Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images

Navjot Kaur1 | Cheng-Chun Lee2 | Ali Mostafavi2 | Ali Mahdavi-Amiri1

1School of Computing Science, Simon Fraser University, British Columbia, Canada
2Urban Resilience AI Lab, Texas A&M University, Texas, USA

Abstract
This paper presents damage assessment using a hierarchical transformer architecture (DAHiTrA), a novel deep-learning model with hierarchical transformers to classify building damages based on satellite images in the aftermath of natural disasters. Satellite imagery provides real-time and high-coverage information and offers opportunities to inform large-scale postdisaster building damage assessment, which is critical for rapid emergency response. In this work, a novel transformer-based network is proposed for assessing building damage. This network leverages hierarchical spatial features of multiple resolutions and captures the temporal differences in the feature domain after applying a transformer encoder to the spatial features. The proposed network achieves state-of-the-art performance when tested on a large-scale disaster damage data set (xBD) for building localization and damage classification, as well as on LEVIR-CD data set for change detection tasks. In addition, this work introduces a new high-resolution satellite imagery data set, Ida-BD (related to 2021 Hurricane Ida in Louisiana) for domain adaptation. Further, it demonstrates an approach of using this data set by adapting the model with limited fine-tuning and hence applying the model to newly damaged areas with scarce data.

1 INTRODUCTION

Rapid and automated damage assessment of buildings and infrastructure in the aftermath of disasters is critical to expedite emergency response and resources. Damage assessments done by ground crews can be time consuming and labor intensive (Spencer et al., 2019). The number of studies focused on damage assessments for civil infrastructures and buildings has grown recently. Several studies are dedicated to identify damage and monitor structure health (Amezquita-Sanchez et al., 2017, 2018; Li et al., 2017; Oh et al., 2017; Perez-Ramirez et al., 2019; Sirca Jr. & Adeli, 2018; Wu et al., 2019; Xu et al., 2021). In addition, recent studies demonstrate the capability of damage assessments on civil infrastructures and buildings with computer-aided approaches, including damage assessment for buildings (Gao & Mosalam, 2018; Sajedi & Liang, 2020; Zhou & Gong, 2018), bridges (Talebinejad et al., 2011; Zhang et al., 2020), concrete structures (Athanasiou et al., 2020; Li et al., 2019; Pan & Yang, 2020; Zou et al., 2022), steel structures (Kong & Li, 2018; Pan & Yang, 2023), structural bolts (Pan & Yang, 2022), and masonry structures (Wang et al., 2018). Due to the advancement of remote sensing, imagery can usually be obtained within a few days after disaster events...
(e.g., WorldView-2 [WV2] satellite has a revisit frequency of 3.7 days or less at 20 degrees off-nadir; WorldView-2, n.d.) and thus provide opportunities for assessing damage conditions more efficiently at a larger scale than manual visual inspection. To investigate expediting building damage assessments, studies have applied computer-vision techniques on building images (Wang et al., 2020; Zou et al., 2022), high-resolution aerial imagery (Cheng et al., 2021; Fujita et al., 2017; Khajewal et al., 2023), and satellite imagery (Cao & Choe, 2020; McCarthy et al., 2020; Tong et al., 2012). While aerial imagery can capture more details about buildings’ conditions due to lower flying altitudes than satellite imagery, it requires extensive local planning and provides a smaller area of building coverage than satellite imagery. Satellite imagery, which provides near real-time and high-coverage information, offers opportunities to assist in large-scale postdisaster building damage assessments (Corbane et al., 2011). Studies (Hao et al., 2021; Shen et al., 2022; Wu et al., 2021b) have shown that by leveraging satellite imagery and deep learning, the process of damage assessments can be accelerated with the generation of high-quality building footprints and damage-level classification for each building. The majority of existing models are built using a single data set (xBD) (Gupta et al., 2019) but have not been tested on data sets that include newly damaged areas. This work overcomes these limitations by creating a novel deep-learning technique to perform damage classification, building segmentation tasks, and testing of the model on a new and higher resolution data set related to building damage, Ida-BD, in Louisiana in the aftermath of Hurricane Ida.

The most common approach for building damage assessment using satellite imagery is to pose the problem as a combination of segmentation and classification tasks and to train deep-learning models on predisaster and postdisaster satellite images. Many researchers have utilized convolutional neural networks (CNNs) to achieve models with acceptable performance (Gupta & Shah, 2021; Hao et al., 2021; Wu et al., 2021b). For example, Weber and Kané (2020) considered building damage assessment as a semantic segmentation task in which damage levels are assigned to different class labels. Pre- and postdisaster image channels are concatenated into a single input, relying on the model to differentiate between the channels of pre- and postdisaster conditions, which complicates the learning process.

The problem of damage classification can be also modeled as change detection where the change is detected over a period of time with a pair of pre- and postdisaster images. Standard practice in the existing literature is to use a CNN-based network with a two-branch architecture for pre- and postdisaster images and directly concatenating the learned features to process pixel localization and classification (Gupta & Shah, 2021; Hao et al., 2021; Wu et al., 2021b).

The proposed method, building damage assessment using a hierarchical transformer architecture (DAHiTrA), argues that directly concatenating features complicates the task of localizing pixels for damage assessment; the network should concentrate on the changes between pre- and postdisaster images. Therefore, the proposed classifier works on the difference between features. Since pre- and postdisaster images are usually obtained at different times and under different lighting or weather conditions, to find meaningful and unbiased differences, both images should be mapped to a common domain. Therefore, the proposed method used difference blocks that map the features of post- and predisaster images onto a common domain to better assess the difference. In addition, as damage is typically at different scales, features should be obtained at multiresolutions. Therefore, the proposed network takes advantage of UNet-based structures to learn features at different scales and build a hierarchy of those features obtained at different scales to localize and classify the changes between two images. In Section 5, ablation studies are performed to validate the proposed design choices, to compare the proposed method with the state-of-the-art techniques, and to demonstrate that it outperforms the alternatives.

Every disaster is unique although they may share some similar characteristics. For example, the building damage caused by a hurricane may not be the same as the damage caused by another hurricane in the future. Therefore, using models based on images directly from previous disaster events to assess the damage of a new event requires the models to be tested on new data sets to evaluate the adaptation performance of the models. To address this domain shift from new disaster events, one approach could be to train the model on a subset of disaster events from xBD data set and then test on other events. However, another challenge with satellite imagery for damaged areas is the possibility of them being assessed from different cameras (for instance, a different satellite) and hence it could have different resolutions or other properties. Further, satellite imagery providers may apply different preprocessing methods. Hence, the approach of training and testing on the imagery from the same source does not handle this challenge (see Section 5.3.3). This has been a major limitation in previous studies since most of the existing models were created based on the xBD data set with limited or no adaptation performance testing. To address this limitation, a new data set, Ida-BD, is introduced with 87 image pairs (1024 × 1024) with a very high resolution (VHR; 0.5 m/pixel) from Hurricane Ida (2021) in Louisiana. Hurricane Ida brought the strongest winds ever recorded in Louisiana and heavy rains, which caused an estimated
$18$ billion in building damage and at least $30$ fatalities in Louisiana (Beven II et al., 2022). Since the near-real-time damage assessment faces the scarcity of labeled data sets, already available data sets can be used for training, and some domain adaptation techniques or fine-tuning can be applied to obtain satisfactory results on the newly damaged areas with scarce data. As a result, the Ida-BD data set can serve as a benchmark for domain adaptation from larger data sets like xBD. A baseline is provided for the domain adaptation task using a simple transfer learning method. Accordingly, the contributions of this work are as follows:

(1) A hierarchical UNet-based architecture is presented, which uses the transformer-based difference to perform well on both the damage classification and building segmentation tasks. This is the first method that explicitly relies on the difference of transformer-encoded features from predisaster and postdisaster images and hierarchically builds the output from multisolution features.

(2) A new data set, Ida-BD, is introduced for evaluating the model’s adaptation performance and providing a baseline for the damage assessment tasks. It also demonstrates the usefulness of the new data set and the efficacy of the proposed network using transfer learning.

