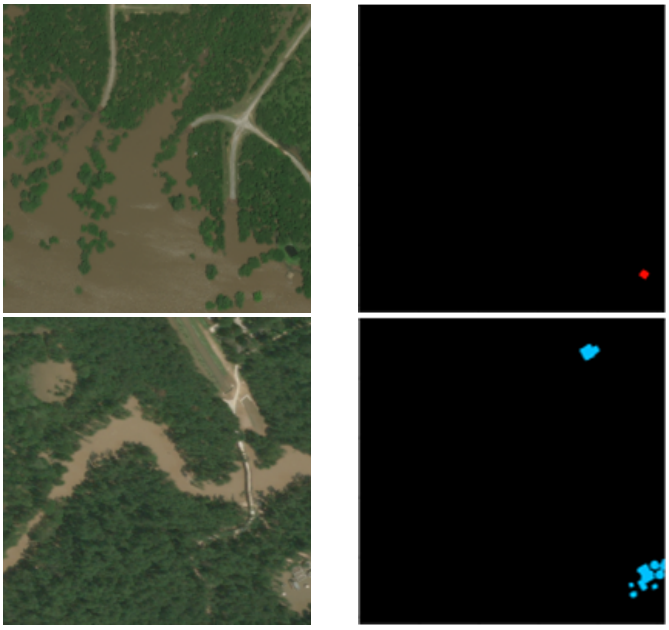


Figure 7: The imbalanced data. In this figure, the "no damage" class is the majority class.

In addition to the above issue, for each sample, i.e., each pair of pre- and post-images, there is an imbalance issue between a background class and the foreground classes, including "no damage", "minor damage", "major damage", and "destroyed". Some examples are shown in Figure 8. In this figure, there are very few pixels belonging to the buildings, compared with the number of pixels in the background. The pixels assigned to "background" are the majority in each image. In general cases, the majority class dominates the loss, leading to a rapid reduction in its error during the early stages of training. However, this can cause the model to overlook or inadequately address errors in the minority classes, resulting in suboptimal performance on these underrepresented categories.



(a) The post-disaster images (b) The Ground Truth

Figure 8: The imbalance problem between background and foreground. In this figure, the left column images are post-disaster images. The right column images are the ground truth of the corresponding post-disaster images. The black, red, and blue colors represent the pixels belonging to the "background", "destroyed", and "major damage", respectively.

4.2 Misalignment Issues

Besides the issues of the imbalanced data, there is another problem with the dataset, namely the misalignment between pre- and post-disaster images caused by shifting and tilting when the images were obtained at different points of time (pre- and post-disaster.) The most typical cases of misalignment are shown in Figure 9. In the first row of images, the building in the post-disaster image is totally outside of the polygon obtained from the projection of the polygon in the pre-disaster image. In the bottom row of images, we can easily tell there exists a misalignment due to tilting between pre- and post-disaster images. Because of these two types of misalignment issues, any hard-coded distance calculation between pre- and post-disaster images, such as L1 distance or L2 distance, is

not reliable for the damage evaluation purpose on this dataset.

(a) The pre-disaster image (b) The post-disaster image

Figure 9: The misalignment issues. In this figure, the left column images are pre-disaster images. The right column images are the post-disaster images.

4.3 Data Cleaning and Augmentation

In order to deal with the imbalanced data mentioned above, we applied data cleaning and augmentation on the original xBD dataset by three steps: in the first step, we eliminate the samples containing "unclassified" class and the samples which do not contain any foreground classes. In the second step, we split the dataset obtained by Step one into 9: 1 to form the basic training dataset and the validation dataset. In the last step, from the basic training dataset, we find the set of images containing at least one minor or major damaged building (minor-major group), and the set in which each image is dominated by major damaged or destroyed buildings (major-destroyed group). We then add one copy of each image in the minor-major group and the major-destroyed group into the training dataset. For example, if an image contains minor, major, and destroyed buildings, and is "destroyed" dominated, there will be three copies of this image including the original one in the training dataset. However, if it is minor-damage dominated, then only two copies including the original image will be in the training set. The testing set remains unaugmented, consisting of 1,866 images provided by the xBD team [8]. The size of the refined dataset is shown in Table 1.

Table 1: The dataset used for model training and testing.

