Bitemporal Attention Transformer for Building Change Detection and Building Damage Assessment

Wen Lu, Lu Wei, and Minh Nguyen

***Abstract*—Building change detection (BCD) holds significant value in the context of monitoring land use, whereas building damage assessment (BDA) plays a crucial role in expediting humanitarian rescue efforts post-disasters. To address these needs, we propose the bitemporal attention module (BAM) as an innovative cross-attention mechanism aimed at effectively capturing spatio-temporal semantic relations between a pair of bitemporal remote sensing images. Within BAM, a shifted windowing scheme has been implemented to confine the scope of the cross-attention mechanism to a specific range, not only excluding remote and irrelevant information but also contributing to computational efficiency. Moreover, existing methods for BDA often overlook the inherent order of ordinal labels, treating the BDA task simplistically as a multiclass semantic segmentation problem. Recognizing the vital significance of ordinal relationships, we approach the BDA task as an ordinal regression problem. To address this, we introduce a rank-consistent ordinal regression loss function to train our proposed change detection network, bitemporal attention transformer. Our method achieves state-of-the-art accuracy on two BCD datasets (LEVIR-CD+ and S2Looking), as well as the largest BDA dataset (xBD).**

***Index Terms*—Building change detection (BCD), building damage assessment (BDA), ordinal regression, transformer.**

I. Introduction

BOTH building change detection (BCD) and building damage assessment (BDA) are subtasks of change detection. BCD aims at identifying structural alterations of buildings over time, it involves allocating binary labels (changed or unchanged) on a pixel level through the analysis of aligned images acquired at different moments. BCD finds applications in urban planning [1], land-cover monitoring [2], [3], and other fields where tracking changes in built structures is crucial for informed decision-making. BDA can be viewed as a multiclass change detection (MCD) task specifically concentrating on the land cover category "building." In this context, it identifies individual buildings and assigns predefined damage degree labels to them. Timely humanitarian assistance and disaster response, especially within the first 72 h, is very crucial for saving lives [4]. By locating and evaluating the damage severity of the buildings, BDA provides critical information for emergency responders to identify damaged zones, plan aid routing, and optimize the deployment of rescue resources within impacted regions. High spatial resolution satellite and aerial imagery can accurately reflect the Earth's surface and rapidly provide large area observations for BCD and BDA tasks. However, analysis of the imagery by experts is laborious and time-consuming, therefore, automatic BCD and BDA are imperative.

It is worth noting that BDA differs from semantic change detection (SCD), which broadens the MCD task by offering not just the locations of changes, but also detailed land cover and land use (LCLU) categories before and after the observation periods. The typical predefined labels for BDA include *no damage*, *minor damage*, *major damage*, *destroyed*, and *background*. Consequently, the predefined LCLU categories consist of only two: "building" and "other." Whereas, the predefined LCLU categories are more diverse for SCD. For example, the SEmantic Change detectiON Dataset (SECOND) [5] includes LCLU categories: "non-vegetated ground surface," "tree," "low vegetation," "water," "buildings," and "playgrounds." The Landsat-SCD dataset [6] includes LCLU categories: "farmland," "desert," "building," and "water." In BDA task, the LCLU categories remain constant, as the primary focus is on evaluating the severity of building damage. Conversely, SCD is designed to identify alterations in LCLU categories.

In contrast to the categorical labels in tasks such as semantic segmentation, land cover classification, and SCD, the labels in the BDA exhibit an ordinal relationship. The joint damage scale [7], developed in collaboration with experts from NASA, CAL FIRE, FEMA, and the California Air National Guard, classifies building damage into four distinct degrees.

1) *No damage* is defined as no sign of water, structural or shingle damage, or burn marks.
2) *Minor damage* is defined as building partially burnt, water surrounding structure, volcanic flow nearby, roof elements missing, or visible cracks.
3) *Major damage* is defined as partial wall or roof collapse, encroaching volcanic flow, or surrounded by water/mud.

Manuscript received 26 November 2023; revised 1 January 2024; accepted 11 January 2024. Date of publication 16 January 2024; date of current version 22 February 2024. *(Corresponding author: Minh Nguyen.)*

Wen Lu is with the School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand, and also with FHE Electrical Ltd., T/A Kinetic Electrical East Tamaki, Auckland 2013, New Zealand (e-mail: wen.lu@autuni.ac.nz).

Minh Nguyen is with the School of Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland 1010, New Zealand (e-mail: minh.nguyen@aut.ac.nz).

Lu Wei is with the School of Information Science and Engineering, Wuchang Shouyi University, Wuhan 430064, China (e-mail: weilu@wsyu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3354310

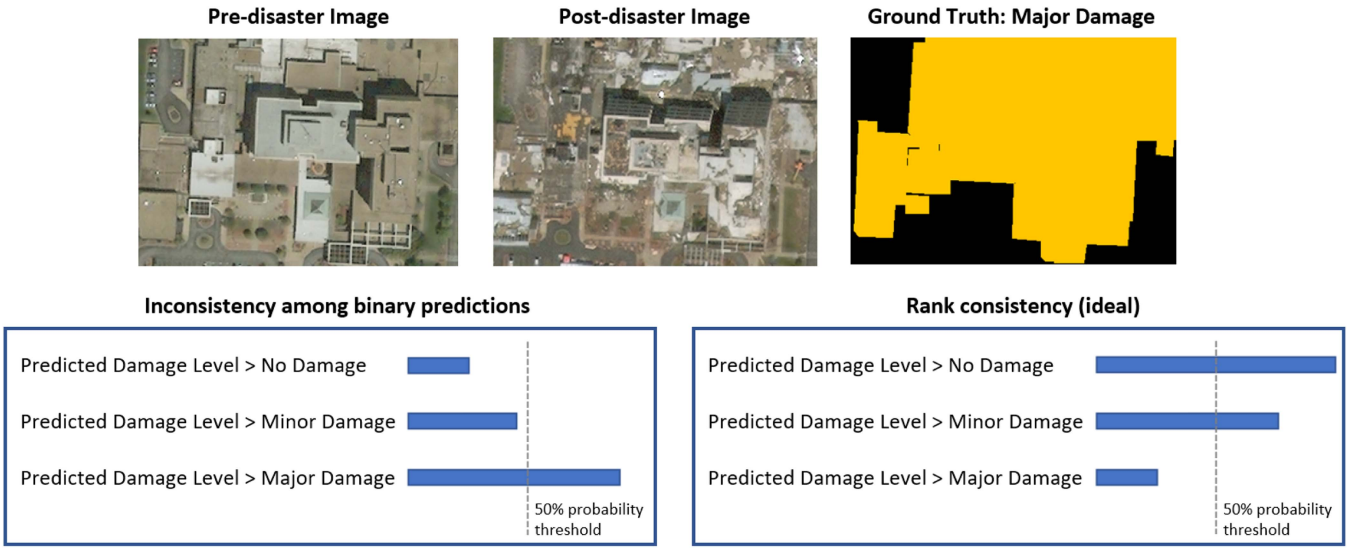

Fig. 1. Left side shows inconsistent predictions, whereas the right side demonstrates ideal predictions where the probabilities consistently decrease.

4) *Destroyed* is defined as scorched, completely collapsed, partially/completely covered with water/mud, or otherwise, no longer present.

Evidently, the differentiation between *No damage* and *Destroyed* is more conspicuous than that between *Major damage* and *Destroyed*. In real-world scenarios, a model that erroneously predicts a *Destroyed* building as *Major damage* would be more accurate and valuable for locating injured individuals than one that misclassifies a *Destroyed* building as *No damage*.

However, existing BDA methods [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] overlook the intrinsic order among ordinal labels and simplistically treat the BDA task as a multiclass semantic segmentation problem. These methods employ traditional classification loss functions, such as crossentropy loss, dice loss, and focal loss, to train their change detection networks. Nevertheless, these loss functions have a drawback: When a building is labeled as *destroyed*, predicting *no damage* or *major damage* incurs the same loss, disregarding the more significant difference between *no damage* and *destroyed*. The BDA task cannot be approached as a metric regression problem either because the distance between ordinal ranks cannot be quantified. For example, the difference between *no damage* and *minor damage* cannot be quantitatively compared with the difference between *minor damage* and *major damage*.

To correctly utilize the ordering information, we approach the BDA task as an ordinal regression problem. In this framework, we transform the K ranks into $K-1$ binary classification problems, where each kth task predicts whether the damage level exceeds rank r_k $(k = 1, \ldots, K-1)$. While all $K-1$ tasks share the same intermediate layers, they possess distinct weight parameters in the output layer. However, traditional ordinal regression methods do not guarantee consistent predictions, leading to potential disagreement among the predictions for individual binary tasks. This inconsistency arises when, for example, the kth binary task indicates that the damage level surpasses *major damage*, whereas a preceding binary task suggests that the damage level falls below *minor damage*, as illustrated in Fig. 1. In order to address these inconsistencies, we employ the conditional ordinal regression for neural networks (CORN) [20] as the loss function to ensure rank-monotonicity and maintain consistent confidence scores. CORN achieves rank consistency through an innovative training scheme that utilizes conditional training sets to obtain unconditional rank probabilities by applying the chain rule for conditional probability distributions.

With aligned preceding and subsequent images, a key question is how to effectively model the spatio-temporal semantic relations between the bitemporal pair? Some convolutional neural networks (CNNs) simply concatenate or subtract bitemporal features to extract change-related information [8], [9], [10], [21], [22]. While concatenation allows for the preservation of the original semantic information within monotemporal images, it fails to incorporate prior knowledge of changes. On the other hand, subtraction enables the acquisition of prior knowledge of changes, but at the cost of losing the original semantic information. Some CNN methods utilize attention mechanisms for bitemporal feature fusion. However, they either apply attention separately to enhance features in each monotemporal image [23], [24], [25], or use attention to reweight the fused bitemporal features in the channel or spatial dimensions [26], [27], [28], [29], [30], [31], [32], [33], [34], instead of using attention mechanisms to model the correlation within the bitemporal image pair.

As shown in Fig. 2, during natural disasters, such as hurricanes and volcanic eruptions, certain damaged buildings exhibit no discernible differences between predisaster and postdisaster images; therefore, the damage levels are evaluated based on their surrounding water or lava. Due to inherently limited receptive field, the CNN features of such a damaged building in both predisaster and postdisaster images would exhibit

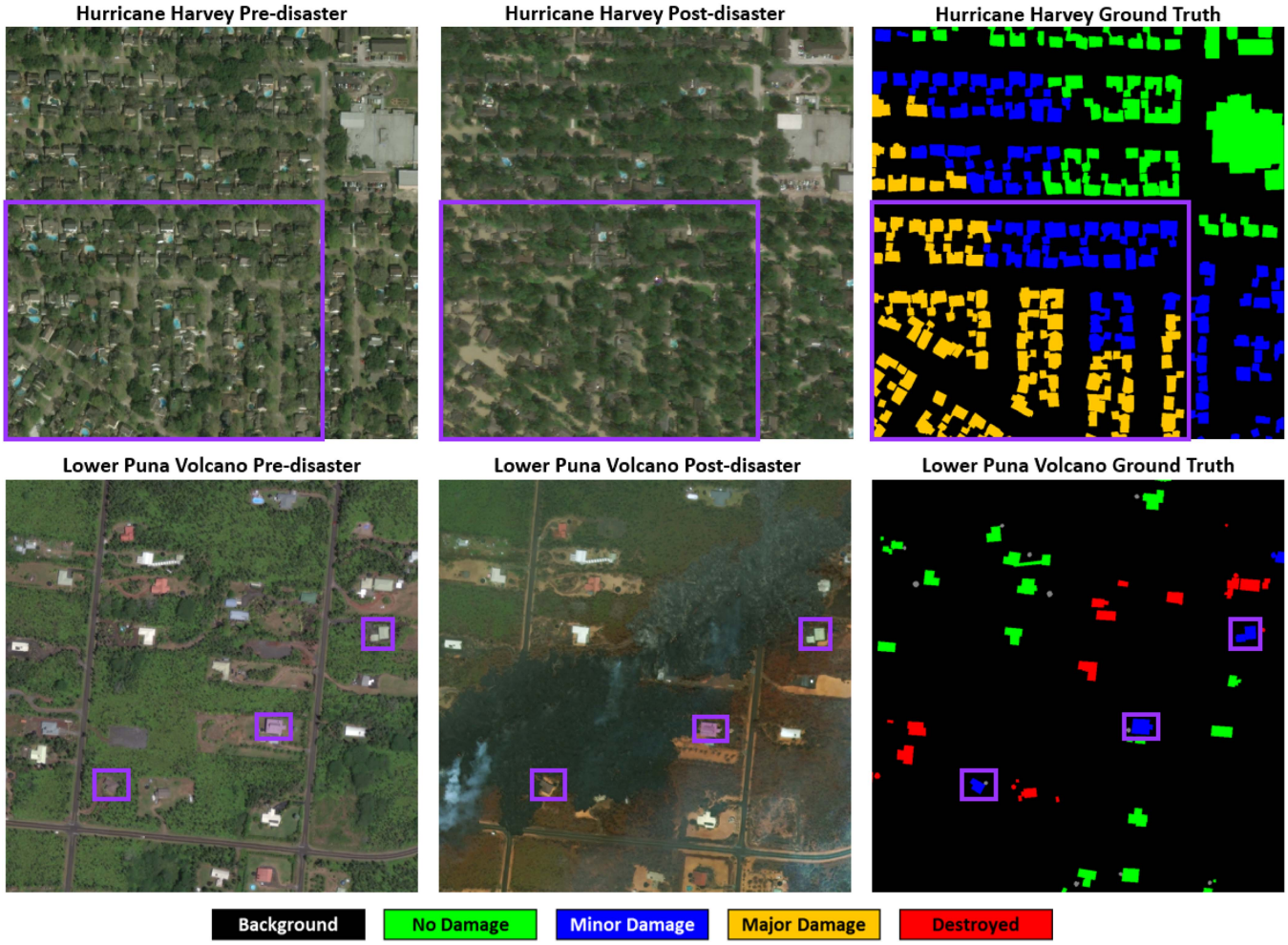

Fig. 2. No discernible difference in the roofs of the damaged buildings within the purple rectangles when comparing predisaster and postdisaster images. The assessment of building damage levels is based on the presence of surrounding water or lava. The images are sourced from the xBD dataset [7].