The rest of the paper is organized as follows. The related works are first discussed in the field of building damage classification and change detection using satellite images in Section 2. In Section 3, the proposed model architecture along with loss functions and training settings are discussed. Section 4 provides an overview of the data sets used for evaluation and the results are presented in Section 5. Section 6 concludes the paper.

2 RELATED WORK

This section discusses recent methods used for building damage classification using satellite images. Since the xView2 challenge (Gupta et al., 2019) in 2019, many researchers have tried various techniques from Siamese networks (Hao et al., 2021; Wu et al., 2021b), to convolutional networks (Gupta & Shah, 2021; Weber & Kané, 2020), to attention mechanisms (Shen et al., 2022) for solving the problem of damage classification.

Earlier models, such as the one done by Hao et al. (2021), combined the UNet model with the Siamese architecture where the UNet model learns the semantic segmentation of buildings and the Siamese network focuses on the damage classification by comparing the segmented outputs. Wu et al. (2021b) further developed the model by using attention-based UNet where attention is applied to the incoming layers from the encoder module before fusing with the up-sampled features in the decoder. These architectures rely on global features learned from pre- and postdisaster images that are then merged to learn the change (here, damage) at the final stage. Since the majority of parameters prioritize learning the global features for a single image, this setting makes it difficult for the model to learn the change from the benefit of multitask supervision.

Gupta and Shah (2021) drew insights from DeepLabv3+ and came up with the RescueNet model. They used a backbone of Dilated ResNet and Atrous Spatial Pyramid Pooling module to generate multiscale features from pre- and postdisaster images and then compared their differences to learn the temporal change. Similarly to the concern with the Siamese networks, a large set of layers of this model focuses on learning the changes in images separately, which limits the learning capacity of the last few layers to make an efficient change detection. Additionally, this model uses multiscale features to capture input image variations, but these features are later simply concatenated and passed into a convolutional block for the final output. Because the change itself could be of varying resolutions, a hierarchical or multiscale analysis would be useful to generate a better quality temporal change mask.

More recently, researchers modeled the temporal relation between pre- and postdisaster images using attention mechanisms. For instance, Shen et al. (2022) introduced BDANet where they use cross-directional attention in a two-stage framework for building segmentation and damage assessment. After modeling for building segmentation using a single UNet branch, the pre- and postdisaster images are fed into two separate branches, and cross-directional attention is used to exchange information between the two parameter branches. It is noted that rather than just making the information available from the other temporal domain, one can explicitly enforce the model to learn the temporal difference by taking their absolute difference in a common feature domain.

To support temporal differences in the common domain, self-attention can be used to map the features from spatial to the feature domain, which can be efficiently applied through transformers. The transformers, although not specifically used for building damage classification tasks, have been successfully applied for building change detection on satellite images. It can be seen that the damage classification problem (multiclass classification) is similar to change detection (binary classification) by viewing the level of damage as change, although the damage classification problem is more difficult due to the higher target dimensionality and higher interclass skewness. Moreover, the damage classes are not mutually exclusive and distinct.
FIGURE 1 Comparing the model architecture of DAHiTra (ours) with two recent works for change detection, ChangeFormer (Bandara & Patel, 2022) and bitemporal image transformer (BiT) (Chen et al., 2022), for reference. (a) Difference blocks: damage assessment using a hierarchical transformer architecture (DAHiTra) and BiT use transformer encoders to map pre- and postchange features into a common domain, take the absolute difference, and then use a transformer decoder to map back the features in the spatial domain. DAHiTra uses four such blocks to get difference features of varying resolutions while BiT only uses a single resolution. In ChangeFormer, the pair of images are passed into transformer encoders and then output features are concatenated to get difference features. (b) Constructing the outputs: DAHiTra constructs hierarchical output from the four difference features; BiT up-samples the single difference feature to construct the output; ChangeFormer concatenated four difference features and passes them into multilayer perceptron stack to get the final output.

as compared to change detection where the presence of new infrastructure directly indicates change. For instance, there is no definitive way to classify roof damage as major damage instead of minor damage.

The state-of-the-art methods for the change detection problem include transformer-based models, as discussed in the following. Chen et al. (2022) introduced bitemporal image transformer (BiT) Figure 1 to model pre- and postchange contexts into visual semantic tokens using spatial attention. The transformer encoder is used to learn the temporal context between the two token sets, followed by a transformer decoder projecting the features back to the pixel space. An explicit difference is then taken on the features to generate the change mask. Although this model has significantly outperformed previous models for change detection, its performance is limited by the feature space of the visual semantic tokens. The tokens are a pair of low-dimension features generated by a CNN module, which are passed into a single-transformer encoder and decoder to learn change and reconstruct a high-dimensional output. Also, the final change prediction mask is generated by a single-step up-sampling from the very low-resolution feature space (1/4th of the final resolution), which makes it difficult to generate high-resolution change masks. In DAHiTra’s architecture, the model considers multiresolution features, which pass into different transformer encoders and decoders to learn change at different resolutions, and then instead of single-step up-sampling, hierarchical up-sampling is used on low- and high-dimensional features to construct the final output.

Another recent work for change detection is ChangeFormer by Bandara and Patel (2022), which uses hierarchically structured transformer encoder modules along with multilayer perceptron (MLP) decoders in a Siamese
network architecture Figure 1. In ChangeFormer, transformers function as feature extractors and are applied on all scales of features; however, as analyzed by Wu et al. (2021a), the transformers are inefficient at the early stages, and the convolutions are better at extracting low-level features. Additionally, the number of parameters for transformers increases four times with the size of the image, which makes the former difficult to use for large-size images as input. Therefore, our work instead uses the convolutions as the base feature extractors and applies the transformers to map these features from different temporal domains to a common domain and then projects them back using transformer decoders. Second, in the work by Bandara and Patel (2022), the output mask is based on parallel concatenation of different resolution change masks, as outputted from the transformer encoder and then passed into an MPL to generate the final output. However, in DAHiTrA, the output is constructed using pairs of transformer encoder and decoder and the change is learned explicitly in the transformer-encoded domain, which is mapped back into the spatial domain. The output is then constructed hierarchically from low-resolution to high-resolution features, which provide the model more guidance for global context and joint optimization of multiresolutions and hence produce better high-resolution results. As discussed in Section 5.2, DAHiTrA outperforms ChangeFormer and other recent works for the damage classification and change detection tasks.

Few other transformer-based change detection works include MSCANet (Liu et al., 2022) with CNN-transformer hybrid architecture with multiscale context aggregation for cropland change detection and EGCTNet (Xia et al., 2022) with a fusion encoder that combines convolutional network with transformer features. Although these networks use transformer encoders for feature encoding, they do not explicitly enforce the network to learn the temporal difference. This work is the first method that explicitly relies on the difference between transformer-encoded features from predisaster and postdisaster images and hierarchically constructs the output from multiresolution features.

3 | METHOD

This work proposes a novel neural network called DAHiTrA for multiclass change detection for the purpose of building damage assessment, which is based on UNet and transformers to learn hierarchical features of pre- and postdisaster images. The multiresolution properties of UNet are utilized to establish a hierarchical feature-based difference block that can achieve highly accurate classification and segmentation results. This is the first method that explicitly relies on the differences of transformer-encoded multiresolution features from predisaster and postdisaster images and then hierarchically builds the output from these multiresolution difference features (Figure 2). Here, DAHiTrA’s architecture and loss functions are explained. To show that DAHiTrA can be used for unseen small data sets, domain adaptation techniques have been discussed.

3.1 | Motivation

DAHiTrA is motivated by the strengths of visual transformer and UNet architecture, which are established for effectively learning context and image segmentation tasks, respectively. The UNet (Ronneberger et al., 2015) is a powerful fully convolutional architecture where the usual contracting network is supplemented by successive layers by replacing pooling operations with up-sampling operators. The contracting layers of the UNet model are made up of convolutional blocks, which are a stack of convolution, batch normalization, nonlinear activation (ReLU), and max pooling (scale of 0.5). The expanding layers of the UNet are made up of similar convolutional blocks but without maxpooling. Instead, up-sampling (scale of 2) is used to double the size of the spatial features at every layer. Hence, for an input of $1024 \times 1024$ image and model with four contracting and four expanding layers, the spatial size reduces to $64 \times 64$ at the bottleneck layer, which is then up-sampled back to generate output with the same size as input. Since the UNet computes pixel-wise output and works well with a small amount of data, it is a simple yet well-fitted network for semantic segmentation problems.