	The dataset after data cleaning and augmentation
Training Dataset	12,030
Validation Dataset	640
Test Dataset	1,866

4.4 The Loss

As we explained above, there exist two main issues in the dataset: first, an imbalance problem between classes "no damage", "minor damage", "major damage", and "destroyed". Second, an imbalance issue between negative class: "background" and the four non-background classes. To address these two issues, we apply two strategies in constructing the loss functions. In the first strategy, the classes in the model are assigned with different weights based on our empirical observations. As Figure 7 shows, the "no damage" class is the majority among the four non-background classes. The size of the "no damage" class is at least one order of magnitude larger than any other four "foreground" classes. Moreover, the "background" class is the majority, compared with the "foreground" classes. Therefore, we assign different weights to these classes, which intends to place more emphasis on these minority classes and the classes that are harder to be distinguished, including the "minor damage" class and the "major damage" class. We reduce the weights for the majority classes, including the "background" class and the "no damage" class. This operation can alleviate the negative impact caused by these two imbalance issues. The weights used in our paper are shown in Table 2. In practice, there exist a lot of ensemble method-based applications that train the models with different weights and sampled the outputs of them as one output. Because of the computational cost, that line of method is not discussed in our paper.

Table 2: The weights for classes. In this table, the "no", "minor", and "major" labels represent the classes "no damage", "minor damage", and "major damage", respectively.

Classes	No	Minor	Major	Destroyed	Background
Weights	0.1	0.3	0.3	0.2	0.1

In addition to assigning a different weight to each class, we further choose a combined loss, which consists of Cross-Entropy Loss, Dice Loss, and Focal Loss. We expect the Dice Loss to alleviate the problem caused by the imbalance between a "background" class and the "foreground" classes. Moreover, because it is hard for models to distinguish between the "minor damage" class and the "major damage" class, we expect to use Focal loss to better address this issue.

The cross-entropy loss is commonly used in segmentation tasks, which can be formulated as Equation (14), where n, m, $y_{i,j}$, and $p_{i,j}$ represent the number of samples, the number of classes, the ground truth label, and the predicted probability for the sample i being class j, respectively.

$$L_{ce} = -\sum_{i}^{n}\sum_{j}^{m} y_{i,j} log(p_{i,j}) \tag{14}$$

The focal loss can be formulated as Equation (15), where $\gamma \geq 0$, and α is a weight hyper-parameter. When an example is misclassified and the corresponding $p_{i,j}$ is small, the loss is large. However, when an example is well-classified, and the corresponding $p_{i,j}$ is close to 1, the loss for this well-classified example is down-weighted [19]. This is helpful for models to distinguish the "minor damage" class from the "major damage" class.

$$L_{focal} = -\sum_{i}^{n}\sum_{j}^{m} \alpha_j (1 - p_{i,j})^{\gamma_j} y_{i,j} log(p_{i,j}) \tag{15}$$

Dice loss is one of the losses which can be used to directly optimize the segmentation metric (F1 score). It aims to minimize the mismatched regions and maximize the overlap regions between ground truth and predicted segmentation [20]. It evolves from the dice coefficient which is shown in Figure 10. In this Figure, the shadow area is the overlap of the ground truth and the prediction. The numerator consists of two times of the overlap, and the denominator is the sum of the ground truth and the prediction. Dice loss can be defined as Equation (16), where ϵ is a small positive infinitesimal quantity.

$$L_{dice} = \sum_{j}^{m}(1 - \frac{2\sum_{i}^{n} p_{i,j} y_{i,j} + \epsilon}{\sum_{i}^{n}(p_{i,j} + y_{i,j}) + \epsilon}) \tag{16}$$

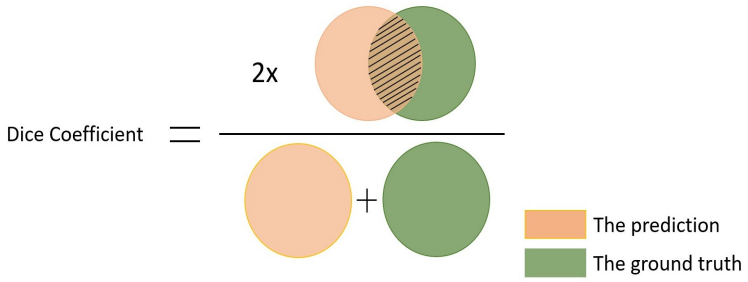


Figure 10: The Dice coefficient. In this figure, The orange color represents the prediction. The green color represents the ground truth of that prediction.