a high degree of similarity. Consequently, without access to extensive receptive context information, CNN faces difficulties in accurately differentiating between varying degrees of damage. Following its dominance in the field of natural language processing, the transformer architecture has shown superior performance compared with CNN in monotemporal computer vision tasks, such as image classification, object detection, and semantic segmentation. Moreover, the transformer has achieved remarkable success in multimodal computer vision tasks, including visual question answering, visual commonsense reasoning, crossmodal retrieval, and image captioning [35]. Compared with CNN, the transformer, with its nonlocal attention mechanism, is better suited for change detection tasks. However, conventional transformer models possess three limitations. First, their self-attention mechanism is designed for monotemporal computer vision tasks, rendering it incapable of capturing the temporal relationships inherent in a bitemporal image pair. Second, their global attention mechanism has quadratic computational complexity relative to the image size. Given the large scale of remote sensing imagery and dense prediction tasks, such as change detection, this computational expense becomes unaffordable. Third, as illustrated in Fig. 2, the surroundings of damaged buildings are often submerged in water or covered in lava, whereas the adjacent areas remain unaffected, with intact buildings. Consequently, distant information becomes irrelevant or possibly misleading, necessitating a focus on middle-range context rather than long-range.

In recent years, specific cross-attention (CA) mechanisms have emerged, tailored to model the temporal relationships within a bitemporal image pair [36], [37], [38], [39], [40], [41]. CA mechanisms enable models to focus on relevant areas in both images and learn the spatio-temporal relationships between them. To effectively and efficiently model the spatio-temporal semantic relations between bitemporal images, we propose a novel CA mechanism called the bitemporal attention module (BAM), which detects discrepancies through bitemporal mutual information. Initially, both the preceding image and the subsequent image are processed by encoder to extract features. The extracted features from one temporal image are used to attend to specific regions in the other temporal image. For each

pixel in the first image, the CA mechanism identifies the most relevant pixels in the second image. The similarity in feature space is used to compute this relevance. Once the CA mechanism has associated pixels in both images, change detection can be performed. Therefore, the BAM is effective in handling misalignments, allowing for more accurate and robust change detection. From another viewpoint, the BAM treats the change detection problem as a question-and-answer (Q&A) scenario. Specifically, a Query token in the subsequent image serves as a question, asking the Key and Value tokens in the preceding image to which change level it belongs. Similarly, a Query token in the preceding image can be seen as a question, asking the Key and Value tokens in the subsequent image how much change has happened to it. To mitigate computational complexity while achieving the desired middle-range context, the BAM also integrates the shifted windowing scheme proposed by Swin Transformer [42]. This strategy confines attention computation within nonoverlapping local windows that partition the image, with a fixed number of patches per window, resulting in linear computational complexity relative to image size.

Some BCD datasets, such as the S2Looking dataset [43], include more than just the normal *building change* labels, it also provides the *demolished* labels, and the *newly built* labels. However, existing change detection models [11], [21], [23], [24], [25], [27], [31], [33], [40], [44], [45], [46], [47], [48], [49] solely support one type of label. They only use the *building change* labels and disregard the valuable information provided by the other two types of labels. Therefore, these methods are limited to predicting only *building change* labels and lack the capability to predict *demolished* or *newly built* labels. In contrast, our BAM can support all three types of labels.

The contributions of this work can be summarized in the following three aspects.

1) We propose the BAM, a novel CA mechanism designed to effectively and efficiently model the spatio-temporal semantic relations between a pair of bitemporal remote sensing images.
2) We construct an efficient semantic segmentation backbone to extract features from buildings and their surroundings, and then integrate the BAM into a Siamese network composed of the backbones, forming a change detection network named bitemporal attention transformer (BAT).
3) We recognize the significance of ordinal relationships and approach the BDA task as an ordinal regression problem. To avoid potential disagreement among the predictions for individual binary tasks, our framework employs a regression loss function with strong theoretical guarantees for rank-monotonicity.

II. Related Work

This section commences with an overview of CNN-based BCD and BDA methods. Following that, we introduce various recently published CA mechanisms and conduct a comparative analysis with our proposed CA mechanism, the BAM.

A. CNN-Based BCD and BDA Methods

Fully convolutional network (FCN) [50] introduced an end-to-end paradigm for pixelwise prediction, starting a new era of applying deep learning for change detection tasks. Change detection through deep learning can be broadly classified into early fusion and late fusion approaches. The early fusion approach is rooted in semantic segmentation, wherein bitemporal images are concatenated and input to a deep learning network, undergoing direct training using ground truth. However, distinct from semantic segmentation, change detection entails the extraction of not just the semantics present in a single image, but also the change-related information derived from dual-phase semantics. In the early fusion approach, the semantic information belonging to an individual temporal image is mixed and confused with the change-related information between the two temporal images. To overcome the issue of semantic confusion, the late fusion approach decouples feature extraction and feature fusion. It commences by individually extracting features from each temporal image, subsequently conducting predictions using either metric-based or classification-based strategy. The metric-based entails the construction of a parameterized embedding space, characterized by a large distance between the changed pixels and a small distance between the unchanged pixels. On the other hand, the classification-based strategy fuses the two temporal features to generate a probability map wherein positions with changes receive higher scores compared with unchanged positions. Regarding the loss function, the former strategy commonly employs the contrastive loss functions such as triplet loss [51], whereas the latter strategy utilizes conventional classification loss functions, such as crossentropy loss or dice loss.

In late fusion approaches, Weber and Kané [8] fused the bitemporal features by concatenation, whereas Gupta and Shah [9] fused them by subtraction. Concatenation can effectively retain building features, but lacks prior knowledge of changes. Conversely, subtraction enables the acquisition of prior knowledge of changes, but leads to the loss of building features and is incapable of handling pseudochanges originating from seasonal variations, weather conditions, differences in illumination, or disparities in image sources. As a strategy of allocating larger weights to informative parts of a feature map, various attention mechanisms have replaced the aforementioned simple fusion methods in recent research studies. For example, DSIFN incorporates a channel attention module and a spatial attention module subsequent to the fusion of bitemporal features and upper-level change features [26]. In STANet, a self-attention mechanism is employed to compute attention weights among pairs of pixels across different temporal instances and spatial locations [27]. ADS-Net introduced a dual-stream attention mechanism subsequent to the fusion of bitemporal features along with their subtraction outcomes [28]. BDANet proposed a crossdirectional attention module to explore the correlations between prediaster and postdisaster images [18]. Siam-U-Net-Attn introduced a self-attention module to incorporate long-range information from the entire image [14]. Deng and Wang [17] used shuffle attention to correlate buildings before and after

the disaster. LGPNet incorporated two general attention mechanisms, the position attention module and the channel attention module [34]. These mechanisms facilitate adaptive selection and enhancement of building features exhibiting high semantic responses. To further enhance its focus on buildings and alleviate the influence of other ground targets, the LGPNet adopts a cross-task transfer learning strategy. This strategic approach significantly boosts the network's performance in isolating and analyzing building features.

However, these methods either apply attention separately to enhance features in each monotemporal image [23], [24], [25], or use attention to reweight the fused bitemporal features in the channel or spatial dimensions [26], [27], [28], [29], [30], [31], [32], [33], [34], instead of using attention mechanisms to model the correlation within the bitemporal image pair.

B. CA Mechanisms

In recent years, specific CA mechanisms have emerged, tailored to model the temporal relationships within a bitemporal image pair. CA mechanisms enable models to focus on relevant areas in both images and learn the spatio-temporal relationships between them. For example, changer includes a series of alternative interaction layers in the feature extractor and proposes a flow-based dual-alignment fusion module, which allows interactive alignment and feature fusion [40]. FCCDN designs a dense connection-based feature fusion module to fuse bitemporal features [41]. PGLF proposes a multiscale spatiotemporal interaction module to model and enhance the spatial and temporal correlations between paired change features and extract robust change representations under the constraint of bidirectional temporal direction changes.

To enable deep and long-range modeling of temporal correlations in the semantic space, SCanNet [36] employed the cross-shaped window transformer mechanism [52], which partitions the input features into vertical and horizontal stripes. However, the vertical and horizontal stripes not only retain distant and unrelated information but also omit crucial features in the adjacent diagonal area. On the contrary, our BAM adopts the shifted windowing scheme that confines attention computation within nonoverlapping local windows that partition the image, not only excluding remote and irrelevant information but also containing all the adjacent features.

CTD-Former utilized a CA mechanism that is based on the differences between similarity matrices of bitemporal images [39]. BiSRNet employed a crosstemporal semantic reasoning (Cot-SR) block to model the temporal correlations [37]. Within the CA mechanism of Cot-SR, the encoded features from one temporal image initially attend to specific regions within itself to generate the attention matrice. Subsequently, the generated attention matrice is matrix multiplied with the encoded features from the other temporal image. Its CA mechanism can be represented as $\mathrm{CA}(Q_{\mathrm{pre}}, K_{\mathrm{pre}}, V_{\mathrm{post}})$ and $CA(Q_{\mathrm{post}}, K_{\mathrm{post}}, V_{\mathrm{pre}})$. This CA mechanism's limitation resides in computing weighted sums of values, determined by relevance derived from attention scores between Queries and Keys within the same temporal feature map. Consequently, it exclusively assesses feature similarities and spatial correlations within a single temporal image, lacking direct comparisons of corresponding features across different temporal images.

In contrast, within the CA mechanism of our BAM, the encoded features from one temporal image initially attend to specific regions in the other temporal image to generate the attention matrice. Our CA mechanism can be represented as $\mathrm{CA}(Q_{\mathrm{pre}}, K_{\mathrm{post}}, V_{\mathrm{post}})$ and $\mathrm{CA}(Q_{\mathrm{post}}, K_{\mathrm{pre}}, V_{\mathrm{pre}})$. Our CA mechanism offers an advantage by directly comparing the similarity of corresponding features across different temporal images, mirroring human behavior and aligning more intuitively with the concept of "cross-attention." Detecting changes necessitates identifying the same building captured from different angles and locating any updated components. The registration process for the bitemporal image pair suffers from inherent inaccuracies owing to varying side-looking angles and terrain undulations [43]. This serves as a test of a change detection model's ability to tolerate minor registration inaccuracies. Another strength of our CA mechanism lies in the integration of a Query vector from one temporal image into the computation of attention scores with Key vectors within a local window from the other temporal image. This integration enhances the mechanism's ability to accommodate minor registration inaccuracies more effectively.

III. Proposed Method

This section begins by presenting the novel CA mechanism called the BAM. Next, we introduce our change detection network, BAT. Finally, we propose an ordinal regression training pipeline along with an object-based prediction pipeline tailored for BDA.