On the other hand, visual transformers (Wu et al., 2021a) extract patches from images and feed them into transformer encoders to learn relationships between different image regions and hence obtain a global representation. The transformer encoders are made up of multthead self-attention (MSA) layers. The basic idea of MSA is to apply multiple independent attention heads in parallel and then concatenate and project the output to the required mapping size through an MLP. There are three inputs to an attention head: key $K$, query $Q$, and value $V$, which are computed from the input $X \in \mathbb{R}^{L \times C}$ ($L$: token length; $C$: channels) as follows:

$$K = X W_k$$
$$Q = X W_q$$
$$V = X W_v$$

where $W_k, W_q$, and $W_v \in \mathbb{R}^{C \times d}$ and are trainable parameters. A single attention head is formulated as:

$$\text{Att}(K, Q, V) = \sigma \left( \frac{QK^T}{\sqrt{d}} \right) V$$

(1)
FIGURE 2 The model architecture for damage detection and classification. A pair of predisaster and postdisaster images is fed into an encoder made up of stacked convolution blocks (red layers), where each convolutional block downsamples the feature space (dashed lines). The output from each block of pre- and postdisaster image encoders is passed jointly (solid lines) into the difference block, which maps the spatial features into the feature domain and then takes the difference followed by a transformer decoder to map the features back into the spatial domain. Next, the output mask is hierarchically built by up-sampling (dashed lines) and concatenating the features from lower dimension to higher dimension layers. The first up-sampling block is an exception that takes in single input instead of two inputs.

The MSA concatenates multiple such attention heads:

\[ \text{MSA}(X) = \text{Concat}(\text{head}_1, ..., \text{head}_n)W_o \]  

where \( \text{head}_j = \text{Att}(K_j, Q_j, V_j) \), \( n \) is the number of attention heads, and \( W_o \in \mathbb{R}^{n_d \times C} \) is the linear projection matrix for output. To project the output back into the spatial domain of the image, transformer decoders are used. Similar to the transformer encoders, the decoder modules are also based on MSA layers, which take the features from the transformer encoder as input. The transformers are well suited for learning complex image embeddings due to their high modeling capacity, global receptive fields, and low inductive bias when compared to CNNs.

Using the insight from Wu et al. (2021a), the transformers are inefficient at the early stages and the convolutions are better at extracting low-level features. In this work, we combine the respective strengths of UNet and transformers by using the convolutions as the base feature extractors and then using the transformers to learn the temporal difference by mapping the features into a common domain. Hence, this work combines the respective strengths of UNet and transformers as we discuss below.

3.2 Model pipeline

Initially, a UNet model is trained for building segmentation tasks on the xBD data set using predisaster images while intermittently giving a few postdisaster images. This model serves as a pretrained backbone for the damage assessment task. In order to learn the difference between predisaster and postdisaster images, the UNet network is split into the encoder \( E \), which is the first half with contracting
levels, and decoder $D$, which is the second half with the expanding levels. The encoder $E$ is replicated with shared weights and the pre- and postimages are independently fed into $E_{\text{pre}}$ and $E_{\text{post}}$. Then, the features from the $i$th contracting level $E_x(i)$, where $x \in \{\text{pre}, \text{post}\}$, are passed into the difference block $Z$ (Section 3.2.1), which gives the joint difference features $z_i$ as output. The number of contracting levels is set to four. Since each contracting level is of a different resolution, we get multisresolution difference features.

$$x_i = Z_i(E_{\text{pre}}(i), E_{\text{post}}(i)) \quad i \in \{1, 4\} \quad (3)$$

Next, the decoder $D$ is used to reconstruct the pixel-wise output. The difference feature $z_i$ is passed to the $i$th expanding level of the decoder $D$, which then hierarchically builds the output mask using up-sampling blocks. Each up-sampling block takes in the output feature from the difference block and the previous up-sampling block. Both the features are concatenated channel-wise and are hence increasing the channel dimensions. The concatenation is followed by convolutional layers to avoid any artifacts due to up-sampling as well as to reduce the channel dimensionality. Next, the feature is up-sampled in spatial space to get an output feature of higher spatial dimension. This is repeated to finally get the output feature of the desired size, that is, the size of the input imagery. This leads to hierarchical development of the final damage output mask of $n$ channels where $n$ is a total number of classes and each channel contains the probability of the corresponding class. While training, a softmax function is applied to get an output whose values lie between $[0 - 1]$ and sum up to 1.

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (4)$$

While evaluation, “argmax” is applied on the final layer of $c$ channels (instead of “softmax”) to get a single channel mask whose each pixel contains the index for the most probable class $c_i$. The “argmax” function simply returns the channel index $i$ for which the probability confidence is maximum. For instance, for an input of a pair of $3 \times 1024 \times 1024$ images and a five-class classification model, the output is $1024 \times 1024$ mask, where each pixel contains 0–4 values. Here, 0 represents the background and 1–4 represent the buildings with no-damage to completely destroyed state.

### 3.2.1 Difference block

An important component of DAHiTrA is the difference block, which is made up of transformer modules. The difference block at level $i$ takes in features $E_x(i)$ where $x \in \{\text{pre, post}\}$ from the $i$th contracting level of UNet encoders $E_x$ with images from the two different temporal domains as inputs (predisaster and postdisaster). The features are independently passed into a transformer encoder $T_{di}$ with shared weights and hence encoding the features from different temporal domains to a common feature domain. An absolute difference of the encoded features is calculated and then the difference is mapped back into the original spatial domain using transformer decoder $T_{di}$.

$$Z_i(E_{\text{pre}}(i), E_{\text{post}}(i)) = T_{di}(K) \quad (5)$$

where $K = \|T_{di}(E_{\text{pre}}(i)) - T_{di}(E_{\text{post}}(i))\|$.

### 3.3 Loss function

This work used a class-weighted loss function provided in Equation (6) with a combination of focal loss (Lin et al., 2017) and dice loss (Milletari et al., 2016) to perform well for the classification task on the unbalanced data set. The focal loss $L_{\text{foc}}(i)$ is used to further address the class imbalance where a modulating term is applied to the cross-entropy loss function to focus the learning on hard misclassified examples. The second loss is dice loss $L_{\text{dice}}(i)$, which tries to maximize the overlap between predicted and ground-truth boundaries. Its denominator considers the total number of boundary pixels at a global scale, while the numerator considers the intersection, implicitly capturing local behavior. An $\alpha$ hyperparameter is used to balance the impact of focal loss and the dice loss. The class weights $w_i$ are assigned as per the pixel distribution of the classes in the data set. Hence for $n$ classes, the loss function is as follows:

$$L_{\text{dmg}}(i) = \sum_i w_i(L_{\text{foc}}(i) + \alpha L_{\text{dice}}(i)) \quad i \in \{0, ..., n\} \quad (6)$$

$$L_{\text{foc}}(i) = \sum_i -(1 - p_i)^\gamma \log(p_i) \quad i \in \{0, 1\}$$

$$L_{\text{dice}}(i) = 1 - \frac{2|\hat{c} \cup c|}{|\hat{c} \cap c|} \quad c \in \{0, 1\}$$

For instance, in the xBD data set discussed in Section 4.1, there are five classes: background (0), no damage (1), minor damage (2), major damage (3), and destroyed (4), and weights $w_i$ were assigned as per the pixel distribution of the classes in xBD data set (Table 1). On the other hand, the building change detection problem is a binary classification problem and we do not use additional weights in the loss function and rely on the focal loss to account for the class imbalance. Hence the loss function for the building change detection problem is a combination...
TABLE 1  
Class-wise pixels count distribution in xBD dataset and the corresponding weights used in training. The notation used: 0, background; 1, no damage; 2, minor damage; 3, major damage; 4, destroyed.

| Class Id | Pixels % | Weight assigned |
|----------|----------|-----------------|
| 0        | 96       | 0.01            |
| 1        | 2.7      | 0.1             |
| 2        | 0.1      | 0.7             |
| 3        | 0.1      | 0.7             |
| 4        | 0.1      | 0.7             |

of focal loss and dice loss as follows:

\[
\mathcal{L}_{chg} = \mathcal{L}_{foc}(i) + \alpha \mathcal{L}_{dice}(i) \quad i \in \{0, 1\}
\]  

3.4 | Training settings

The model is trained using Adam optimizer with a learning rate starting with a value of 1e-4, which is gradually reduced by a factor of 0.6 following a multistep learning rate scheduler. For the xBD data set, the model takes in 1024 × 1024 images as input and is trained using a batch size of 8 for 50 epochs. For the LEVIR data set, the model takes in 256 × 256 cropped images and is trained using a batch size of 16 for 200 epochs. Geometric and photometric data augmentations are heavily used to make the model robust to variations and noise in the data set. Further, small random shifts and rotation of images are used to handle any shift in pixels among the imagery pair. A 10% of the data are held for validation and the model is trained on the remaining set. The training has been done using p3.8xlarge instances of Nvidia GPUs on an AWS cloud environment and took nearly 6 h for 200 epochs on the xBD data set.