The definition of the total loss for the proposed Siamese based model is shown in Equation (17). The w_{ce}, w_f, and w_d represents the weights for the Cross-Entropy Loss, the Focal Loss, and the Dice Loss, respectively. Because the Focal loss and the Dice loss serve only auxiliary roles in improving the performance of the model on imbalanced data and hard samples, we assign a larger weight on Cross-Entropy loss. In this study, the weights for the Cross-Entropy Loss, the Focal Loss, and the Dice Loss are assigned as 0.8, 0.1, and 0.1, respectively.

$$Loss_{total} = w_{ce} L_{ce} + w_f L_{focal} + w_d L_{dice} \tag{17}$$

5 Experiments

5.1 Model Performance Comparison Experiment

Our comprehensive evaluation was conducted on the curated dataset shown in Table 1, featuring a comparative analysis among five distinct models: the xBD baseline model [8], the Siamese-based model [21], the Siamese-Attention-based model, the single-branch prompt-based model, and our novel dual-branch prompt-based model.

1. **Baseline model:** Provided by the xBD team [8], the baseline employs a two-stage approach. The first stage uses a U-Net based model for semantic segmentation to identify building locations in pre-disaster images. The second stage applies a dual-stream classification model for determining buildings' damage levels, with performance metrics reported as F1 scores.

2. **Siamese-based model:** Our previous two-stage method [21]. The first stage uses a MaskR-CNN [22] model for object detection to identify building locations in pre-disaster images. The second stage reuses the MaskRCNN's backbone with a Siamese network architecture for the building damage evaluation task.

3. **Siamese-Attention-based model:** An improvement of the previous Siamese-based model [21]. We applied an attention mechanism to enhance feature extraction. Training occurred across four NVIDIA Tesla P100 GPUs for 100 epochs, beginning with a learning rate of 0.001.

4. **Single-branch prompt-based model:** In this configuration, only post-disaster images are utilized for the building damage evaluation task. The pre-disaster flow in VPG is removed and the other parts remain untouched. This model serves as a comparative baseline, illustrating the benefits of our fully prompt-based approach. It was trained under the same conditions as the Siamese-Attention-based model, with an initial learning rate of 0.0001.

5. **Dual-branch prompt-based model:** Our proposed model utilizes a dual-branch architecture to process both pre- and post-disaster images through the SAM encoder, guided by VPG's visual prompts. The model was trained under the same setup as the single-branch prompt-based model.

Each model underwent training with a learning rate decay strategy applied to optimize performance. We further adopt the data augmentation operation during the training process. This technique increases the diversity and the amount of data available for the training process, without actually collecting any new data. In each step of the training process, we obtain variations of samples by using up-down flips, left-right flips, the rotation with 90 degrees, 180 degrees, and 270 degrees, and Gaussian noise, etc.

5.2 Misalignment Assessment Experiment

To explicitly assess how image misalignment might affect performance, we devised an experiment utilizing a basic alignment strategy. This strategy involves creating variations of each post-disaster

image by shifting it n pixels in four directions: upwards, downwards, left, and right. The newly exposed borders resulting from these shifts are filled with black pixels to ensure no additional noise is introduced into the analysis area.

By pairing each of the four shifted post-disaster image variations with the unchanged pre-disaster image, we generate four "well-aligned" sample pairs. Alongside the original unaltered pair, these five samples undergo evaluation by the model. The highest F1 score (defined in Section 4.4) from these evaluations can be used to identify the most suitable alignment for the given area, with the best-performing alignment hinting at the optimal strategy for mitigating misalignment impacts.

These strategically aligned samples are compiled into an "aligned dataset," which is then used to benchmark the performance of our duel-branch prompt-based model. This experiment aims to quantify the misalignment's detriment to building damage evaluation, potentially offering insights into corrective measures that enhance model accuracy. We tested three different values for n, including 3 pixels, 10 pixels and 15 pixels. The experimental result is presented in Section 6.3.

5.3 Metrics

F1 score is chosen as the basic component of the evaluation metric. We calculate the F1 Score for each of the four damage classes. To deal with the class imbalance problem explained in Section 2, we use the Harmonic Mean of these 4 F1 Scores as the final Score.