A. Bitemporal Attention Module

In contrast to bitemporal fusion methods, such as concatenation, subtraction, channel attention, and self-attention, BAM detects discrepancies through bitemporal mutual information. Initially, both the preceding image and the subsequent image are processed by encoder to extract features. The extracted features from one temporal image are used to attend to specific regions in the other temporal image. For each pixel in the first image, the CA mechanism identifies the most relevant pixels in the second image. The similarity in feature space is used to compute this relevance. Once the CA mechanism has associated pixels in both images, change detection can be performed. Therefore, the BAM is effective in handling misalignments, allowing for more accurate and robust change detection. From another viewpoint, the BAM captures change features by treating the change detection problem as a Q&A scenario. In this scenario, a Query token in the subsequent image asks the Key and Value tokens in the preceding image about its change level. Similarly, a Query token in the preceding image asks the Key and Value tokens in the subsequent image about the amount of change that has occurred to it. The upper part of Fig. 3 illustrates how the CA mechanism in the BAM designates one temporal image as the source and the other as the inquirer, enabling the detection of differences through mutual information. Furthermore, to eliminate remotely

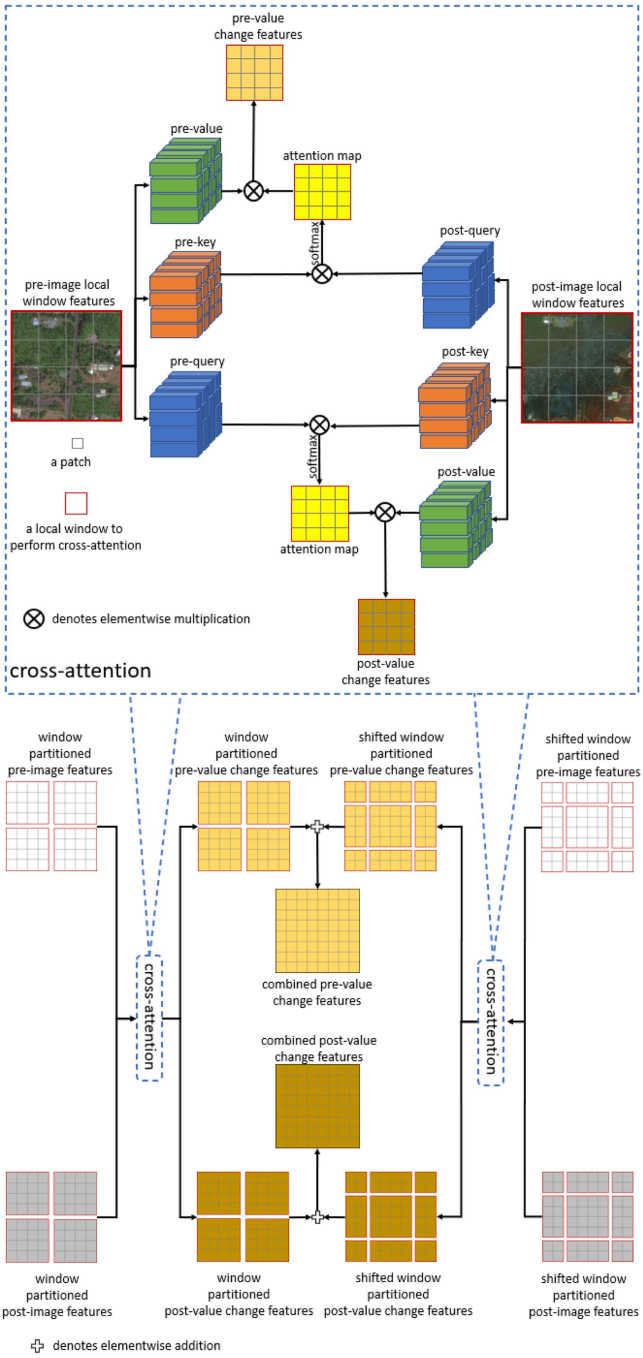

Fig. 3. Structure of BAM. Initially, a bitemporal feature pair undergoes partitioning into nonoverlapping local windows/shifted windows, followed by CA within these windows.

irrelevant information and reduce computational complexity, the BAM integrates the shifted windowing scheme proposed by Swin Transformer [42] to restrict the CA mechanism within a specific range. This strategy confines attention computation within nonoverlapping local windows that partition the image, with a fixed number of patches per window, resulting in linear computational complexity relative to image size. The CA mechanism within a local window is formulated as

$$\mathrm{CA}(Q_{\mathrm{pre}}, K_{\mathrm{post}}, V_{\mathrm{post}}) = \sigma(Q_{\mathrm{pre}} K_{\mathrm{post}}^{T}/\sqrt{d} + B)V_{\mathrm{post}} \quad (1)$$

$$\mathrm{CA}(Q_{\mathrm{post}}, K_{\mathrm{pre}}, V_{\mathrm{pre}}) = \sigma(Q_{\mathrm{post}} K_{\mathrm{pre}}^{T}/\sqrt{d} + B)V_{\mathrm{pre}} \quad (2)$$

where $Q_{\mathrm{pre}}, K_{\mathrm{pre}}, V_{\mathrm{pre}} \in \mathbb{R}^{M^2 \times d}$ are the predisaster Query, Key, and Value matrices, $Q_{\mathrm{post}}, K_{\mathrm{post}}, V_{\mathrm{post}} \in \mathbb{R}^{M^2 \times d}$ are the postdisaster Query, Key, and Value matrices, σ is softmax operation, d is the Query/Key dimension, M^2 is the number of patches in a window, and B is the relative position bias. In the CA mechanism, $Q_{\mathrm{pre}} K_{\mathrm{post}}^{T}$ and $Q_{\mathrm{post}} K_{\mathrm{pre}}^{T}$ are considered as bitemporal mutual information, whereas V_{post} and V_{pre} are considered as the monotemporal image features.

As illustrated in the lower part of Fig. 3, the CA mechanism accepts either a window or shifted window partitioned bitemporal feature pair as input, producing a change feature pair (consisting of a prevalue change feature map and a postvalue change feature map) that is also window or shifted window partitioned. Afterward, to establish connections among the windows, the window partitioned change features and the shifted window partitioned change features are combined through addition, resulting in combined prevalue change features and combined postvalue change features.

B. Bitemporal Attention Transformer

To avoid the issue of semantic confusion in the early fusion approach, we adopt a late fusion approach that separates the processes of feature extraction and feature fusion. Our approach utilizes a Siamese architecture that begins by extracting features individually from each monotemporal image, and subsequently integrates the bitemporal features through CA. Specifically, we construct an efficient semantic segmentation backbone to extract features from buildings and their surroundings, and then integrate the BAM into a Siamese network composed of the backbones, forming a change detection network named BAT.

Existing methods [8], [14], [18] utilize a heavyweight backbone for the Siamese branch, which necessitates cropping an original 1024×1024 pixel image into 512×512 pixel or 256×256 pixel patches due to limitations in GPU memory. Nonetheless, this image preprocessing operation inevitably leads to the splitting of some complete buildings into partial fragments, resulting in the loss of middle-range context information. Moreover, resizing the predicted results back to their original size introduces additional latency [53]. In order to overcome these limitations, we construct a lightweight semantic segmentation backbone incorporated as a branch within the Siamese architecture. This hybrid backbone combines the high efficiency of CNNs with the powerful and nonlocal modeling capability of transformers.

As illustrated in the upper part of Fig. 4, the semantic segmentation backbone utilizes EfficientNetV2 [54] as the encoder to extract multiscale features, and then, employs sequential Swin Transformer blocks [42] as the decoder to fuse these features. The EfficientNetV2 Stage 5 output features are upsampled $2\times$ and concatenated with Stage 3 output features before sending to the decoder for fusion and window and shifted window attention. The decoder output is considered as the monotemperal features.

As illustrated in the lower part of Fig. 4, both Siamese branches share a common backbone, facilitating the equitable

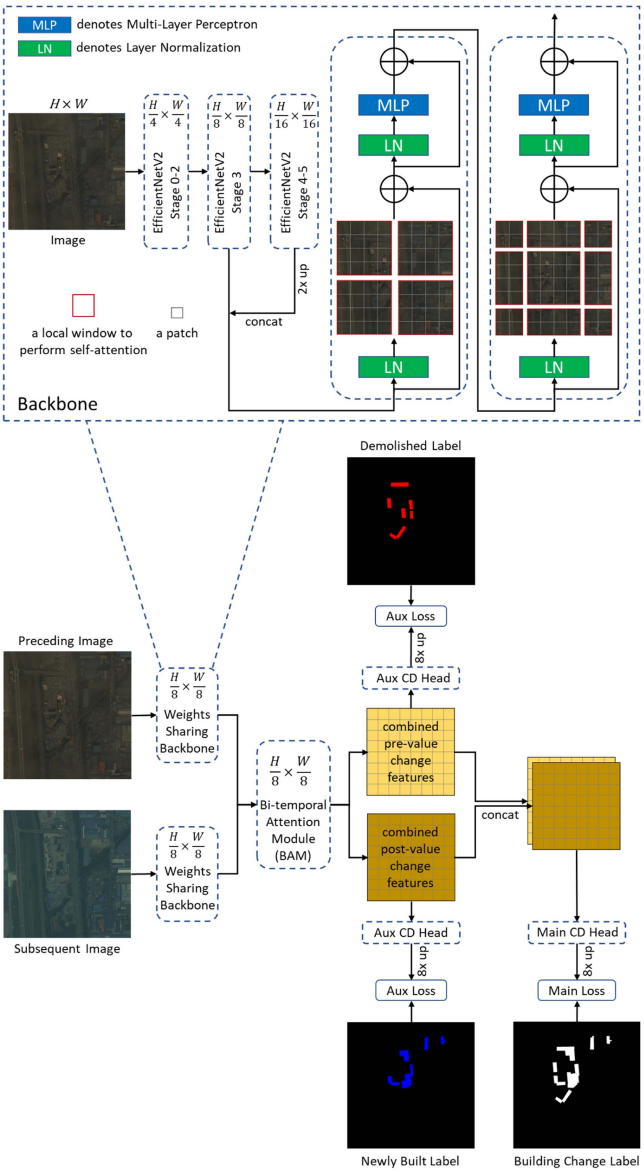

Fig. 4. Architecture of the BAT. BAT utilizes a Siamese architecture composed of weights sharing backbones to extract features individually from each monotemporal image, and subsequently integrates the bitemporal features through the BAM.

extraction of features. Subsequently, the monotemperal features from each Siamese branch are routed to the BAM, responsible for extracting spatio-temporal information through window and shifted window CA.

The outputs of the BAM consist of combined prevalue change features and combined postvalue change features. Due to their stronger association with preceding image features, we train the combined prevalue change features using the *demolished* labels. Conversely, the combined postvalue change features, which exhibit a closer relationship with subsequent image features, are trained using the *newly built* labels. We also fuse the combined prevalue change and combined postvalue change features through concatenation, forming the dual-perspective change sensitivity features, and train them using the normal *building change* labels. The pseudocode of the "forward" method of BAT is shown as follows.

```
# backbone encoder features
x1 = self.EfficientNetV2(x1)
x2 = self.EfficientNetV2(x2)
# window partitioned pre-image features
x1 = self.SwinTransformerBlocks(x1)
# window partitioned post-image features
x2 = self.SwinTransformerBlocks(x2)
# window partitioned pre-value change features
w_pre_features = self.CrossAtt1(x1, x2, shift=False)
# shifted window partitioned pre-value change features
sw_pre_features = self.CrossAtt1(x1, x2, shift=True)
# combined pre-value change features
combined_pre_features = w_pre_features + sw_pre_features
# window partitioned post-value change features
w_post_features = self.CrossAtt2(x2, x1, shift=False)
# shifted window partitioned post-value change features
sw_post_features = self.CrossAtt2(x2, x1, shift=True)
# combined post-value change features
combined_post_features = w_post_features + sw_post_features
# dual-perspective change sensitivity features
dual-perspective_change_features = torch.concat([
     combined_pre_features, combined_post_features])

return combined_pre_features, combined_post_features, dual-
     perspective_change_features
```

A simple 1×1 convolution layer serves as either the auxiliary or the main change detection head, with logits upsampled by a factor of 8 through bilinear interpolation before being directed to the auxiliary or the main loss function. The total loss L_t is the weighted sum of the main loss L_m and the two auxiliary losses L_a

$$L_t = w_1 \times L_m + w_2 \times L_a. \tag{3}$$

Fine-tuning the hyperparameters w_1 and w_2 typically leads to improved results. However, to ensure generality, we refrained from fine-tuning the hyperparameters in our experiments. Instead, we used straightforward and intuitive values. Given that building change detection is the primary task for most datasets, we assigned greater emphasis to the main loss. This was achieved by setting $w_1 = 1$ and $w_2 = 0.5$ in the subsequent experiments.

In the case of certain BCD datasets, such as the S2Looking dataset [43], which include not only the normal *building change* labels but also the *demolished* labels and the *newly built* labels, our BAT distinguishes itself from existing change detection models (e.g., [11], [21], [23], [27], [31], [40], [45], [46], [47], [48], and [49]), which exclusively support *building change* labels. BAT is capable to harness the valuable information offered by these additional label types, enabling it to predict *demolished* and *newly built* labels.

For other BCD or BDA datasets, such as LEVIR-CD+ [43] BCD dataset and xBD BDA dataset [7], which include only a single type of labels, BAT's auxiliary change detection heads and corresponding loss functions are omitted.