3.5 | Domain adaptation

This paper demonstrates that the model is capable of handling other data sets using a simple transfer learning approach. This also shows a path on how to make use of small data sets for other disasters including the one we have provided (Ida-BD, introduced in Section 4.2). The diverse nature of satellite imagery captured from different locations, terrains, and weather conditions makes machine learning algorithms difficult to generalize, which means that the model trained on the xBD data set does not generalize well for other data sets. Additionally, training a model on small data sets from scratch is difficult for highly complex tasks such as damage classification. Several domain adaptation algorithms have been proposed (Wang & Deng, 2018) to enable models trained on images from one data set (source) to work on images from another data set (target). Here, the Ida-BD data set is used as the target domain and since it has a domain shift from the xBD data set, this work applies the discrepancy-based method to fine-tune the network with a small labeled target data. Since the percentage of the destroyed category is relatively lower than the rest of the classes (Table 3), the destroyed class is merged with the major damage class. This is handled by replacing the output layer head with four neurons head instead of five neurons and thus framing the problem as a four-class classification (i.e., background, no damage, minor damage, major damage). As a result, the Ida-BD data set also poses the challenge of concept shift (five-class to four-class mapping) along with the domain shift (different terrain). Multiple settings are experimented with to effectively train a model on the Ida-BD data set, which include using the pretrained model directly from xBD data set on the target domain (works with unlabeled data), fine-tuning the pretrained model on the target data set (works only after some labeling), and training a model from scratch (works only after some labeling). The performance is compared in Section 5.2.3. For fine-tuning experiments, the pretrained model is trained on the Ida-BD data set for 10 epochs using a fixed learning rate of 1e-6.

4 | DATA SETS

The xBD data set (Gupta et al., 2019) was used for model training and testing for disaster damage detection and classification tasks. Later, a new data set of higher resolution (but smaller size) from Hurricane Ida is introduced for damage assessment and to explore the domain adaptation from xBD to these data. Additionally, this work also uses the LEVIR-CD data set to evaluate the model for the change detection tasks. In all these data sets, the images from pre-event and post-event are coordinate-registered by the imagery providers.

4.1 | xBD

This work uses the xBD data set (Gupta et al., 2019) for disaster damage detection and classification tasks. The xBD data set is a large-scale building damage data set that publicly offers high-resolution (0.8 m/pixel) satellite imagery with building segmentation and damage-level labels. The xBD data set provides images paired from pre- and post-disaster images collected from 19 disaster events such as floods and earthquakes with an image size of 1024 × 1024 pixels. The data set uses polygons to represent building instances and provides four damage categories—no damage, minor damage, major damage, and destroyed—for each building. As described in Table 2, the minor damage pixels represent visible roof cracks or partially burnt structures while the major damage represents partial wall, roof collapse, or structure surrounded by water. The destroyed label means that the building structure has completely
TABLE 2  Class-wise representation of damage (Gupta et al., 2019).

| Class name | Description |
|------------|-------------|
| No damage  | Undisturbed. No sign of water, structural, or burn damage. |
| Minor damage | Building partially burnt, water surrounding structure, roof elements missing, or visible cracks. |
| Major damage | Partial wall or roof collapse, water surrounded. |
| Destroyed  | Completely collapsed, completely covered with water/mud or no longer present. |

collapsed, scorched, or is no longer present. The work makes use of tier 1 and tier 3 data partitions where the labels are available. Using stratified sampling based on the damage class distribution, 10% of these data are kept aside for validation and another 10% for final testing. The rest of the data are used for training the model.

4.2  Ida-BD

A new data set, Ida-BD, obtained from the WV2 satellite with VHR images taken in November 2020 and July 2021 for predisaster and September 2021 for postdisaster is offered along with this paper. The satellite imagery was collected close to New Orleans, Louisiana, one of the most heavily impacted areas during Hurricane Ida in late August 2021 (Figure 3a). The WV2 satellite provides panchromatic images with a spectral resolution of 450–800 nm and a spatial resolution of 46 cm. The panchromatic images in this data set were first orthorectified by Apollo Mapping to 0.5 m/pixel. The work then created 87 image pairs (pairs of pre- and postdisaster images at the same location) with a size of 1024 × 1024 pixels. The resolution of these images is finer than the ones in the xBD data set. Similar to xBD, polygons were used to represent building segments and provided four damage categories. All annotations were done by the in-house team with quality control procedures using Labelbox (Labelbox, 2022). The annotation procedure is shown in Figure 3b. Building polygons were first annotated in predisaster images to avoid incorrect building boundaries due to damage. Then, by overlapping the annotation of building boundaries, building damage was classified for each building based on postdisaster images. All annotations, including building boundaries and damage levels, were reviewed several times by experts on the team. The damage levels were labeled as per the criterion introduced by the xBD data set and are specified as no damage, minor damage, major damage, and destroyed, for each building. Figure 4 shows the examples of buildings identified as destroyed, major damage, and minor damage in the Ida-BD data set.

The comparison of damage class distribution in the xBD and Ida-BD data sets is shown in Table 3. Also, the statistical summary of Ida-BD is presented in Table 4. Ida-BD shows more damaged buildings at all levels in terms of pixel counts compared with xBD except for the class of destroyed buildings. Due to the small size of the data set.
TABLE 3  Class-wise pixel percentage comparison of xBD data set with Ida-BD data set. The notation used: 0, background; 1, no damage, 2, minor damage, 3, major damage, 4, destroyed.

| Class Id | 0  | 1  | 2  | 3  | 4  |
|----------|----|----|----|----|----|
| xBD      | 96.1 | 2.7 | 0.1 | 0.1 | 0.1 |
| Ida-BD   | 81.7 | 11.9 | 4.6 | 1.6 | 0.05 |

TABLE 4  Statistical summary of the Ida-BD data set.

| Damage class | Building polygons | Pixel count | Pixel % |
|--------------|-------------------|-------------|---------|
| No damage    | 13,667            | 10,879,251  | 11.9    |
| Minor damage | 3247              | 4,214,728   | 4.6     |
| Major damage | 1120              | 1,524,746   | 1.6     |
| Destroyed    | 49                | 52,884      | 0.05    |

set, it is hard to use it directly for training a model from scratch; therefore, this work leveraged domain adaptation using a pretrained model on the xBD data set for training and evaluation on this data set. Since the xBD data set includes various disaster types, the class distribution of xBD data set is quite different from the Ida-BD data set. To address this shift in target distribution, the class weights are updated in the loss function according to the new distribution. Also, aggressive data augmentation is used for scaling and normalization to model transfer knowledge from the data set with different resolutions and red green blue distributions. We have made this data set publicly available at DesignSafe-CI (Lee et al., 2022).

4.3 | LEVIR-CD

LEVIR-CD consists of bitemporal satellite images from Texas capturing significant land-use changes such as new construction or building decline. This data set has 637 pairs of VHR (0.5 m/pixel) satellite imagery with periods of 5 to 14 years and a size of 1024 × 1024 pixels. There are nearly 31K building instances within this data set, including various building types such as warehouses, garages, and apartment complexes. This data set provides binary masks as labels with values 1 for changed and 0 for unchanged. Similar to the split for the xBD data set, this work keeps aside 10% of the data for validation and another 10% for testing while using the rest for model training.