The four damage levels, i.e., no damage, minor damage, major damage, and destroyed, are each assigned a value 1, 2, 3, and 4, respectively.The Harmonic F1 Score is defined in Equation (18), in which $F1_i$ where $i \in [1, 4]$ represents the F1 score of the ith damage level. ϵ is a small positive infinitesimal quantity.

$$F1_{damage} = \frac{4}{\sum_{i=1}^{4} (F1_i + \epsilon)^{-1}} \tag{18}$$

6 Results and Discussions

The final harmonic F1 scores are shown in Table 3. The final score obtained by our duel-branch prompt-based model is 0.7014, which is 20 times and 22.37% better than that of the baseline model and the Siamese-based model proposed in our previous work, respectively. Moreover, the F1 score for each individual class obtained by the proposed duel-branch prompt-based model is higher than that of the other models. It is worth mentioning that the baseline model's harmonic F1 score is nearly zero, highlighting the substantial improvement achieved by our dual-branch prompt-based model with the VPG module.

6.1 Ablation Study on Visual Prompt Generator

In order to quantitatively evaluate the impact of the Visual Prompt Generator (VPG) on the performance of the ViDE model, we conducted an ablation study, comparing two configurations of the model: 1. ViDE with VPG: this configuration represents our complete proposed framework. Both VPG and the task head are trainable, allowing the model to fully utilize the capabilities of VPG in

Table 3: The F1 score of each model. In the table, the "Baseline" is the baseline model provided by xBD team. The "Siamese" is the previously proposed Siamese based model. We can see that duel-branch prompt-based model obtains the best scores in each row.

Damage Level	Baseline [8]	Siamese [21]	Siamese-Attention [21]	**Prompt (duel-branch)**
No	0.6631	0.714	0.8623	**0.8839**
Minor	0.1435	0.3955	0.4115	**0.5082**
Major	0.0094	0.6037	0.6702	**0.7241**
Destroyed	0.4657	0.7181	0.7766	**0.8180**
Mean	0.320	0.6078	0.6802	**0.7335**
Harmonic Mean	0.0342	0.5732	0.6280	**0.7014**

enhancing the processing of satellite imagery. 2. ViDE without VPG: the model relies solely on the pre-trained SAM image encoder for processing pre- and post-disaster images. Only the task head is trainable, omitting the visual prompts provided by VPG.

Table 4: $F1$ scores comparison highlighting VPG's impact

Damage Level	w/o VPG	VPG
No	0.4216	**0.8839**
Minor	0.0607	**0.5082**
Major	0.0742	**0.7241**
Destroyed	0.0123	**0.8180**
Mean	0.1422	**0.7335**
Harmonic Mean	0.0352	**0.7014**

The results in Table 4 elucidate the critical role of VPG in the ViPDE framework. The model without VPG performs 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 VPG 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 effectiveness of generated visual prompts in bridging the gap between foundation model capabilities and the specific needs of damage assessment tasks. By facilitating better integration and contextualization of satellite image data, VPG 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.

6.2 Ablation Study on the Utilization of Pre-Disaster Imagery

To assess the impact of integrating pre-disaster imagery into our contrastive learning framework, we conducted an ablation study comparing two models: the single-branch prompt-based model and the dual-branch prompt-based model. These models are described in detail in Section 5.

The results of this study in Table 5 underscore the significant benefits of employing a dual-branch contrastive learning-based architecture, which utilizes both pre- and post-disaster imagery to perform detailed semantic analysis. The dual-branch prompt-based model, by leveraging high-

Table 5: F1 scores comparison underscoring the dual-branch model's effectiveness

Damage Level	Baseline	Prompt (single-branch)	Prompt (duel-branch)
No	0.6631	0.8594	**0.8839**
Minor	0.1435	0.4299	**0.5082**
Major	0.0094	0.6583	**0.7241**
Destroyed	0.4657	0.7943	**0.8180**
Mean	0.320	0.6855	**0.7335**
Harmonic Mean	0.0342	0.6382	**0.7014**

level features extracted from pre-disaster data alongside post-disaster assessments, demonstrates superior performance in accurately classifying damage levels. This model not only outperforms the single-branch counterpart but also highlights the robustness of our contrastive learning approach in enhancing disaster damage evaluation. These findings suggest that incorporating comprehensive pre-event data into the model significantly improves its predictive accuracy and sets a new benchmark in disaster response strategies.