The BAT is suitable for both the BCD and BDA tasks. Given that BCD is a binary-class semantic segmentation problem, conventional loss functions, such as binary crossentropy loss and dice loss are applicable. However, it is inappropriate to oversimplify BDA by regarding it as a multiclass semantic segmentation task and applying conventional classification loss functions, such as crossentropy loss, dice loss, and focal loss, due to the intrinsic ordinal relationships within the labels. Consequently, we address the BDA task as an ordinal regression

problem and propose the subsequent ordinal regression training pipeline along with an object-based prediction pipeline tailored for BDA.

C. Ordinal Regression Training Pipeline for BDA

Since BDA is a combination of two subtasks: Building extraction and damage classification. For the task of building extraction, the backbone of BAT along with a simple 1×1 convolution layer served as semantic segmentation head are utilized and trained with predisaster images and their corresponding building footprint labels by binary crossentropy loss. Another two semantic segmentation heads are added upon the Stages 3 and 5 output features in the training phase, respectively, serving as auxiliary losses to enhance feature extraction ability. In the inference phase, the two auxiliary heads are discarded without incurring the additional computational cost.

Given that BDA is a combination of two subtasks: Building extraction and damage classification. For the building extraction task, we employ the backbone of BAT (as illustrated in the upper part of Fig. 4), augmented by a simple 1×1 convolution layer functioning as the semantic segmentation head. During the training phase, two additional semantic segmentation heads are introduced, one connected to the Stage 3 and the other to the Stage 5 output features. These heads function as auxiliary losses, enhancing feature extraction capability. In the inference phase, the two auxiliary heads are discarded, incurring no additional computational cost. This network is trained using predisaster images and their corresponding building footprint labels, and employs binary crossentropy as loss function.

For the damage classification task, we leverage the trained weights acquired from the building extraction task to initialize the BAT backbone. BAT is trained using bitemporal image pairs and their corresponding building damage labels, and employs a regression loss function named CORN, which guarantees rank-monotonicity to avoid rank inconsistencies among the binary tasks.

Let $D = \{\mathbf{x}^{[i]}, y^{[i]}\}_{i=1}^{N}$ denote a dataset containing N samples, in which $\mathbf{x}^{[i]} \in \mathcal{X}$ denotes the inputs of the ith sample, $y^{[i]}$ denotes its corresponding class label, and K denotes the number of classes. In ordinal regression, $y^{[i]}$ is referred as rank, where $y^{[i]} \in \mathcal{Y} = \{r_1, r_2, \ldots r_K\}$ with rank order $r_1 \prec r_2 \prec \ldots \prec r_{K-1} \prec r_K$. CORN applies a label extension to the rank labels $y^{[i]}$, such that the resulting binary label $y_k^{[i]} \in \{0, 1\}$ indicates whether $y^{[i]}$ exceeds rank r_k. CORN employs $K-1$ binary tasks in the output layer of a neural network. CORN estimates a series of conditional probabilities using conditional training subsets, such that the output of the kth ($k > 1$) binary task $f_k(\mathbf{x}^{[i]})$ represents the conditional probability

$$f_k(\mathbf{x}^{[i]}) = \hat{P}(y^{[i]} > r_k | y^{[i]} > r_{k-1}) \tag{4}$$

where the events are nested, $\{y^{[i]} > r_k\} \subseteq \{y^{[i]} > r_{k-1}\}$.

When $k = 1$, $f_k(\mathbf{x}^{[i]})$ represents the initial unconditional probability

$$f_1(\mathbf{x}^{[i]}) = \hat{P}(y^{[i]} > r_1). \tag{5}$$

The equivalent unconditional probabilities are computed by applying the chain rule

$$\hat{P}(y^{[i]} > r_k) = \prod_{j=1}^{k} f_j(\mathbf{x}^{[i]}). \tag{6}$$

Since $\forall j, 0 \leq f_j(\mathbf{x}^{[i]}) \leq 1$, we have

$$\hat{P}(y^{[i]} > r_1) \geq \hat{P}(y^{[i]} > r_2) \geq \ldots \geq \hat{P}(y^{[i]} > r_{K-1}) \tag{7}$$

which guarantees rank consistency among the $K-1$ binary tasks.

The neural network aims to estimate the initial unconditional probability $f_1(\mathbf{x}^{[i]})$ and the conditional probabilities $f_2(\mathbf{x}^{[i]}), \ldots, f_{K-1}(\mathbf{x}^{[i]})$. Estimating $f_1(\mathbf{x}^{[i]}) = \hat{P}(y^{[i]} > r_1)$ is a classic binary classification task with the binary label $y_1^{[i]}$. To estimate the conditional probability $f_k(\mathbf{x}^{[i]}) = \hat{P}(y^{[i]} > r_k | y^{[i]} > r_{k-1})$, only the subset of the dataset where $y^{[i]} > r_{k-1}$ is needed.

Let $f_j(\mathbf{x}^{[i]})$ denote the predicted value of the jth node in the output layer of the network, and let $|S_j|$ denote the size of its conditional training set. The loss function $L(\mathbf{X}, \mathbf{y})$is

$$-\frac{1}{\sum_{j=1}^{K-1}|S_j|}\sum_{j=1}^{K-1}\sum_{i=1}^{|S_j|}\Bigg[\log\Big(f_j(\mathbf{x}^{[i]})\Big)\cdot\mathbb{1}\Big\{y^{[i]} > r_j\Big\} + \log\Big(1 - f_j(\mathbf{x}^{[i]})\Big)\cdot\mathbb{1}\Big\{y^{[i]} \leq r_j\Big\}\Bigg] \tag{8}$$

where $\mathbb{1}$ denotes indicator function.

To improve the numerical stability of the loss gradients during training, the following alternative formulation $L(\mathbf{Z}, \mathbf{y})$ of the loss is implemented:

$$-\frac{1}{\sum_{j=1}^{K-1}|S_j|}\sum_{j=1}^{K-1}\sum_{i=1}^{|S_j|}\Bigg[\log\Big(\sigma(\mathbf{z}^{[i]})\Big)\cdot\mathbb{1}\Big\{y^{[i]} > r_j\Big\} + \Big(\log\Big(\sigma(\mathbf{z}^{[i]})\Big) - \mathbf{z}^{[i]}\Big)\cdot\mathbb{1}\Big\{y^{[i]} \leq r_j\Big\}\Bigg] \tag{9}$$

where $\mathbf{Z}$ are the inputs of the last layer, σ is softmax operation, and $\log(\sigma(\mathbf{z}^{[i]})) = \log(f_j(\mathbf{x}^{[i]}))$.

The rank index q of the ith sample is obtained by thresholding the predicted probabilities corresponding to the $K-1$ binary tasks and summing the binary labels as follows:

$$q^{[i]} = 1 + \sum_{j=1}^{K-1}\mathbb{1}\Big(\hat{P}(y^{[i]} > r_j) > 0.5\Big) \tag{10}$$

where the predicted rank is $r_{q^{[i]}}$.

Besides taking advantage of ordinal information, another benefit of employing CORN as the loss function lies in the reduced number of output logits channels, which is $K-1$, compared with the conventional loss functions where it is K. Consequently, CORN consumes less memory than alternative loss functions.

In order to make BAT focus on building change detection, only the pixels within the building footprint contribute to the loss calculation, whereas those in the background are disregarded.

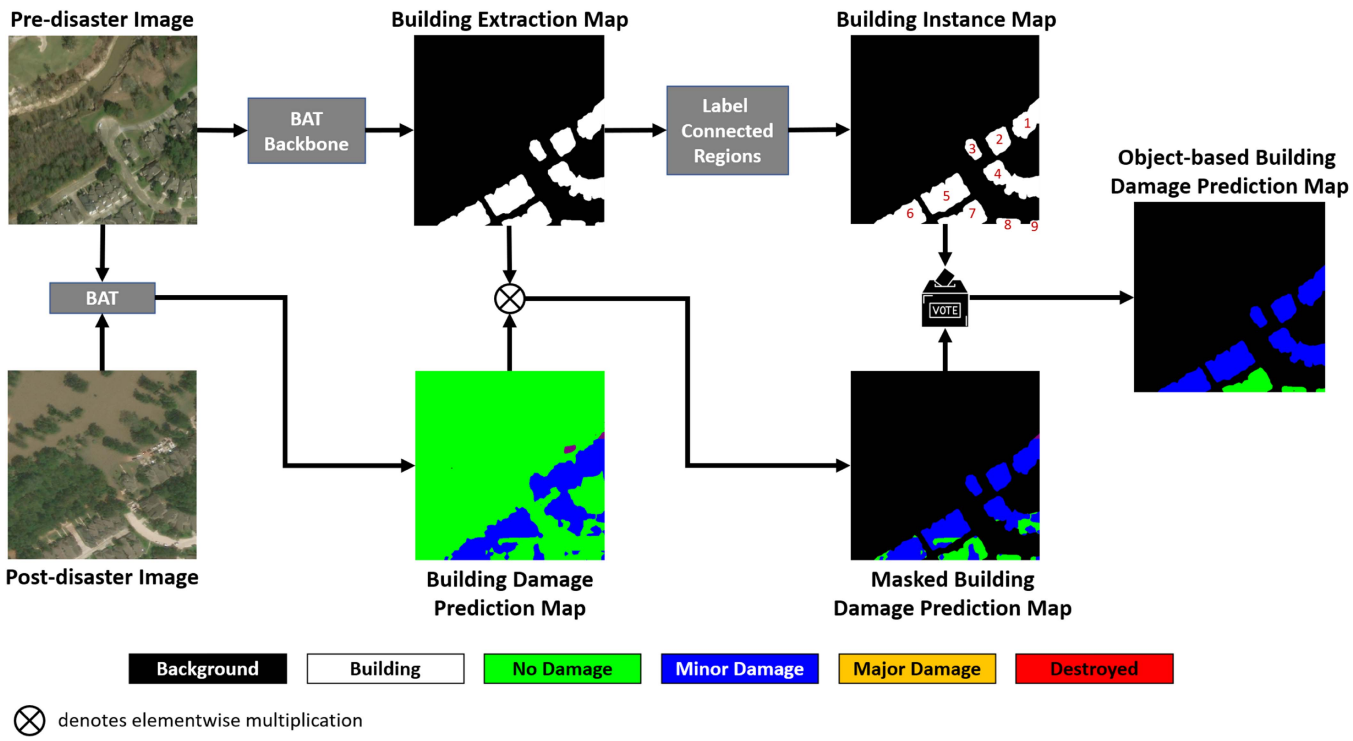

Fig. 5. Object-based prediction pipeline for BDA.

D. Object-Based Prediction Pipeline for BDA

The object-based prediction pipeline comprises the sequential steps depicted in Fig. 5 and detailed in Algorithm 1.

- Step 1: Building Extraction. The predisaster image is input to a BAT backbone network with pretrained frozen weights for the first subtask, yielding a building extraction map.
- Step 2: Instance Segmentation. The connected component labeling algorithm [55], [56] is applied on the building extraction map from Step 1 to assign a distinct label to each extracted building. This process transforms the semantic segmentation map from Step 1 into an instance segmentation map.
- Step 3: Pixel-based Building Damage Classification. The bitemporal image pair is fed into a BAT with pretrained frozen weights for the second subtask, yielding a pixel-based building damage prediction map.
- Step 4: Background Removal. To eliminate the background, the building extraction map from Step 1 serves as a mask, which is then applied to the building damage prediction map from Step 3 through multiplication, resulting in a masked building damage prediction map.
- Step 5: Object-based Building Damage Classification. In order to ensure label consistency within individual buildings, majority voting is carried out within individual building instances. This process transforms the pixel-based predictions from Step 4 into object-based predictions.

Algorithm 1: Object-Based Prediction Pipeline for BDA.

Input: Pre-disaster Image $\mathbf{X1}$, Post-disaster Image $\mathbf{X2}$, Connected Component Labeling Algorithm $\mathbf{CCL}$.
Output: Object-based Prediction $\hat{\mathbf{Y}}$.

$\#Step1 : Building\ Extraction$
$\mathbf{Y_b} \leftarrow \mathbf{BAT's\ backbone(X1)}$
$\#Step2 : Instance\ Segmentation$
$\mathbf{Y_i} \leftarrow \mathbf{CCL(Y_b)}$
$\#Step3 :$
$Pixel-based\ Building\ Damage\ Classification$
$\mathbf{Y_d} \leftarrow \mathbf{BAT(X1, X2)}$
$\#Step4 : Background\ Removal$
$\mathbf{Y_m} \leftarrow \mathbf{Y_b} \otimes \mathbf{Y_d}$
$\#Step5 :$
$Object-based\ Building\ Damage\ Classification$
$\hat{\mathbf{Y}} \leftarrow \mathbf{Vote(Y_m, Y_i)}$
return $\hat{\mathbf{Y}}$
$\otimes$ denotes elementwise multiplication.

IV. BCD Experimental Results

To assess BAT's efficacy in BCD tasks, we conducted experiments on two BCD datasets, comparing it with various methods.