5 | EVALUATION

In this section, the paper presents the performance of the model on the data sets listed in Section 4. The metrics used for the evaluation are shared, and the qualitative and quantitative results are presented. The impact of components of our model and loss functions are examined through ablation studies.

5.1 | Metrics

The results are evaluated on the xBD data set using the combined F1-score metrics from the XView2 Challenge 2019 (Gupta et al., 2019). The metrics include a weighted average of the F1-scores of building segmentation results and the harmonic mean of the F1-scores of building damage classification results. The higher weight is assigned to the classification score since it is a multiclass classification problem with highly skewed labels and hence more challenging to achieve.

\[
Score = 0.3 \times F_{1_{Loc}} + 0.7 \times F_{1_{Class}}
\]

\[
F_{1_{Loc}} = \frac{2 | X \cup Y |}{| X | \cap | Y |}
\]

\[
F_{1_{Class}} = \frac{1}{\frac{1}{F_{1_{i}}} + \cdots + \frac{1}{F_{1_{n}}}}
\]

Here, \(F_{1_{Loc}}\) is the F1-score for building segmentation. \(X\) and \(Y\) denote background and buildings interchangeably. \(F_{1_{i}}\) denotes class-specific F1-scores for each damage level. Since this metric heavily penalizes overfitting on the over-represented classes and given that xBD data set is heavily skewed toward the no-damage class, it is a challenging metric to use for evaluation.
5.2 | Quantitative and qualitative results

Here, the results are discussed for the three tasks: damage detection and classification, change detection, and domain adaptation tasks. This work evaluates the results for these tasks on the xBD data set (damage classification), LEVIR data set (change detection), and the Ida-BD (domain adaptation) data set.

5.2.1 | Results for damage classification

The performance of the proposed model for the damage detection and classification task on the xBD data set is evaluated by comparing the F1-score and intersection over union (IoU) metrics. The results are compared with multiple benchmarks and recent works.

1. Two-stage CNN-based models: The Siamese UNet (Ronneberger et al., 2015) is a fully convolutional and siamese-based network and generates a segmentation mask and classification mask in two stages. RescueNet (Gupta & Shah, 2021) uses a dilated ResNet and atrous spatial pyramid pooling framework to generate multiscale features and then uses independent heads for segmentation and classification tasks.

2. Fusion-based models: The Dual-HRNet (Ku et al., 2020) uses fusion blocks to exchange the features of predisaster and postdisaster images. The BDANet (Shen et al., 2022) uses cross-directional attention to exchange the features of predisaster and postdisaster images.

3. Transformer-based model: BiT (Chen et al., 2022) uses a single-transformer encoder to learn the temporal context between the predisaster and postdisaster, project the features back to the pixel space, and then generate the classification map using up-sampling.

As presented in Table 5, the proposed model outperforms the existing methods for the damage classification task. First, since the Siamese UNet (Wu et al., 2021b) is based on a comparison of pre- and postfeatures at the very end of the model layers, it does not perform well for the classification task (gain of 0.076 is observed with the proposed model). Since the RescueNet (Gupta & Shah, 2021) model also concatenates the features from pre- and postlayers at the very end of the model, it provides a comparatively low final score (gain of 0.053 is observed with the proposed model) The Dual-HRNet (Ku et al., 2020), which is one of the top-performing architectures in xview2-challenge, also provides a low score for F1 metrics, which is reasonable because it does not account for multiresolution features during comparison or feature construction (gain of 0.05 is observed with the proposed model). The BDANet (Shen et al., 2022) incorporates the attention-based exchange of information from pre and post and uses multiresolution input features and performs well for major and destroyed classes; however, the proposed model provides a gain of 0.013 in the overall score by using richer features extracted via transformers. The BiT uses a single-transformer model and uses one-step up-sampling to generate the final mask, which makes it difficult to get a good IoU score as well as an F1-score for high-resolution images. All the above discussed models use a batch size of 8 while training. As an exception, we train ChangeFormer using a batch size of 1 and compare against DAHiTrA using the same batch size of 1. The batch size is reduced due to the significantly higher computational requirements for the ChangeFormer model. As presented in Table 5, DAHiTrA outperforms ChangeFormer in segmentation as well as damage classification, which could be due to hierarchical reconstruction of the output and explicit learning of the change in transformer-encoded domain, respectively, in the former model.

The qualitative results are displayed in Figure 5, where the proposed model assigns the same damage level to almost all pixels within a building boundary, in contrast to other models. Figure 6 presents additional results from rural regions and images with poor quality. Please note that the damage classification problem is significantly difficult for images with very good quality. Further addition of noise from clouds and dust makes it more difficult to get accurate results. Though there are some performance gaps on the images with clouds (rows 3 and 4) in the precision of the building boundaries, the proposed model performs decent comparatively to the other models. A few challenges posed in the damage classification on satellite imagery that have not been addressed in this work include the damages that are not visible clearly in the images, for example, the buildings that are covered under trees before or after the disaster or the damaged regions, which are fully covered under the clouds. Additionally, the last row in Figure 6 presents zoomed in results for an imagery with moving objects like cars. All the models effectively ignore such objects and focus more on identifying building structures. This is because the xBD data set has been labeled only for the building polygons and hence the models learn to distinguish building structures from rest of the moving objects.

5.2.2 | Results for change detection

The performance of change detection is evaluated on the LEVIR data set and compared against the following methods:
TABLE 5 Average quantitative results for damage classification on the xBD dataset. Our model outperforms the state-of-the-art methods in the overall IoU and F1-score metrics as well as class-specific F1-scores for minor damage class and no-damage class. Although BDANet performs better for major damage and destroyed classes, our model has higher overall scores.

| Model      | Score | IoU  | F1-score | Class F1-scores |
|------------|-------|------|----------|-----------------|
|            |       |      |          | No damage | Minor damage | Major damage | Destroyed |
| Siam-UNet  | 0.743 | 0.824| 0.709    | 0.955     | 0.576       | 0.744        | 0.662      |
| RescueNet  | 0.766 | 0.840| 0.735    | 0.883     | 0.563       | 0.771        | 0.808      |
| Dual-HRNet | 0.769 | 0.834| 0.741    | 0.898     | 0.590       | 0.737        | 0.809      |
| BDANet     | 0.806 | 0.864| 0.782    | 0.925     | 0.616       | 0.788        | 0.876      |
| Bit        | 0.764 | 0.816| 0.742    | 0.971     | 0.631       | 0.723        | 0.719      |
| Ours       | 0.819 | 0.872| 0.796    | 0.978     | 0.711       | 0.765        | 0.772      |
| ChangeFormer* | 0.458 | 0.678| 0.364    | 0.778     | 0.389       | 0.493        | 0.196      |
| Ours*      | 0.495 | 0.734| 0.393    | 0.725     | 0.412       | 0.487        | 0.232      |

*Here, the models are trained with batch size of 1 due to GPU compute constraints for ChangeFormer. The rest of the models are trained with a batch size of 8.

Bold terms are indicating the best performance among all models.

TABLE 6 Average quantitative results for change detection on the LEVIR-CD dataset. Our model performs significantly better than the state-of-the-art methods in overall F1-score and IoU as well as class-specific F1-scores.

| Model       | IoU  | F1-score | Class F1-scores |
|-------------|------|----------|-----------------|
|             |      |          | 0 class | 1 class |
| Siam-UNet   | 0.813| 0.859    | 0.987   | 0.788  |
| BiT         | 0.820| 0.899    | 0.990   | 0.807  |
| Changeformer| 0.828| 0.898    | 0.990   | 0.806  |
| Ours + with conv | 0.833| 0.901    | 0.991   | 0.812  |
| Ours        | **0.842** | **0.908** | **0.991** | **0.825** |

Bold terms are indicating the best performance among all models.

1. CNN-based model: The Siamese UNet (Ronneberger et al., 2015) is a fully convolutional and siamese-based network and is used as a simple baseline for the comparison.

2. Transformer-based models: BiT (Chen et al., 2022) uses a single-transformer encoder to learn the temporal context between the predisaster and postdisaster, project the features back to the pixel space and then generate the classification map using up-sampling. Changeformer (Bandara & Patel, 2022) uses hierarchically structured transformer encoder modules along with MLP decoders.