6.3 Misalignment Assessment

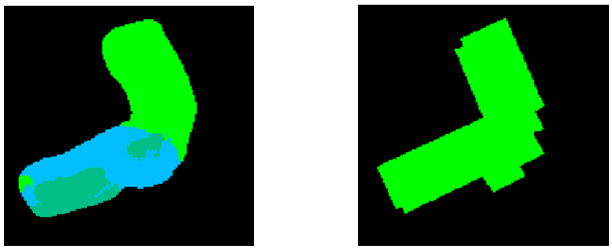
As mentioned in Section 5.2, this experiment investigates the influence of image misalignment on the performance of damage assessment models. The alignment adjustments applied to the post-disaster images were set at shifts of 3 pixels, 10 pixels, and 15 pixels to address misalignment issues. The results, as depicted in Table 6, revealed a direct correlation between alignment precision and model performance. A shift of 3 pixels yielded a positive improvement in model accuracy, while a 10-pixel shift resulted in the most significant performance boost. This enhancement indicates that a modest realignment can correct enough misalignment to positively affect model output. However, an overshoot to 15 pixels demonstrated a decline in performance, suggesting an overcompensation that reintroduced misalignment. This pattern indicates the model's sensitivity to alignment nuances, underscoring the necessity for a balanced and precise alignment methodology to optimize damage evaluation in remote sensing imagery. These findings underscore the importance of image alignment, setting a direction for future work to focus on refining pre- and post-disaster image alignment techniques.

Table 6: Misalignment assessment results. ("improve" indicates the percentage increase from the original scores.)

	3 pixels		10 pixels		15 pixels	
F1 Score	Score	Improve	Score	Improve	Score	Improve
Harmonic Mean	0.7064	0.7129%	0.7091	**1.0921%**	0.6968	-0.6601%
Mean	0.7408	0.9918%	0.7430	**1.2914%**	0.7356	0.2808%

6.4 Discussion on the Ground Truth of the Dataset

In the ground truth of the dataset, each building polygon can belong to only one single class, which means each building can only have one damage level. In our proposed models, the predictions are semantic segmentation maps, which are pixel-wise predictions, i.e., for a building, there might be different damage levels in the different parts of this building, which actually matches real-world scenarios better but does not necessarily have one single damage class label throughout the building polygon area, leading to discrepancy at the pixel level (and "errors") between the segmentation map and the ground truth. To further examine the performance of our model on the polygon area level (instead of the pixel-level), we implement a projection test to transfer the pixel-wise segmentation map to a polygon-wise map. We project the boundary of polygons from the corresponding ground truth to the prediction segmentation map. We assign each polygon to a class, which is the majority class of the pixels in the polygon. One example is shown in Figure 11. The left image is the prediction from our models. There are three classes in this building, including "no damage", "minor damage", and "major damage". Because the "no damage" class, which is green, is the majority, the whole polygon is classified as "no damage" in the end. We use our best model, duel-branch prompt-based model, for this experiment. The result is shown in Table 7. In this Table, we can see that the F1 scores for "no damage", "major damage" and "destroyed" classes are increased by at least 3.74%. The F1 score of "no damage" is improved to 0.9293 from 0.8839, which is impressive. However, the F1 score for "minor damage" slightly declined, a reflection of the inherent challenge in assessing buildings with minor yet significant enough damage. In many cases, the affected portions of a building classified as "minor damage" represent only a small fraction of the structure's total area, which can be more accurately measured by pixel-level measures as shown before. However, the majority rule applied during the projection test may override these findings, resulting in a classification of "no damage" for buildings predominantly unaffected but with noticeable minor damage areas. The current practice of assigning a single damage level to an entire building fails to



(a) The original prediction (b) The new prediction

Figure 11: A sample of the generated new segmentation map. In (a), the "no damage" class is the majority, which is green. The black, dark green, and blue represent the pixels belonging to "background", "minor damage", and "major damage", respectively.

capture the nuanced reality of disaster impact, where parts of a building may exhibit different levels of damage. This limitation not only challenges the precision of our model but also suggests a pressing need for refining ground truth datasets. A more granular approach to labeling, which acknowledges

the varying degrees of damage a single structure can sustain, would not only enhance the accuracy of models like ours but also provide a more realistic assessment of disaster effects. Adopting such detailed labeling practices would significantly improve the correspondence between model predictions and real-world conditions, ultimately leading to more effective and targeted disaster response efforts.

Table 7: F1 score comparison with and without applying a projection test, where the majority-vote approach assigns a single damage class per building polygon in the projection test. Percentage improvements over the original F1 scores are shown.

	F1 Score		
	Original	Majority Vote	Improve
No	0.8839	0.9293	5.1363%
Minor	0.5082	0.5053	-0.5706%
Major	0.7241	0.7512	3.7426%
Destroyed	0.8180	0.8486	3.7408%
Harmonic Mean	0.7014	0.7188	**2.4808%**
Mean	0.7335	0.7586	**3.4219%**

7 Conclusion

This work introduces the Visual Prompt Damage Evaluation (ViPDE), a pioneering approach that integrates prompt learning and contrastive learning with building damage evaluation to enhance post-disaster analysis through satellite imagery. By employing the Segment Anything Model (SAM) and innovatively utilizing pre- and post-disaster imagery, our approach demonstrates a significant advancement in the field of disaster management technology.