A. Experimental Setup

1) BAT Parameter Setting: We employed EfficientNetV2S as the backbone encoder, and configured the CA mechanism in BAM with a window size of 16.

2) Training Details: Network training was conducted on an NVIDIA RTX 3090 GPU within a PyTorch environment. We employed AdamW as the optimizer, with a batch size of 7 and a base learning rate of 0.0001, utilizing cosine decay. Crossentropy was used as the loss function. The networks underwent 200 epochs of training, incorporating a warmup strategy during the initial 50 epochs.

3) Data Augmentation: During the training process, we applied random flipping, random rotation, random scaling with rates (0.8, 0.9, 1.0, 1.25, 1.5), random cropping to the size of 768×768 pixels, and color jittering operations to the input images.

4) Evaluation Metrics: In accordance with previous research, we employed Precision, Recall, and F1-score to assess the effectiveness of different methods. The three metrics are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

where TP represents the count of true-positive pixels, FP represents the count of false-positive pixels, and FN represents the count of false-negative pixels.

Recall measures the method's effectiveness in identifying the regions that have undergone changes. Precision assesses how effectively the method filters out irrelevant and unchanged structures from the prediction results. The F1-score offers a comprehensive assessment of the prediction results.

B. Compared Methods

We make a comparison to representative and state-of-the-art BCD methods, which are described as follows.

FC-EF, FC-Siam-Conc, and FC-Siam-Diff [21] belong to the category of classification-based UNet-like models. FC-EF employs early fusion by directly concatenating bitemporal images, whereas FC-Siam-Conc utilizes Siamese encoders and concatenation for feature fusion. FC-Siam-Diff, on the other hand, employs Siamese encoders and difference for feature fusion.

DTCDSCN [23] incorporates a dual-attention module for capturing interdependencies among channels and spatial positions, thus, enhancing the representation of features.

STANet [27] is a Siamese network with a self-attention mechanism to compute attention weights among pairs of pixels across different temporal instances and spatial locations. STANet-BAM includes the basic spatial-temporal attention module, whereas STANet-PAM includes the pyramid spatial-temporal attention module.

CDNet [45] is a Siamese CNN with instance-level data augmentation. Through generative adversarial training, the instance-level data augmentation can generate bitemporal images that contain changes involving numerous and diverse synthesized building instances.

IFNet [26] extends the design principles of FC-Siam-Conc, it incorporates a channel attention module and a spatial attention module subsequent to the fusion of bitemporal features and upper-level change features.

SNUNet [31] employs a nested U-Net architecture, incorporating dense skip connections between the Siamese encoder and multiple subdecoders. This design choice is made to mitigate the loss of spatial position information within the deep decoder layers.

ChangerEx [40] includes a series of alternative interaction layers in the feature extractor and proposes a flow-based dual-alignment fusion module, which allows interactive alignment and feature fusion.

ChangeStar [44] presents a scalable multitemporal remote sensing change data generator via generative modeling, it decouples the complex simulation problem into change event simulation and semantic change synthesis.

BiT [11] is a hybrid of CNN and transformer, it employs convolutional blocks in the shallow layers and transformer blocks in the deeper layers.

ChangeFormer [47] is a Siamese network based on the transformer architecture. It integrates a hierarchical transformer encoder with an MLP decoder to capture and represent long-range details.

TransUNetCD [49] is a hybrid of UNet and transformer, it uses a difference enhancement module to generate a difference feature map containing rich change information.

C. Experimental Results on the LEVIR-CD+ Dataset

As the expansion of the LEVIR-CD dataset [27], the LEVIR-CD+ BCD dataset [43] comprises 985 near-nadir satellite image pairs, each with dimensions of 1024×1024 pixels and a spatial resolution of 0.5 m/pixel. It spans 20 regions within various cities in Texas, USA. Each pair of bitemporal images captures a time span of five years. The dataset is officially divided into a training set, consisting of 637 pairs of bitemporal images, and a test set, comprising 348 pairs of bitemporal images. Following established practices, we employed the official training set for network training and the official test set for reporting results. As this dataset only provides the normal *building change* labels, we excluded the two auxiliary heads and corresponding losses, training the BAT solely with this single type of label.

As demonstrated in Table I, BAT exhibits a notably higher recall rate, substantiating the efficacy of the CA mechanism in modeling the change detection process by the Q&A scenario. Despite a higher recall rate, BAT attains a precision rate on par with the leading performance methods, ultimately resulting in the highest F1 Score.

Fig. 6 displays the predicted building changes. This visualization highlights BATs' ability to not only differentiate building alterations from seasonal and land-cover variations, but also accurately reconstruct the boundary details of the modified structures.

In addition, we conducted an assessment of BAT's inference speed employing an NVIDIA RTX 2060 Max-Q 6 G Mobile GPU with a computational capacity of 4.55 TFLOPS in FP32, which closely matches the computational capability of

TABLE I
COMPARISON OF BAT WITH OTHER METHODS ON THE LEVIR-CD+ DATASET

Method	Precision	Recall	F1
FC-EF [21]	0.6130	0.7261	0.6648
FC-Siam-Conc [21]	0.6624	0.8122	0.7297
FC-Siam-Diff [21]	0.7497	0.7204	0.7348
DSAMNet [57]	0.6976	0.8031	0.7466
DTCDSCN [23]	0.8036	0.7503	0.7760
L-Unet [58]	0.7899	0.7918	0.7909
STANet-PAM [27]	0.7462	0.8454	0.7931
SNUNet [31]	0.7951	0.8142	0.8045
CDNet [45]	0.8896	0.7345	0.8046
TFI-GR [59]	0.7972	0.8345	0.8154
BiT [11]	0.8274	0.8285	0.8280
A2Net [32]	0.8525	0.8127	0.8321
MSCANet [60]	0.8580	0.8124	0.8346
DCAT [61]	0.8472	0.8334	0.8402
IFN [26]	0.8582	0.8324	0.8451
Hu et al. [46]	0.8874	0.8363	0.8611
AR-CDNet [22]	0.8662	0.8618	0.8640
FHD [48]	**0.8960**	0.8383	0.8662
BAT	0.8829	**0.8622**	**0.8724**

The bold entities denote the best performance.

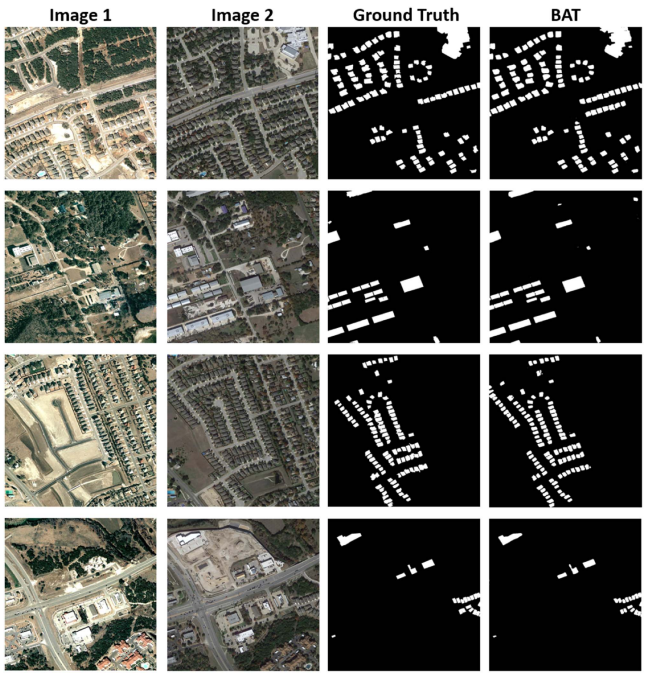

Fig. 6. Predicted building changes by BAT on the LEVIR-CD+ dataset.

an NVIDIA Jetson AGX Orin embedded system (FP32: 5.33 TFLOPS). BAT achieves an inference speed of 6.1 frames per second (FPS) when analyzing bitemporal image pairs with dimensions of 1024×1024 pixels. This high efficiency is attributed to the shifted windowing scheme that is integrated in the BAM, which mitigates computational complexity from quadratic in terms of image size to linear in terms of image size while achieving the desired middle-range context.

D. Experimental Results on the S2Looking Dataset

The S2Looking dataset [43] stands out as both the largest and the most challenging BCD dataset to date. Comprising 5000 bitemporal image pairs, this dataset is officially divided into a training set of 3500 pairs, a validation set of 500 pairs, and a test set of 1000 pairs. In contrast with the LEVIR-CD+ dataset, which focuses on urban areas at near-nadir angles, the S2Looking dataset primarily centers on rural areas captured from varying large off-nadir angles. The S2Looking dataset exhibits significantly sparser changes in buildings compared with the LEVIR-CD+ dataset. Following established practices, we utilized the official training set for training the network, the official validation set for validation purposes, and the official test set for reporting the results. As this dataset provides not only the normal *building change* labels but also the *demolished* labels and the *newly built* labels, we kept the two auxiliary heads and corresponding losses, training the BAT with all the three types of labels, as illustrated in the lower part of Fig. 4. To evaluate the impact of the two auxiliary heads and corresponding losses, we constructed a variant network that lacked these components. This variant network is referred to as *BAT without aux heads*.

As indicated in Table II, on this particularly challenging dataset, BAT significantly strengthens its superiority in terms of recall rate, ultimately achieving the highest F1 Score. Although FC-EF and FC-Siam-Diff attain a high level of precision, their recall rates are notably deficient, resulting in the lowest F1 Scores. This illustrates their limited capacity to detect only the most conspicuous building changes. In contrast, our BAT can identify a substantial portion of building changes while maintaining a precision rate comparable to that of the top-performing methods. This advantage can be attributed to the generation of change-sensitive features by the CA mechanism within the BAM.

As presented in Table II, BAT with aux heads stands out as the sole model capable of disentangling demolished structures from newly constructed ones. This capability empowers urban administrators to analyze these distinct changes independently. Through the extraction of change-related information embedded within the *demolished* and *newly built* labels, the auxiliary heads, along with their associated loss functions, effectively enhance both precision and recall rates.

As shown in Fig. 7, BAT effectively identifies the majority of building changes, while excluding unchanged structures and pseudochanges originating from seasonal variations, weather conditions, differences in illumination, or disparities in image sources.

Fig. 8 illustrates the comparison of predicted building changes using different methods. In the first set of bitemporal images, BAT exhibits superior building boundary recovery and a reduced occurrence of pseudochanges. For the second pair of bitemporal images, BAT stands out as the only model whose predictions closely align with the ground truth, whereas other methods exhibit numerous pseudochanges but fail to capture real alterations. In the case of the third set of bitemporal images, BAT effectively upholds a high recall rate for detecting building changes while mitigating the occurrence of pseudochanges.

V. BDA EXPERIMENTAL RESULTS

A. Experimental Setup

In order to evaluate the effectiveness of BAT, along with our proposed training and prediction pipelines, we conducted

TABLE II
COMPARISON OF BAT WITH OTHER METHODS ON THE S2LOOKING DATASET

Method	#Param (M)	MACs (G)	FLOPs (G)	Building Change			Demolished			Newly Built		
				Preci	Recall	F1	Preci	Recall	F1	Preci	Recall	F1
FC-EF [21]	1.4	12.5		0.8136	0.0895	0.0765	-	-	-	-	-	-
FC-Siam-Diff [21]	1.4	17.1		**0.8329**	0.1576	0.1319	-	-	-	-	-	-
FC-Siam-Conc [21]	1.5	19.5		0.6827	0.1852	0.1354	-	-	-	-	-	-
STANet-BAM(ResNet18) [27]	12.2	49.2		0.3119	0.5291	0.3924	-	-	-	-	-	-
STANet-PAM(ResNet18) [27]	12.2	50.2		0.3875	0.5649	0.4597	-	-	-	-	-	-
AMIO-Net [24]				0.6394	0.4925	0.5334	-	-	-	-	-	-
DTCDSCN(SE-Res34) [23]	41.1	60.9		0.6858	0.4916	0.5727	-	-	-	-	-	-
L-Unet [58]	8.5			0.5995	0.5859	0.5926	-	-	-	-	-	-
CDNet [45]	14.3			0.6748	0.5493	0.6056	-	-	-	-	-	-
MSCANet [60]	16.4			0.6463	0.5767	0.6095	-	-	-	-	-	-
BiT(ResNet18) [11]	3.0	35.0		0.7264	0.5385	0.6185	-	-	-	-	-	-
Hu et al. [46]				0.7253	0.5453	0.6225	-	-	-	-	-	-
SNUNet [31]	3.0	46.9		0.7194	0.5634	0.6319	-	-	-	-	-	-
ChangeFormer(MiT-b1) [47]	13.9	26.4		0.7282	0.5613	0.6339	-	-	-	-	-	-
IFN(VGG-16) [26]	36.0	316.5		0.6646	0.6195	0.6413	-	-	-	-	-	-
FHD [48]	11.8			0.7409	0.5671	0.6425	-	-	-	-	-	-
ChangeStar(MiT-b1) [44]	18.4		67.3	0.6930	0.5990	0.6430	-	-	-	-	-	-
CGNet [33]				0.7018	0.5938	0.6433	-	-	-	-	-	-
Xu et al. [25]	61.4			0.6968	0.6154	0.6536	-	-	-	-	-	-
ChangerEx(ResNet18) [40]	11.4	23.9		0.7359	0.6015	0.6620	-	-	-	-	-	-
ChangeStar(ResNet18) [44]	16.4		65.3	0.7090	0.6220	0.6630	-	-	-	-	-	-
TransUNetCD [49]	95.5			0.7641	0.5970	0.6703	-	-	-	-	-	-
ChangerEx(MiT-b0) [40]	3.5	8.5		0.7301	0.6204	0.6708	-	-	-	-	-	-
BAT without aux heads	6.9	36.5	73.5	0.7050	0.6399	0.6709						
BAT with aux heads	6.9	36.5	73.5	0.7225	**0.6415**	**0.6796**	0.6347	0.4409	0.5203	0.7525	0.6668	0.7071

The MACs are calculated using an RGB input image pair with a resolution of 512×512 pixels.
- means not supported by the method or not reported by the authors.
The bold entities denote the best performance.