The proposed model performs significantly better in both overall F1-score and IoU metrics, as demonstrated in Table 6. Since the BiT model uses a single-transformer encoder and decoder pair and then generates the output mask by up-sampling the decoder features, it is difficult to get a high-resolution pixel-level classification mask. On the other hand, Changeformer uses hierarchically structured transformer encoder modules but the features are not projected back into the spatial domain using the transformer decoders. Instead, the proposed network explicitly takes the difference in the transformer-encoded feature domain and then maps back the features into the spatial domain to reconstruct the final output. This leads to a better performing framework for change detection. In addition, adding a convolution layer after every merge of outputs from the transformer decoder and lower layer provides smoother results and fewer artifacts. The accuracy achieved under this setting is slightly lower than our final model, though higher than other recent works. The qualitative results are displayed in Figure 7, where it can be seen that the model produces finer building boundaries and captures a few additional buildings that had been missed by other models.

5.2.3 Results for domain adaptation

The purpose of this section is to demonstrate a simple approach to using the model on new and small data sets like Ida-BD. Please note that the model is not necessarily better than the rest of the models for this task and the authors only present one of the methods for transferring learning on new data sets. To evaluate the model adaptation performance on the Ida-BD data set, this work experiments with three training settings for our model using a subset of Ida-BD as test data. Siamese UNet is used as a baseline model for reference on performance. For both models, in the first experiment, the model is trained on Ida-BD data from scratch, assuming all the new imagery is labeled. In the second setting, the model is trained on only xBD (without any fine-tuning on the Ida-BD), assuming that there are no labels available for the new data set. In the last experiment, the model is pretrained on xBD data and...
FIGURE 5  Qualitative results for damage classification (evaluation on the xBD data set). As highlighted in the boundary boxes, our model assigns a more accurate damage level to almost all pixels within a building boundary as compared to Siamese UNet and Dual-HRNet.

TABLE 7  Domain adaptation on the Ida-BD data set. Three different settings are tested on Ida-BD data: (1) model trained only on Ida-BD data (source and target) from scratch; (2) model trained on only xBD data without any fine-tuning; (3) model pretrained on xBD data (source) and fine-tuned on Ida-BD data (target). The last setting, that is, pretraining on xBD and fine-tuning on the new data set, performs the best for both models.

| Model | Pretraining | Training | IoU  | Class F1-scores | F1-score | No damage | Minor damage | Major damage |
|-------|-------------|----------|------|-----------------|----------|-----------|--------------|--------------|
|       |             |          |      |                 |          |           |              |              |
| Siam-UNet | -           | Ida      | 0.697| F1-score        | 0.472    | 0.906     | 0.313        | 0.483        |
|         | xBD         | -        | 0.584| 0.047           | 0.921    |           | 0.409        | 0.151        |
|         | xBD         | Ida      | 0.748| 0.507           | 0.846    |           | 0.322        | 0.609        |
| Ours   | -           | Ida      | 0.778| 0.541           | 0.916    |           | 0.384        | 0.538        |
|         | xBD         | -        | 0.674| 0.023           | 0.921    |           | 0.432        | 0.071        |
|         | xBD         | Ida      | 0.805| 0.585           | 0.910    |           | 0.439        | 0.577        |

Bold terms are indicating the best performance among all models.

then fine-tuned using a subset of Ida-BD imagery, assuming that the new imagery is partially labeled. Finally, as shown in Table 7 for both the models, the best performance was obtained by pretraining the model on xBD data and then fine-tuning it on the Ida-BD data set. This shows that to use models for small data sets like Ida-BD, pretraining on the large-scale xBD data set provides a good prior, and hence models are able to better generalize to new data sets. The qualitative results are displayed in Figure 8.

5.2.4  | Limitations

This work has used a simple domain adaptation approach for the Ida-BD data set. Further experiments can be done
FIGURE 6  Additional figures for qualitative results for damage classification (evaluation on the xBD data set), containing images from hard terrains, images with poor quality, and rural regions. As observed in the figure, the model performs well on these images as compared to other baselines. The last row contains zoomed in results from imagery with moving objects like vehicles. The models effectively ignore such objects while identifying the building structures.
using the domain adaptation methods like knowledge distillation (Hinton et al., 2015), few-shot learning (Guo et al., 2020), and adversarial-based domain adaptation using generative models generative adversarial networks (GANs) (Tzeng et al., 2017).

5.3 Ablation studies

In this subsection, the paper discusses a few experiments on the performance and impact of distortions and different loss functions on our model.

5.3.1 Impact of distortion

The pre-event and post-event images from the xBD, LEVIR-CD, and Ida-BD data sets are coordinate-registered by the imagery providers and hence they are perfectly aligned. However, the impact of slight translation or angular deviation is discussed by training on images with random translational distortion and angular deviation. As presented in Table 8, training the model on images randomly translated up to 1 m gives a small decrease of 0.008 in F1-score and 0.003 in IoU metrics. Similarly, angular deviation until 0.1 degrees gives a very small decrement of 0.004 in F1-score and 0.004 in IoU metrics. This could be explained using the property of transformers that it learns global features and is not limited to the local context. However, further distortions up to 2 m or angular deviation of 1-degree lead to F1-score decrement of 0.03 and 0.05, respectively.

5.3.2 Computational analysis

The computational cost and inference speed analysis of the network is compared with the recent models, using the images from xBD data set. As seen in Table 9, the transformer-based models, BiT and DAHiTrA, have a lower number of parameters than CNN models. The inference speed of Siam-UNet is the lowest since it is fully convolutional and can be parallelized more efficiently.
FIGURE 8 Qualitative results for domain adaptation (evaluation on the Ida-BD data set). Our model assigned significantly more accurate damage classes than Siamese UNet. Please compare the highlighted boxes as a reference.

TABLE 9 Computational analysis of the network.

| Model      | Params (M) | Inference (i/s) | FLOPs (G) |
|------------|------------|-----------------|-----------|
| Siam-UNet  | 25.7       | 0.620           | 26.37     |
| BDANet     | 34.5       | 0.681           | 82.22     |
| BiT        | 12.4       | 0.636           | 11.52     |
| ChangeFormer | 41.1     | 2.832           | 209.41    |
| Ours       | 13.2       | 0.625           | 11.95     |

However, the transformer-based models are faster than fusion-based models including BDANet and Dual-HRNet.

5.3.3 Data set choice for domain adaptation

As discussed in the introduction, adapting the model to new imagery is challenging and there are two different alternatives for choosing a data set to study domain adaptation. One method is to use imagery from one of the disaster events in the xBD data set as test data while training on the remaining imagery. Another method is to use data sets like Ida-BD as a test set, which is from a different imagery source and different resolution, and train on the xBD data set. In this experiment, we first assumed that no labeled data are available for the “new” disaster event and compared the models with no fine-tuning on the “new” disaster class. Next, we have compared the performance by assuming that 30% of imagery for the “new” disaster event is labeled and hence the model is fine-tuned using 30% of imagery from the “new” disaster class. As seen in Table 10, in the absence of the training data for fine-tuning, the model does not work well in both cases but a small amount of fine-tuning improves the performance significantly for both the data sets. The test performance for
TABLE 10 Performance of DAHiTrA for different choices of the dataset for domain adaptation study. Here, xBD-T stands for xBD data set except for Hurricane Matthew imagery and xBD-V stands for xBD data set except for Hurricane Matthew imagery, along with 30% imagery of test class.

| Test              | Train  | IoU   | F1-score |
|-------------------|--------|-------|----------|
| Hurr. Matthew (xBD) | xBD-T  | 0.553 | 0.191    |
| Ida-BD            | xBD-T  | 0.628 | 0.026    |
| Hurr. Matthew (xBD) | xBD-V  | 0.720 | 0.724    |
| Ida-BD            | xBD-V  | 0.773 | 0.578    |

TABLE 11 Impact of loss functions (xBD data).

| Loss function                  | IoU   | F1-score |
|--------------------------------|-------|----------|
| Focal + Dice                   | 0.816 | 0.796    |
| Focal + Dice + Ordinal         | 0.807 | 0.791    |
| Buildings only Cross-entropy   | 0.798 | 0.803    |

Bold terms are indicating the best performance among all comparisons.