Our learnable Visual Prompt Generator (VPG) module, a cornerstone of ViPDE, showcases the utility of semantic visual prompts in directing pre-trained vision foundation models for detailed damage evaluation. This method not only streamlines the adaptation process of these models to the task of disaster damage assessment but also circumvents the limitations posed by conventional, computationally intensive data fusion techniques.

Extensive experiments validate the effectiveness of ViPDE, with our model outperforming state-of-the-art methods. The ablation study further underscores the significance of incorporating pre-disaster imagery, affirming the dual-branch model's superior performance.

ViPDE establishes a new standard in disaster management, providing a scalable, precise, and versatile tool for post-disaster damage assessment. Looking forward, we aim to expand this approach to encompass a broader range of disaster types, enhancing the model's applicability and effectiveness. Future work will also focus on developing sophisticated alignment techniques to address the misalignment issues prevalent in the dataset and on testing the adaptability of our VPG within other contrastive learning-based tasks in diverse domains.

8 Acknowledgments

Data Availability Statement

The authors thank the team responsible for the xBD dataset described in [8], which was utilized in this study. Access to this xBD dataset is restricted and requires permission from the dataset owners in [8]. Interested researchers should contact the dataset owners directly as referenced in [8] to request access and complete any necessary agreements.

References

1. Wallemacq P. Economic losses, poverty & disasters: 1998-2017. Centre for Research on the Epidemiology of Disasters, CRED, 2018.
2. Mimura N, Yasuhara K, Kawagoe S, Yokoki H, and Kazama S. Damage from the Great East Japan Earthquake and Tsunami-a quick report. Mitigation and adaptation strategies for global change 2011;16:803–18.
3. Cao QD and Choe Y. Deep learning based damage detection on post-hurricane satellite imagery. arXiv preprint arXiv:1807.01688 2018.
4. Duarte D, Nex F, Kerle N, and Vosselman G. SATELLITE IMAGE CLASSIFICATION OF BUILDING DAMAGES USING AIRBORNE AND SATELLITE IMAGE SAMPLES IN A DEEP LEARNING APPROACH. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 2018;4.
5. Brunner D, Lemoine G, and Bruzzone L. Earthquake damage assessment of buildings using VHR optical and SAR imagery. IEEE Transactions on Geoscience and Remote Sensing 2010;48:2403–20.
6. Wei D and Yang W. Detecting damaged buildings using a texture feature contribution index from post-earthquake remote sensing images. Remote Sensing Letters 2020;11:127–36.
7. Xu JZ, Lu W, Li Z, Khaitan P, and Zaytseva V. Building damage detection in satellite imagery using convolutional neural networks. arXiv preprint arXiv:1910.06444 2019.
8. Gupta R, Hosfelt R, Sajeev S, et al. xbd: A dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296 2019.
9. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems 2022;35:27730–44.
10. Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 2023.
11. Bao H, Dong L, Piao S, and Wei F. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 2021.
12. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 2020.
13. Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023:4015–26.

14. Lester B, Al-Rfou R, and Constant N. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 2021.

15. Jia M, Tang L, Chen BC, et al. Visual prompt tuning. In: *European Conference on Computer Vision*. Springer. 2022:709–27.

16. Chen S, Ge C, Tong Z, et al. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. arXiv preprint arXiv:2205.13535 2022.

17. Woo S, Park J, Lee JY, and Kweon IS. Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018:3–19.

18. Zhang Z, Wang X, Chen S, et al. MLMSA: Multi-Level and Multi-Scale Attention for Lesion Detection in Endoscopy. In: *2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom)*. IEEE. 2023:144–50.

19. Lin TY, Goyal P, Girshick R, He K, and Dollár P. Focal loss for dense object detection. In: *Proceedings of the IEEE international conference on computer vision*. 2017:2980–8.

20. Ma J. Segmentation Loss Odyssey. arXiv preprint arXiv:2005.13449 2020.

21. Zhao F and Zhang C. Building damage evaluation from satellite imagery using deep learning. In: *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE. 2020:82–9.

22. He K, Gkioxari G, Dollár P, and Girshick R. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*. 2017:2961–9.