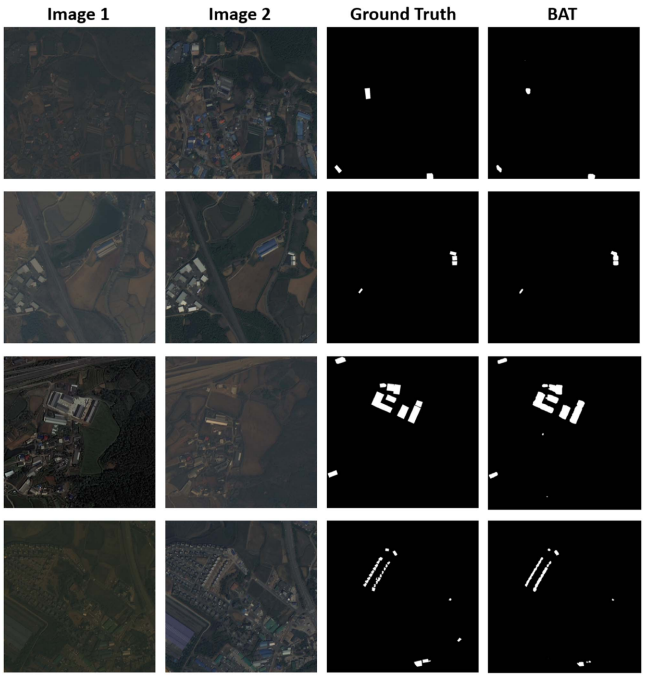

Fig. 7. Predicted building changes by BAT on the S2looking dataset.

experiments on the xBD dataset [7], the most extensive BDA dataset available in the field, and compared its performance against various methods.

The xBD dataset encompasses more than 800 000 building annotations, spanning across an area of over 45 000 square kilometers. Designed to facilitate the development of a versatile model suitable for a wide range of disaster scenarios, xBD comprises a diverse collection of disasters that occurred in various global regions between 2011 and 2019. These disasters encompass volcano eruptions, hurricanes, wildfires, floods, tsunamis, earthquakes, monsoons, and tornadoes. The dataset categorizes building damage into four classes: *no damage*, *minor damage*, *major damage*, and *destroyed*. The xBD dataset is officially divided into a training set, a test set, and a holdout set. The training set comprises 9168 pairs of aligned predisaster and postdisaster images, each sized 1024×1024 pixels. The test and holdout sets each consist of 933 pairs. Notably, the distribution of building damage classes is heavily skewed toward *no damage*, which is represented over eight times more than the other classes. Fig. 9 displays some predisaster images, postdisaster images, and corresponding ground truth masks. As prior studies, we utilized the official training set for network training, employed the official holdout set for validation, and used the official test set for reporting the results.

The xBD dataset official evaluation metric ($F1_s$) is a weighted average of the building segmentation F1 score ($F1_b$) and the harmonic mean of classwise damage classification F1 scores ($F1_d$)

$$F1_s = 0.3 \times F1_b + 0.7 \times F1_d \tag{14}$$

in which $F1_d$ is defined as

$$F1_d = \frac{n}{\sum_{i=1}^{n} 1/F1_{C_i}} \tag{15}$$

where $F1_{C_i}$ denotes the F1 score of each damage class.

This weighted F1 score, which balances precision and recall in a harmonic mean, proves particularly effective for assessing imbalanced datasets such as xBD. Using accuracy alone as a

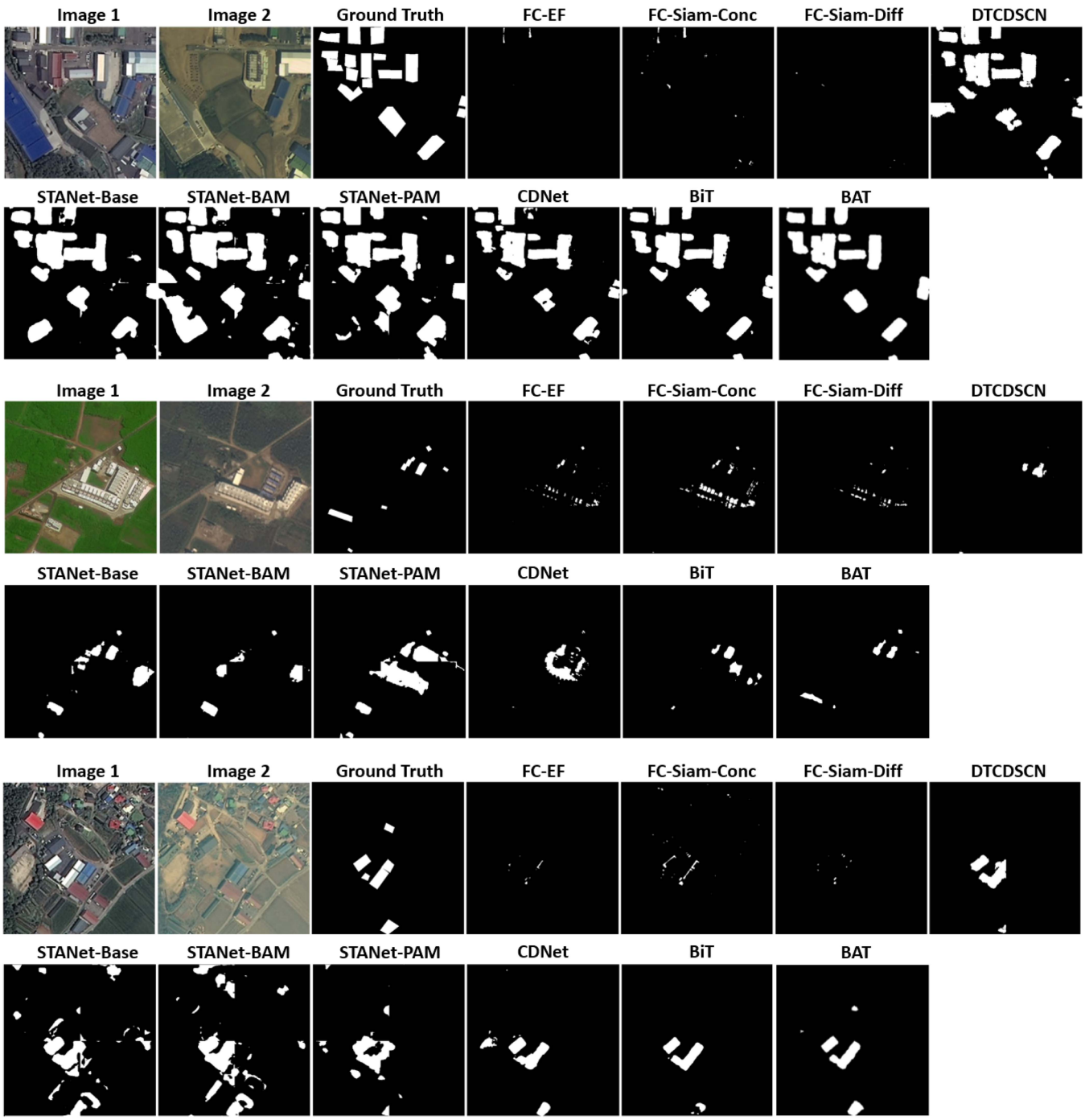

Fig. 8. Comparison of predicted building changes by various methods on the S2looking dataset.

metric is problematic, as a classifier that consistently predicted "no damage" for all images would yield a misleadingly high 75% accuracy. Hence, this metric is reasonable as well as challenging because it heavily penalizes overfitting to overrepresented classes.

As detailed in Section III-C, we utilized the BAT's backbone, complemented by two auxiliary heads (as depicted in the upper portion of Fig. 4) for building extraction. This network was trained using predisaster images and their corresponding building footprint labels, and employed binary crossentropy as loss function. Based on the dataset's spatial resolution, the window size of sequential Swin Transformer blocks within the decoder was set to 16. The encoder in BAT's backbone was initialized with weight values from EfficientNetV2 pretrained on ImageNet [62], and remained frozen for the initial 20 epochs. The network was trained for 120 epochs with a warmup strategy in the first 20 epochs. We applied identical settings for other training details and data augmentation as those used in the preceding BCD experiments.

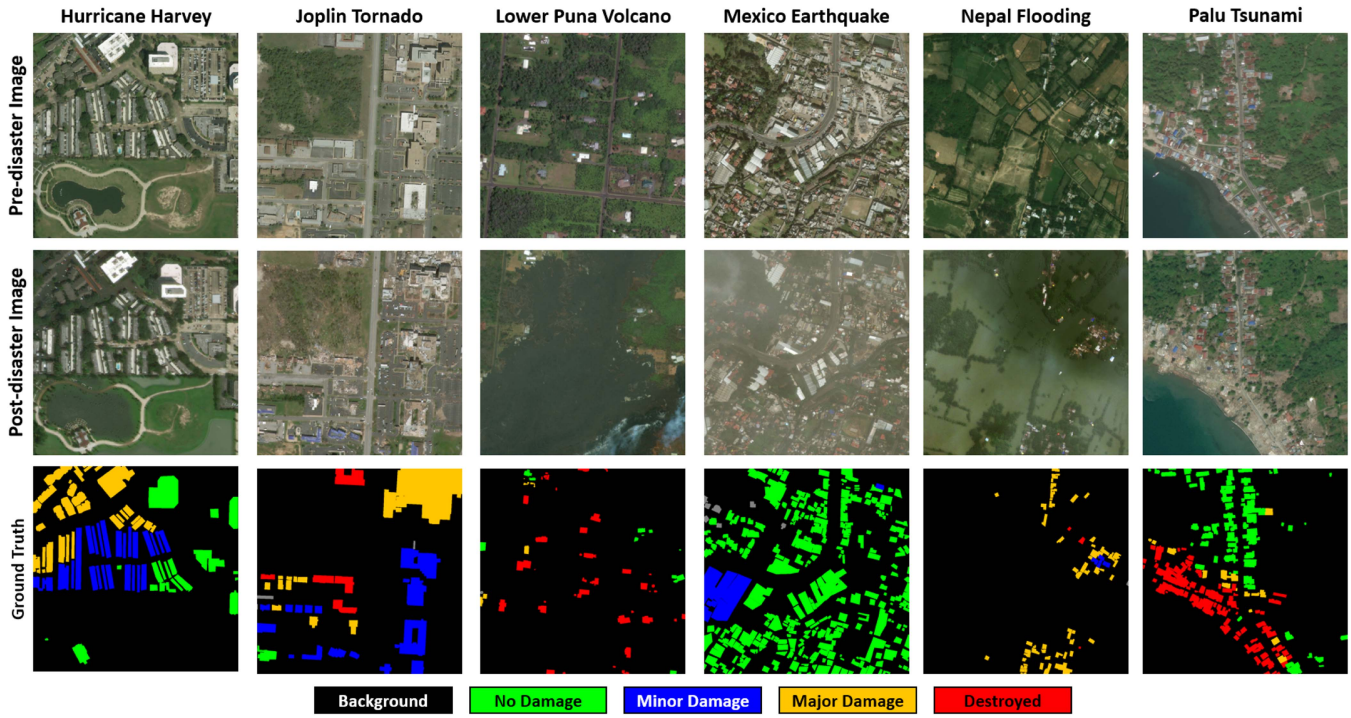

Fig. 9. Predisaster images, postdisaster images, and ground truth of the xBD dataset.

As explained in Section III-C, for the damage classification task, we utilized the pretrained weights from the building extraction task to initialize BAT's backbone. BAT was trained with bitemporal image pairs and their associated building damage labels, utilizing CORN as the loss function. The network was trained for 150 epochs with a warmup strategy in the first 30 epochs. We also applied the same settings for other training details and data augmentation as those employed in the previous BCD experiments.