5.3.4 Loss functions

This work experimented with different combinations of loss functions (Table 11). The localization-aware loss (Gupta & Shah, 2021) gave better F1-scores for damage classes where pixels are accounted for in the cross-entropy loss calculation only in the presence of buildings. The ordinal loss is also experimented with in the form of mean square error between the ground truth and predicted output for buildings only by converting the values from a scale of 1 to 4 to a scale of 0 to 1. This loss did not work well for our task, which could be because of the highly imbalanced and sparse class distribution. The weighted sum of focal and dice loss provided the best performance in class-wise F1-scores as well as building boundary precision since the former explicitly addresses the class imbalance and the latter ensures crisp building boundaries.

5.3.5 Transformer layers

This work experimented with adding different numbers of transformer blocks between the encoder and decoder; three iterations gave us the best results. It can be observed that adding transformer blocks at all levels, except the highest resolution layer, provided an incremental improvement in accuracy. An additional transformer block in the final layer does not perform well, which could be because the input dimension is very high and it becomes hard for transformers to learn in this space, as discussed in Wu et al. (2021a). Hence, this work used only the convolutional layer in the highest resolution layer. On the other hand, if the transformer is not used in any of the layers, the performance highly degrades, indicating the importance of transformer-encoded features. The test result values from the LEVIR-CD data set are shared in Table 12. Also, using a combination of transformer encoder and decoder along with the difference module, instead of just transformer encoder followed by difference module, provided a bump of accuracy by 0.8% for the xBD data set.

TABLE 12 Effect of the number of transformer layers used (LEVIR data).

| Number | IoU     | F1-score |
|--------|---------|----------|
| 0      | 0.818   | 0.843    |
| 1      | 0.837   | 0.897    |
| 2      | 0.843   | 0.903    |
| 3      | 0.842   | 0.908    |
| 4      | 0.838   | 0.901    |

Bold terms are indicating the best performance among all comparisons.

6 CONCLUDING REMARKS

This paper presents a hierarchical UNet-based architecture that uses transformer-based differences to perform well on both the damage classification and building segmentation tasks. This is the first method that explicitly relies on the difference between transformer-encoded features from predisaster and postdisaster images and hierarchically builds the output from multiresolution features by using UNet as a backbone. The proposed method yields state-of-the-art results on the xBD data set and LEVIR-CD data set for damage classification and change detection tasks, respectively. Additionally, this work introduces a new data set, Ida-BD, for evaluating the model’s adaptation performance and providing a baseline for the damage assessment tasks. It also demonstrates the usefulness of the new data set and the efficacy of the proposed network using transfer learning. The domain adaptation results indicate that the model can be adapted to a new event with slight fine-tuning, which is critical for the application of the model for future events. In addition to application to building damage assessment, the proposed network architecture with hierarchical transformers can be used in
other civil infrastructure and urban systems applications; for example, this model architecture is expected to perform well on problems such as road damage classification, structure change detection, and urban land cover change classification problems.

In future work, the authors plan to improve the building boundaries using either GAN loss or exponential boundary loss, which could help to split different building instances. The work can be further extended by using dynamic algorithms like neural dynamic classification (Rafiei & Adeli, 2017) and dynamic ensemble learning (Alam et al., 2020) and can be optimized for faster learning using finite element machines (Pereira et al., 2020).

ACKNOWLEDGMENTS
The authors would like to acknowledge funding support from the Texas A&M X-Grant Presidential Excellence Fund.

CODE AVAILABILITY
The code of the proposed model in this study is available at https://github.com/nka77/DAHiTra.

DATA AVAILABILITY STATEMENT
xBD data set is publicly available at https://xview2.org/dataset. LEVIR-CD data set can be acquired from https://justchennao.github.io/LEVIR/. Ida-BD is publicly available on DesignSafe-CI at https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRI-3563.

REFERENCES
Alam, K. M. R., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. Neural Computers and Applications, 32(12), 8675–8690. https://doi.org/10.1007/s00521-019-04359-7
Amezquita-Sanchez, J. P., Park, H. S., & Adeli, H. (2017). A novel methodology for modal parameter identification of large smart structures using MUSIC, empirical wavelet transform, and Hilbert transform. Engineering Structures, 147, 148–159. https://www.sciencedirect.com/science/article/pii/S0141029617300445
Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., & Adeli, H. (2018). Wireless smart sensors for monitoring the health condition of civil infrastructure. Scientia Iranica, 25(6), 2913–2925. http://scientiairanica.sharif.edu/article_21136.html
Athanasiou, A., Ebrahimmghanlou, A., Zaborac, J., Hrynyk, T., & Salamone, S. (2020). A machine learning approach based on multifractal features for crack assessment of reinforced concrete shells. Computer-Aided Civil and Infrastructure Engineering, 35(6), 565–578. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12509
Bandara, W. G. C., & Patel, V. M. (2022). A transformer-based siamese network for change detection. https://arxiv.org/abs/2201.01293
Beven II, J. L., Hagen, A., & Berg, R. (2022). National hurricane center tropical cyclone report—Hurricane Ida. Technical Reports AL092021, Nationa Hurricane Center. https://www.nhc.noaa.gov/data/tcr/AL092021_Ida.pdf
Cao, Q. D., & Choe, Y. (2020). Building damage annotation on post-hurricane satellite imagery based on convolutional neural networks. Natural Hazards, 103(3), 3357–3376.
Chen, H., Qi, Z., & Shi, Z. (2022). Remote sensing image change detection with transformers. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14.
Cheng, C.-S., Behzadan, A. H., & Noshadravan, A. (2021). Deep learning for post-hurricane aerial damage assessment of buildings. Computer-Aided Civil and Infrastructure Engineering, 36, 695–710. https://doi.org/10.1111/mice.12658
Corbane, C., Cartron, D., Lemoine, G., & Broglia, M. (2011). Comparison of damage assessment maps derived from very high spatial resolution satellite and aerial imagery produced for the Haiti 2010 earthquake. Earthquake Spectra, 27(suppl I), 199–218.
Fujita, A., Sakurada, K., Imaizumi, T., Ito, R., Hikosaka, S., & Nakamura, R. (2017). Damage detection from aerial images via convolutional neural networks. In 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan (pp. 5–8). IEEE. https://doi.org/10.23919/MVA.2017.7986759
Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image-based structural damage recognition. Computer-Aided Civil and Infrastructure Engineering, 33(9), 748–768. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12363
Guo, Y.,Codecella, N. C.,Karlinsky, L.,Codella, J. V., Smith, J. R., Saenko, K., Rosing, T., & Feris, R. (2020). A broader study of cross-domain few-shot learning. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.). Computer vision – ECCV 2020 (pp. 124–141). Springer International Publishing.
Gupta, R., Hosfelt, R., Sajeev, S., Patel, N., Goodman, B., Doshi, J., Heim, E. T., Choset, H., & Gaston, M. E. (2019). XB: A dataset for assessing building damage from satellite imagery. arXiv preprint arXiv:1911.09296.
Gupta, R., & Shah, M. (2021). RescueNet: Joint building segmentation and damage assessment from satellite imagery. In 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy (pp. 4405–4411). IEEE. https://doi.org/10.1109/ICPR48806.2021.9412295
Hao, H., Bairreddy, S., Bartusiak, E. R., Konz, L., LaTourrette, K., Gribbons, M., Chan, M., Delp, E. J., & Comer, M. L. (2021). An attention-based system for damage assessment using satellite imagery. In 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium (pp. 4396–4399). IEEE. https://doi.org/10.1109/IGARSS47720.2021.954054
Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. https://arxiv.org/abs/1503.02531
Khajwali, A. B., Cheng, C.-S., & Noshadravan, A. (2023). Post-disaster damage classification based on deep multi-view image fusion. Computer-Aided Civil and Infrastructure Engineering, 38, 528–544. https://doi.org/10.1111/mice.12890
Kong, X., & Li, J. (2018). Vision-based fatigue crack detection of steel structures using video feature tracking. Computer-Aided Civil and Infrastructure Engineering, 33(9), 783–799. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12535
Ku, J., Seo, J., & Jeon, T. (2020). Dualhrnet for building localization and damage classification. https://github.com/SIAAnalytics/dualhrnet/blob/master/figures/xView2_White_Paper_SI_Analytics.pdf
Labelbox (2022). Labelbox. https://labelbox.com
Lee, C.-C., Kaur, N., Mahdavi-Amiri, A., & Mostafavi, A. (2022). Ida-BD: Pre- and post-disaster high-resolution satellite imagery for building damage assessment from Hurricane Ida. https://www.designsafe-ci.org/data/browser/public/designsafe.
storage.published/PRJ-3563
Li, S., Zhao, X., & Zhou, G. (2019). Automatic pixel-level multiple damage detection of concrete structure using fully convolutional network. Computer-Aided Civil and Infrastructure Engineering, 34(7), 616–634. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12433
Li, Z., Park, H. S., & Adeli, H. (2017). New method for modal identification of super high-rise building structures using discretized synchrosqueezed wavelet and Hilbert transforms. The Structural Design of Tall and Special Buildings, 26(3), e1312. https://onlinelibrary.wiley.com/doi/abs/10.2014/tal.1312
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2017). Focal loss. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2980–2988). IEEE.
Liu, M., Chai, Z., Deng, H., & Liu, R. (2022). A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 4297–4306.
McCarthy, M. J., Jessen, B., Barry, M. J., Figueura, M., McIntosh, J., Murray, T., Schmid, J., & Muller-Karger, F. E. (2020). Mapping hurricane damage: A comparative analysis of satellite monitoring methods. International Journal of Applied Earth Observation and Geoinformation, 91, 102134.
Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 fourth international conference on 3D vision (3DV) (pp. 565–571). IEEE.
Oh, B. K., Kim, K. J., Kim, Y., Park, H. S., & Adeli, H. (2017). Evolutionary learning based sustainable strain sensing model for structural health monitoring of high-rise buildings. Applied Soft Computing, 58, 576–585. https://www.sciencedirect.com/science/article/pii/S1568494617302922
Pan, X., & Yang, T. Y. (2020). Postdisaster image-based damage detection and repair cost estimation of reinforced concrete buildings using dual convolutional neural networks. Computer-Aided Civil and Infrastructure Engineering, 35(5), 495–510. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12549
Pan, X., & Yang, T. Y. (2023). 3D vision-based out-of-plane displacement quantification for steel plate structures using structure-from-motion, deep learning, and point-cloud processing. Computer-Aided Civil and Infrastructure Engineering, 38, 547–561. https://doi.org/10.1111/mice.12906
Pan, X., & Yang, T. Y. (2022). Image-based monitoring of bolt loosening through deep-learning-based integrated detection and tracking. Computer-Aided Civil and Infrastructure Engineering, 37(10), 1207–1222. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12797
Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. a. P., & Adeli, H. (2020). FEMA: A finite element machine for fast learning. Neural Computers and Applications, 32(10), 6393–6404. https://doi.org/10.1007/s00521-019-04146-4
Perez-Ramirez, C. A., Amezquita-Sanchez, J. P., Valtierra-Rodriguez, M., Adeli, H., Dominguez-Gonzalez, A., & Romero-Troncoso, R. J. (2019). Recurrent neural network model with Bayesian training and mutual information for response prediction of large buildings. Engineering Structures, 178, 603–615. https://www.sciencedirect.com/science/article/pii/S0141029618307235
Rafiei, M. H., & Adeli, H. (2017). A new neural dynamic classification algorithm. IEEE Transactions on Neural Networks and Learning Systems, 28(12), 3074–3083.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III (pp. 234–241). Springer International Publishing.
Sajedi, S. O., & Liang, X. (2020). Vibration-based semantic damage segmentation for large-scale structural health monitoring. Computer-Aided Civil and Infrastructure Engineering, 35(6), 579–596. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12523
Sircar, G. F. Jr., & Adeli, H. (2018). Infrared thermography for detecting defects in concrete structures. Journal of Civil Engineering and Management, 24(7), 508–515. https://journals.vilniustech.lt/index.php/ICEM/article/view/6186
Shen, Y., Zhu, S., Yang, T., Chen, C., Pan, D., Chen, J., Xiao, L., & Du, Q. (2022). Bdanet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images. IEEE Transactions on Geoscience and Remote Sensing, 60, 1–14.
Spencer, B. F., Hoskere, V., & Narazaki, Y. (2019). Advances in Computer Vision-Based Civil Infrastructure Inspection and Monitoring. Engineering, 5(2), 199–222.
Talebinejad, I., Fischer, C., & Ansari, F. (2011). Numerical evaluation of vibration-based methods for damage assessment of cable-stayed bridges. Computer-Aided Civil and Infrastructure Engineering, 26(3), 239–251. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8667.2010.00684.x
Tong, X., Hong, Z., Liu, S., Zhang, X., Xie, H., Li, Z., Yang, S., Wang, W., & Bao, F. (2012). Building damage detection using pre- and post-seismic high-resolution satellite stereo imagery: A case study of the May 2008 Wenchuan earthquake. ISPRS Journal of Photogrammetry and Remote Sensing, 68, 13–27.
Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) (pp. 7167–7176). IEEE.
Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. Neurocomputing, 312, 135–153. https://www.sciencedirect.com/science/article/pii/S0925231218306684
Wang, N., Zhao, Q., Li, S., Zhao, X., & Zhao, P. (2018). Damage segmentation and measurement of glazed tiles in historic buildings via deep learning. Computer-Aided Civil and Infrastructure Engineering, 35(3), 277–291. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12488
Weber, E., & Kané, H. (2020). Building disaster damage assessment in satellite imagery with multi-temporal fusion. arXiv preprint arXiv:2004.05525.
WorldView-2. (n.d.). https://resources.maxar.com/data-sheets/worldview-2

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J. E., Keutzer, K., & Vajda, P. (2021a). Visual transformers: Where do transformers really belong in vision models? In Proceedings of the IEEE/CVF international conference on computer vision (ICCV) (pp. 599–609). IEEE.

Wu, C., Zhang, F., Xia, J., Xu, Y., Li, G., Xie, J., Du, Z., & Liu, R. (2021b). Building damage detection using u-net with attention mechanism from pre- and post-disaster remote sensing datasets. Remote Sensing, 13(5), 905. https://doi.org/10.3390/rs13050905

Wu, R.-T., Singla, A., Jahanshahi, M. R., Bertino, E., Ko, B. J., & Verma, D. (2019). Pruning deep convolutional neural networks for efficient edge computing in condition assessment of infrastructures. Computer-Aided Civil and Infrastructure Engineering, 34(9), 774–789. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12449

Xia, L., Chen, J., Luo, J., Zhang, J., Yang, D., & Shen, Z. (2022). Building change detection based on an edge-guided convolutional neural network combined with a transformer. Remote Sensing, 14(18), 4524. https://doi.org/10.3390/rs14184524

Xu, Y., Lu, X., Cetiner, B., & Taciroglu, E. (2021). Real-time regional seismic damage assessment framework based on long short-term memory neural network. Computer-Aided Civil and Infrastructure Engineering, 36(4), 504–521. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12628

Zhang, C., Chang, C.-c., & Jamshidi, M. (2020). Concrete bridge surface damage detection using a single-stage detector. Computer-Aided Civil and Infrastructure Engineering, 35(4), 389–409. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12500

Zhou, Z., & Gong, J. (2018). Automated analysis of mobile LiDAR data for component-level damage assessment of building structures during large coastal storm events. Computer-Aided Civil and Infrastructure Engineering, 33(5), 373–392. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12345

Zou, D., Zhang, M., Bai, Z., Liu, T., Zhou, A., Wang, X., Cui, W., & Zhang, S. (2022). Multicategory damage detection and safety assessment of post-earthquake reinforced concrete structures using deep learning. Computer-Aided Civil and Infrastructure Engineering, 37(9), 1188–1204. https://onlinelibrary.wiley.com/doi/abs/10.1111/mice.12815

How to cite this article: Kaur, N., Lee, C.-C., Mostafavi, A., & Mahdavi-Amiri, A. (2023). Large-scale building damage assessment using a novel hierarchical transformer architecture on satellite images. Computer-Aided Civil and Infrastructure Engineering, 38, 2072–2091. https://doi.org/10.1111/mice.12981