B. Compared Methods

We make a comparison to representative and state-of-the-art BDA methods, which are described as follows.

Siam-U-Net-Attn [14] not only employs the U-Net architecture for local information extraction but also utilizes skip connections to preserve global information. In addition, it incorporates a self-attention module to capture long-range information spanning the entire image.

RescueNet [9] employs a segmentation head and a change detection head on the dilated ResNet50 backbone. The network is simultaneously trained to fulfill the tasks of building extraction and damage classification.

Dai et al. [15] employs SE-ResNeXt-50 along with an attention gate module in the initial stage for building segmentation. In the subsequent stage, adjustments are made to the network's output layer to accommodate the damage classification task.

Deng and Wang [17] develops a two-stage BDA network based on the U-Net architecture. The initial stage employs an independent U-Net for precise building segmentation, succeeded by a Siamese U-Net dedicated to building damage classification. To enhance the network's capability in segmenting buildings across various scales, the architecture incorporates extra skip connections and asymmetric convolution blocks. In addition, the network employs shuffle attention to focus on the correlation between buildings before and after the disaster.

ChangeOS [16] integrates building localization and damage classification within a cohesive framework using a partial Siamese FCN architecture. This approach facilitates interaction at the feature representation level. ChangeOS offers an advantage by enabling end-to-end training and inference.

BDANet [18] uses a two-branch multiscale U-Net as backbone, where pre and postdisaster images are fed into the network separately. To investigate correlations between these images, a crossdirectional attention module has been introduced. In addition, the application of CutMix data augmentation addresses the difficulties associated with challenging classes.

C. Experimental Results for Stage 1: Building Extraction

As displayed in Table III, the backbone of BAT demonstrates superior performance, yielding the highest building segmentation F1 score ($F1_b$). This superiority is attributed both to the encoder's powerful feature extraction capability and to the decoder's constrained self-attention within window partitions, which are commensurate in size with buildings and their surrounding environments. Because building damage degrees are predicted based on the extracted buildings, enhanced accuracy in building extraction serves as a crucial prerequisite for the subsequent stage.

TABLE III
COMPARISON OF BAT'S BACKBONE WITH OTHER METHODS IN STAGE 1: BUILDING EXTRACTION

Method	$F1_b$
xBD Baseline [7]	0.790
W-Net [12]	0.817
Siam-U-Net-Attn [14]	0.823
Weber et al. [8]	0.835
Improved UNet++ [13]	0.838
RescueNet [9]	0.840
LRBNet [10]	0.850
ChangeOS [16]	0.854
BDANet [18]	0.864
Dai et al. [15]	0.864
Dual-HRNet [19]	0.866
Deng et al. [17]	0.874
BAT's backbone	**0.882**

The bold entities denote the best performance.

TABLE IV
COMPARISON OF BAT WITH OTHER METHODS IN STAGE 2: DAMAGE CLASSIFICATION

Method	No Damage	Minor	Major	Destroyed	$F1_d$
xBD Baseline [7]	0.721	0.024	0.011	0.426	0.030
Weber et al. [8]	0.906	0.493	0.722	0.837	0.697
W-Net [12]	0.884	0.518	0.684	0.855	0.703
Improved UNet++ [13]	0.877	0.513	0.715	0.857	0.707
LRBNet [10]	0.908	0.522	0.706	0.820	0.707
Siam-U-Net-Attn [14]	0.955	0.576	0.744	0.662	0.709
RescueNet [9]	0.885	0.563	0.771	0.808	0.740
Dual-HRNet [19]	0.898	0.590	0.737	0.809	0.741
BiT [11]	**0.971**	0.631	0.723	0.719	0.742
Dai et al. [15]	0.935	0.585	0.755	0.856	0.745
Deng et al. [17]	0.952	0.578	0.754	0.834	0.754
ChangeOS [16]	0.927	0.601	0.742	0.835	0.756
BDANet [18]	0.925	0.616	0.788	**0.876**	0.782
BAT	0.908	**0.643**	**0.803**	0.834	**0.784**

The bold entities denote the best performance.

D. Experimental Results for Stage 2: Damage Classification

Weber and Kané [8] employed a straightforward bitemporal feature fusion strategy via concatenation. Despite this simplicity, as presented in Table IV, their method attains competitive F1 scores for the evident damage levels (*no damage* and *destroyed*). However, distinguishing between the intermediate damage levels (*minor damage* and *major damage*) remains challenging due to subtle visual discrepancies, resulting in significant confusion among these classes for most methods. In contrast, BAT exhibits a more balanced performance compared with other methods across all building damage degrees, resulting in the highest $F1_d$ score (the harmonic mean of classwise damage classification F1 scores). Furthermore, it attains the highest accuracy in recognizing the intermediate damage levels (*minor damage* and *major damage*), which pose a greater challenge compared with the more evident building statuses (*no damage* and *destroyed*).

Fig. 10 displays BAT's predictions for building damage classification across diverse disasters. Due to the strong building extraction capabilities of its backbone, BAT precisely reconstructs boundaries even for very small structures. Some damaged buildings exhibit intact roofs, posing a significant challenge that requires a change detection model to assess damage degrees by considering the environmental factors, such as accumulated water, around the structures. BAT addresses this challenge by confining the CA mechanism to local windows that enclose both the buildings and their surrounding environments. Consequently, BAT accurately identifies the building damages that are subtle from an aerial perspective. In addition, our ordinal regression approach plays a crucial role in categorizing damage severity levels. For instance, it accurately labels buildings with a small portion of roof damaged in hurricanes as *minor damage*, those with a significant portion of roof damaged in hurricanes as *major damage*, and those burnt down in wildfires as *destroyed*.

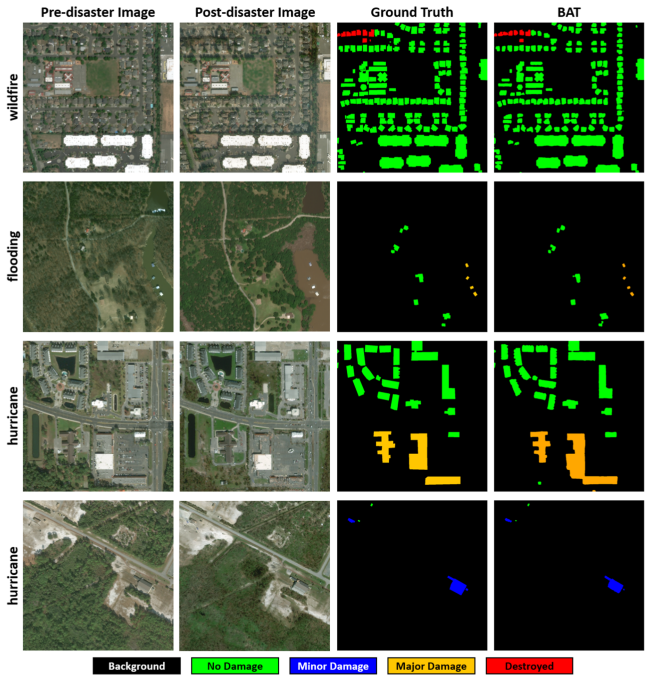

Fig. 10. Predicted building damage classification by BAT.

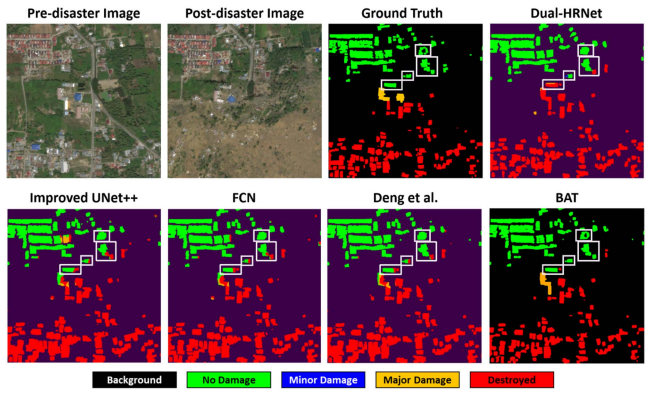

Fig. 11. Comparison of predicted BDAs by various methods on the xBD dataset.

Fig. 11 presents the predicted BDAs by various methods, it is obvious that BAT correctly labels the damage levels of the majority of buildings situated at the boundary between the undamaged and destroyed zones. Predictions from other

TABLE V
COMPARISON OF BAT WITH OTHER METHODS IN TERMS OF THE TOTAL SCORE

Method	$F1_b$	$F1_d$	$F1_s$
xBD Baseline [7]	0.790	0.030	0.260
W-Net [12]	0.817	0.703	0.737
Weber et al. [8]	0.835	0.697	0.741
Improved UNet++ [13]	0.838	0.707	0.746
Siam-U-Net-Attn [14]	0.823	0.714	0.747
LRBNet [10]	0.850	0.707	0.749
Dual-HRNet [19]	0.866	0.726	0.768
RescueNet [9]	0.840	0.740	0.770
Dai et al. [15]	0.864	0.745	0.781
ChangeOS [16]	0.854	0.756	0.786
Deng et al. [17]	0.874	0.754	0.790
BDANet [18]	0.864	0.782	0.806
BAT	**0.882**	**0.784**	**0.813**

The bold entities denote the best performance.

methods often exhibit inconsistent pixelwise labels within individual buildings, which presents a challenge to statistics. We adopt an object-based prediction pipeline where pixelwise labels within individual buildings achieve consensus through majority voting, thereby advancing predictions to an instancewise level.

E. Experimental Results for the Entire Task

As presented in Table V, BAT's superior performance in both the building extraction stage and the damage classification stage leads to the highest total score.

VI. Ablation Analysis

The BAM, ordinal regression approach, and object-based prediction are core components in our pipeline. Consequently, we conducted ablation analysis on the xBD dataset to assess their effectiveness.

A. Ablation Analysis of the BAM

Bi-SRNet [37] also utilizes a CA mechanism (Cot-SR) to model the temporal correlations. Within the CA mechanism of Cot-SR, the encoded features from one temporal image initially attend to specific regions within itself to generate the attention matrice. Subsequently, the generated attention matrice is matrix multiplied with the encoded features from the other temporal image. Its CA mechanism can be represented as $\mathrm{CA}(Q_{\mathrm{pre}}, K_{\mathrm{pre}}, V_{\mathrm{post}})$ and $\mathrm{CA}(Q_{\mathrm{post}}, K_{\mathrm{post}}, V_{\mathrm{pre}})$. In contrast, within the CA mechanism of our BAM, the encoded features from one temporal image initially attend to specific regions in the other temporal image to generate the attention matrice. Our CA mechanism can be represented as $\mathrm{CA}(Q_{\mathrm{pre}}, K_{\mathrm{post}}, V_{\mathrm{post}})$ and $\mathrm{CA}(Q_{\mathrm{post}}, K_{\mathrm{pre}}, V_{\mathrm{pre}})$. To compare the effectiveness of these two CA mechanisms, we only replaced the CA mechanism of BAM with that of Cot-SR and formed a variant network. As demonstrated in Table VI, the CA mechanism of BAM can better model the spatio-temporal semantic relations than that of Cot-SR.

TABLE VI
COMPARISON OF THE CA MECHANISMS OF COT-SR AND BAM

Method	No Damage	Minor	Major	Destroyed	$F1_d$
Cot-SR	0.906	0.617	0.787	**0.836**	0.771
BAM	**0.908**	**0.643**	**0.803**	0.834	**0.784**

The bold entities denote the best performance.

TABLE VII
COMPARISON OF DIFFERENT LOSS FUNCTIONS EMPLOYED BY BAT IN STAGE 2: DAMAGE CLASSIFICATION

Loss Function	No Damage	Minor	Major	Destroyed	$F1_d$
Dice	0.899	0.577	0.766	0.815	0.744
Cross-Entropy	0.901	0.576	0.769	0.827	0.747
CORN	**0.908**	**0.643**	**0.803**	**0.834**	**0.784**

The bold entities denote the best performance.

TABLE VIII
COMPARISON OF SEMANTIC SEGMENTATION AND INSTANCE SEGMENTATION

Method	No Damage	Minor	Major	Destroyed	$F1_d$
Semantic Seg.	0.907	0.627	0.792	**0.840**	0.777
Instance Seg.	**0.908**	**0.643**	**0.803**	0.834	**0.784**

The bold entities denote the best performance.

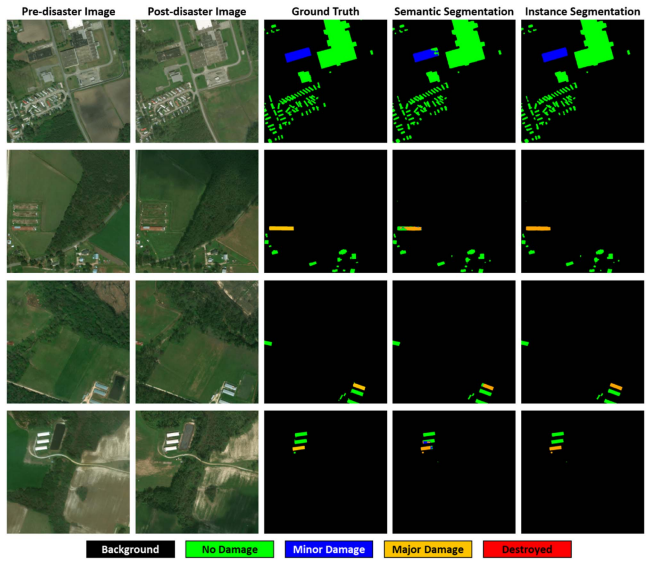

Fig. 12. Comparison of semantic segmentation approach and instance segmentation approach on the xBD dataset.

B. Ablation Analysis of the Ordinal Regression Approach

As shown in Table VII, when compared with conventional multiclass classification loss functions, such as dice and crossentropy, the ordinal regression loss function CORN prominently boosts accuracy for intermediate damage levels (*minor damage* and *major damage*). This demonstrates that BDA is more appropriately treated as an ordinal regression problem rather than a multiclass classification problem.

C. Ablation Analysis of the Object-Based Prediction

As presented in Table VIII, the object-based postprocessing procedure primarily improves accuracy for intermediate damage levels, resulting in F1 score increases of 1.6% for *minor damage* and 1.1% for *major damage*. This is evidenced by Fig. 12, where

inconsistent pixelwise predictions within individual buildings are rectified through a majority voting mechanism.

VII. Discussion

Change detection finds applications in various domains, enhancing decision-making and facilitating the understanding of dynamic processes. With aligned preceding and subsequent images, a key question is how to effectively model the spatio-temporal semantic relations between the bitemporal image pair. In response to this question, we propose a novel CA mechanism that emulates the question-and-answer pattern observed in human interactions. We integrate this CA mechanism into a Siamese network, forming a change detection network named BAT. Subsequently, we applied BAT to BCD task to evaluate its effectiveness. Considering the building scale, irrelevant information removal, and computational efficiency, we introduce the shifted windowing scheme to the CA mechanism, confining it to a defined range. BAT distinguishes itself from existing change detection models, which exclusively support *building change* labels. BAT is capable to harness the valuable information offered by *demolished* and *newly built* labels, enabling it to predict these additional label types.

In contrast to existing BDA methods, which overlook the intrinsic order among ordinal targets and simplistically treat the BDA task as a multiclass semantic segmentation problem, we recognize the significance of ordinal relationships and approach the BDA task as an ordinal regression problem. Therefore, we design an ordinal regression training pipeline and an object-based prediction pipeline for BDA. The effectiveness of our proposed pipelines is not only substantiated through quantitative metrics and ablation analysis, but also evidenced by visual interpretations.

Despite the aforementioned advantages, a limitation worth noting is that our CA mechanism is constrained to analyzing bitemporal image pairs and incapable to model interrelationships within a multitemporal image set. A further limitation to be acknowledged is that the backbone of BAT consumes a majority of the inference time, accounting for 71.1%, whereas the CA mechanism occupies only 28.5% of the inference time. Therefore, the pursuit of real-time processing speed on embedded devices may be advanced by designing a powerful yet lightweight backbone to supplant the existing one. Optimizing BAT with TensorRT may also result in a significant speed boost.

VIII. Conclusion

In this work, we propose a novel CA mechanism to effectively and efficiently model the spatio-temporal semantic relations between a pair of bitemporal remote sensing images. We also recognize the significance of ordinal relationships and approach the BDA task as an ordinal regression problem. Our method achieves state-of-the-art accuracy on two BCD datasets (LEVIR-CD+ and S2Looking), as well as the largest BDA dataset (xBD). This study focuses on the BCD and BDA tasks, our future research direction involves the application, adaptation, and enhancement of the BAT framework for addressing change detection tasks within the domains of agriculture and climate change.

Acknowledgment

Conceptualization, Wen Lu and Minh Nguyen; Data curation, Wen Lu; Formal analysis, Wen Lu; Funding acquisition, Minh Nguyen; Investigation, Wen Lu; Methodology, Wen Lu; Project administration, Minh Nguyen; Resources, Minh Nguyen; Software, Wen Lu; Supervision, Minh Nguyen; Validation, Minh Nguyen; Visualization, Wen Lu; Writing—original draft, Wen Lu; Writing—review and editing, Lu Wei and Minh Nguyen. All authors have read and agreed to the published version of the manuscript. We gratefully acknowledge the support of FHE Electrical Ltd. T/A Kinetic Electrical East Tamaki for their invaluable contributions to this research. Their technical assistance has been instrumental in the successful completion of this study. We also extend our appreciation to Andrew Bryson for his expertise and guidance throughout the project. This work would not have been possible without his assistance.

References

[1] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.

[2] H. A. Afify, "Evaluation of change detection techniques for monitoring land-cover changes: A case study in new Burg EL-Arab area," *Alexandria Eng. J.*, vol. 50, no. 2, pp. 187–195, 2011.

[3] B. Demir, F. Bovolo, and L. Bruzzone, "Updating land-cover maps by classification of image time series: A novel change-detection-driven transfer learning approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 300–312, Jan. 2013.

[4] M. E. Hodgson, B. A. Davis, and J. Kotelenska, "Remote sensing and GIS data/information in the emergency response/recovery phase," *Geospatial Techniques in Urban Hazard and Disaster Analysis*. 2010, pp. 327–354.

[5] K. Yang et al., "Semantic change detection with asymmetric siamese networks," 2020, *arXiv:2010.05687*.

[6] P. Yuan, Q. Zhao, X. Zhao, X. Wang, X. Long, and Y. Zheng, "A transformer-based Siamese network and an open optical dataset for semantic change detection of remote sensing images," *Int. J. Digit. Earth*, vol. 15, no. 1, pp. 1506–1525, 2022.

[7] R. Gupta et al., "xBD: A dataset for assessing building damage from satellite imagery," 2019, *arXiv:1911.09296*.

[8] E. Weber and H. Kané, "Building disaster damage assessment in satellite imagery with multi-temporal fusion," 2020, *arXiv:2004.05525*.

[9] R. Gupta and M. Shah, "RescueNet: Joint building segmentation and damage assessment from satellite imagery," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 4405–4411.

[10] Y. Zhang et al., "An efficient change detection method for disaster-affected buildings based on a lightweight residual block in high-resolution remote sensing images," *Int. J. Remote Sens.*, vol. 44, no. 9, pp. 2959–2981, 2023.

[11] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[12] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From W-Net to CDGAN: Bitemporal change detection via deep learning techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1790–1802, Mar. 2020.

[13] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.

[14] H. Hao et al., "An attention-based system for damage assessment using satellite imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4396–4399.

[15] B. Dai, H. Xiao, and M. Zhang, "A novel two-stage network for building localization and damage level assessment," in *Proc. 6th Int. Conf. Big Data Inf. Analytics*, 2020, pp. 265–269.

[16] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, 2021, Art. no. 112636.

[17] L. Deng and Y. Wang, "Post-disaster building damage assessment based on improved U-Net," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 15862.

[18] Y. Shen et al., "BDANet: Multiscale convolutional neural network with cross-directional attention for building damage assessment from satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5402114.

[19] J. Koo, J. Seo, K. Yoon, and T. Jeon, "Dual-HRNet for building localization and damage classification," 2020. [Online]. Available: https://github.com/DIUxxView/xView2_fifth_place/blob/master/figures/xView2_White_Paper_SI_Analytics.pdf

[20] X. Shi, W. Cao, and S. Raschka, "Deep neural networks for rank-consistent ordinal regression based on conditional probabilities," *Pattern Anal. Appl.*, vol. 26, no. 3, pp. 941–955, 2023.

[21] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[22] Z. Li, C. Tang, X. Li, W. Xie, K. Sun, and X. Zhu, "Towards accurate and reliable change detection of remote sensing images via knowledge review and online uncertainty estimation," 2023. *arXiv:2305.19513*.

[23] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[24] W. Gao, Y. Sun, X. Han, Y. Zhang, L. Zhang, and Y. Hu, "AMIO-Net: An attention-based multiscale input–output network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2079–2093, 2023.

[25] C. Xu et al., "Progressive context-aware aggregation network combining multi-scale and multi-level dense reconstruction for building change detection," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 1958.

[26] C. Zhang et al., "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.

[27] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.

[28] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, 2021, Art. no. 102348.

[29] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2020.

[30] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 484.

[31] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8007805.

[32] Z. Li et al., "Lightweight remote sensing change detection with progressive feature aggregation and supervised attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5602812.

[33] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, 2023.

[34] T. Liu et al., "Building change detection for VHR remote sensing images via local–global pyramid network and cross-task transfer learning strategy," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4704817.

[35] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–41, 2022.

[36] L. Ding, J. Zhang, K. Zhang, H. Guo, B. Liu, and L. Bruzzone, "Joint spatio-temporal modeling for semantic change detection in remote sensing images," 2022, *arXiv:2212.05245*.

[37] L. Ding, H. Guo, S. Liu, L. Mou, J. Zhang, and L. Bruzzone, "Bi-temporal semantic reasoning for the semantic change detection in HR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5620014.

[38] S. Hou et al., "Stable prototype guided single-temporal supervised learning for change detection and extraction of building," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4406622.

[39] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611615.

[40] S. Fang, K. Li, and Z. Li, "Changer: Feature interaction is what you need for change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5610111.

[41] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, "FCCDN: Feature constraint network for VHR image change detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.

[42] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[43] L. Shen et al., "S2Looking: A satellite side-looking dataset for building change detection," *Remote Sens.*, vol. 13, no. 24, 2021, Art. no. 5094.

[44] Z. Zheng, S. Tian, A. Ma, L. Zhang, and Y. Zhong, "Scalable multi-temporal remote sensing change data generation via simulating stochastic change process," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 21761–21770.

[45] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.

[46] R. Hu, G. Pei, P. Peng, T. Chen, and Y. Yao, "Feature difference enhancement fusion for remote sensing image change detection," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, S. Yu et al., Eds., 2022, pp. 510–523.

[47] W. G. C. Bandara and V. M. Patel, "A transformer-based siamese network for change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 207–210.

[48] G. Pei and L. Zhang, "Feature hierarchical differentiation for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 6514105.

[49] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[50] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[51] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, "Triplet-based semantic relation learning for aerial remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[52] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12114–12124.

[53] Z. Zhang, W. Lu, J. Cao, and G. Xie, "MKANet: An efficient network with sobel boundary loss for land-cover classification of satellite remote sensing imagery," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4514.

[54] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.

[55] C. Fiorio and J. Gustedt, "Two linear time union-find strategies for image processing," *Theor. Comput. Sci.*, vol. 154, no. 2, pp. 165–181, 1996.

[56] K. Wu, E. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," in *Proc. Med. Imag.: Image Process.*, 2005, pp. 1965–1976.

[57] M. Liu and Q. Shi, "DSAMNet: A deeply supervised attention metric based network for change detection of high-resolution images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 6159–6162.

[58] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021.

[59] Z. Li, C. Tang, L. Wang, and A. Y. Zomaya, "Remote sensing change detection via temporal feature interaction and guided refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5628711.

[60] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[61] Y. Zhou, C. Huo, J. Zhu, L. Huo, and C. Pan, "DCAT: Dual cross-attention-based transformer for change detection," *Remote Sens.*, vol. 15, no. 9, 2023, Art. no. 2395.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2012.

Wen Lu received the B.Eng. degree in materials physics from Wuhan University of Technology, Wuhan, China, in 2007, the M.Eng. degree in computer science and technology from Hubei University of Technology, Wuhan, in 2023. He is currently working toward the Ph.D. degree in computer and information sciences with the School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand.

His research interests include computer vision, remote sensing, machine learning, and deep learning.

Lu Wei received the B.Eng. degree in software engineering from Wuhan University of Technology, Wuhan, China, in 2006, and the M.Eng. degree in computer technology from Wuhan University, Wuhan, in 2019.

She was a Senior Engineer with Huawei Technologies Corporation, and is currently an Associate Professor with School of Information Science and Engineering, Wuchang Shouyi University, Wuhan. Her research interests include image radiance correction, multisensor image fusion, image quality assessment, and intelligent image processing.

Minh Nguyen received the B.Sc. degree in computer science, and the M.Sc. and Ph.D. degrees in computer vision from The University of Auckland, Auckland, New Zealand, in 2007, 2010, and 2014 respectively.

Since 2017, he has codirected the Centre for Robotics and Vision with Auckland University of Technology (AUT). Currently, he is the Head of the Department of Computer Science and Software Engineering with AUT, leading a team of 40 faculty members. His research interests include computer vision, AI, virtual/augmented reality, computer–human interaction, and knowledge representation and machine learning.