BuildMon: Building Extraction and Change Monitoring in Time Series Remote Sensing Images

Yuxuan Wang, *Graduate Student Member, IEEE*, Shuailin Chen, Ruixiang Zhang, *Student Member, IEEE*, Fang Xu, Shuo Liang, Yujing Wang, and Wen Yang, *Senior Member, IEEE*

***Abstract*—Building extraction and change monitoring in remote sensing (RS) imagery play pivotal roles in various applications, including urban planning, disaster management, and infrastructure monitoring. While significant progress has been made in single and bitemporal RS images, effectively harnessing the rich temporal information of time series RS images remains a challenge. Time series RS images offer an extended temporal span for monitoring dynamic changes in building instances. However, they often exhibit noticeable appearance discrepancies and feature variations, presenting substantial obstacles to effective multitemporal information aggregation. To address these challenges, we introduce a Building Extraction and Change Monitoring Network (BuildMon), which jointly explores the segmentation masks, location tracking, and construction status of building instances. Our approach incorporates a spatial-temporal transformer to model relationships between images at different time spans. The windowed attention module within it can capture spatial-temporal context for a larger scope of feature aggregation. For enhancing the performance on both tasks, we adopted ground truth masks and semantic change information together as supervisory signals for BuildMon. This is complemented by the specially designed change-guided loss function, which specifically highlights regions of change and assigns targeted weights to building areas within the imagery. To validate the effectiveness of our method, we conduct comprehensive experiments on the SpaceNet 7 dataset. The results showcase the state-of-the-art performance of our approach, achieving mIoU and SpaceNet Change and Object Tracking metrics of 67.90 and 39.73, respectively.**

***Index Terms*—Building extraction, change-guided loss, change monitoring, spatial-temporal (ST) transformer, time series images.**

I. Introduction

BUILDINGS, as significant human-made structures, are crucial targets in Earth observation. The locations, shapes, construction states, and demolition times of buildings hold immense value, supporting various applications such as population estimation [1], urban planning [2], and disaster risk assessment [3]. Remote sensing (RS) images provide essential visual characteristics of buildings, including color and texture, crucial for interpretation tasks like building detection, segmentation, and change detection. However, single or bitemporal images lack sufficient temporal dimension information for identifying building changes over time. Fortunately, the number of satellites forming constellations with consistent revisit cycles has increased over decades, offering multitemporal observation capabilities [4]. This advancement has led to studies beyond single and paired images, exploring dynamic data such as time series images [5] and videos [6]. They are often characterized by longer spans and higher temporal resolutions, facilitating automated extraction of changing trends and temporal points of change occurrence. Given the construction and demolition of buildings typically span several days or months, utilizing time series images with longer temporal resolutions is more appropriate for revealing the static and dynamic characteristics of buildings [7]. To delineate the positions, shapes, and construction status of buildings, solely extracting buildings from each image or detecting changes from bitemporal image pairs is not enough. Change detection [8], [9], [10] aims at deciding whether new buildings are constructed or existing buildings are demolished, without considering the distribution of the buildings. It only predicts the change information but does not output all building instances. To overcome the dilemma, this article focuses on simultaneously handling building extraction and building change monitoring in Time Series Remote Sensing Images (TSRSI).

TSRSI, akin to time-varying visual information, can be viewed as a video with very low frame rates, enabling techniques of video segmentation [11], [12], [13], [14], [15] to be directly applied for dynamic building extraction. Nevertheless, compared to TSRSI, video frames are usually captured in a much shorter period, so there is little appearance discrepancy and semantic changes across the temporal images. Conversely, TSRSI exhibit varying appearances due to factors like seasons, weather, and shadows, despite acquiring in the same physical areas. In addition, targets in adjacent video frames exhibit relative motion, often described and modeled through techniques like optical flow [16]. As the buildings are stable artificial land objects with no movement, there is no relative motion between buildings in adjacent images.

Manuscript received 5 February 2024; revised 13 April 2024; accepted 13 May 2024. Date of publication 24 May 2024; date of current version 14 June 2024. This work was supported in part by the NSFC Regional Innovation and Development Joint Fund under Grant U22A2010, in part by the NSFC general program under Grant 62271355, and in part by the CETC key laboratory of aerospace information applications under Grant SKX222010021. *(Yuxuan Wang and Shuailin Chen contributed equally to this work.) (Corresponding author: Wen Yang.)*

Yuxuan Wang, Shuailin Chen, Ruixiang Zhang, and Wen Yang are with the School of Electronic Information, Wuhan University, Wuhan 430072, China (e-mail: wyx2017@whu.edu.cn; slchen1997@whu.edu.cn; zhangruixiang@whu.edu.cn; yangwen@whu.edu.cn).

Fang Xu is with the School of School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: xufang@whu.edu.cn).

Shuo Liang and Yujing Wang are with the 54th Research Institution of CETC, Shijiazhuang 050081, China (e-mail: liangs@nuaa.edu.cn; afreeboat@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3404781

For analysis of TSRSI, traditional methods are often based on the perspective of statistics and adopt handcrafted operators. Due to the semantic complexity in RS images, traditional methods are inadequate for handling both tasks of building extraction and change monitoring in TSRSI. With the advancement of deep learning technology and its success in visual tasks, RS intelligent interpretation has also benefited from it. However, most existing deep learning-based methods on TSRSI [17], [18], [19], [20], [21] are designed for classification of land cover and agricultural crops. They often assume that spatial features alone are insufficient for classification tasks, thus they utilize multitemporal features as a complement. Nevertheless, these methods often assume that the semantic categories of observed ground objects do not change across the time dimension, neglecting the dynamic building distribution throughout the time series. To explore how to associate the time series images for building extraction and change monitoring, in this article we propose to predict results for each temporal image, taking into consideration temporal variations within the time series.

Motivated by the above analysis, we propose a novel Building Extraction and ChAnge Monitoring network (BuildMon), addressing building extraction and change monitoring tasks simultaneously in TSRSI. BuildMon integrates a spatial-temporal (ST) Transformer to enhance temporal context and aggregate attention in spatial and temporal dimensions. To mitigate appearance variation, we design a change-guided loss function and implement a simple yet effective modification, enhancing the network's change detection ability. In summary, our contributions are as follows.

1) We propose a novel framework BuildMon to handle both tasks of building extraction and change monitoring in TSRSI. Our method jointly explores the positions, shapes, location tracking, and change monitoring of building instances.
2) We introduce a transformer-based ST context module and integrate it into the segmentation network. This component utilizes ST attention to establish relationships between different frames, enhancing the feature representation for TSRSI.
3) We devise a change-guided loss for the learning process, serving as auxiliary supervision. This loss function promotes temporal consistency in the feature space, benefiting both building extraction and change monitoring tasks.
4) We conduct extensive experiments on the SpaceNet 7 Dataset, demonstrating the superiority of our BuildMon over other approaches, achieving state-of-the-art performance.

The rest of this article is organized as follows. In Section II we introduce the related work of our research. In Section III we describe the proposed BuildMon in detail. The experimental results are reported in Section IV. Finally, some discussions are presented in Section V. Finally, Section VI concludes this article.

II. Related Work

Our research aims at extracting buildings and monitoring their changes in TSRSI. The proposed method handles both tasks simultaneously and explores the inter-image relationship by modeling temporal attention. The following subsections will introduce the literature on the two tasks, i.e., building extraction and change detection in RS and traditional analysis methods of TSRSI.

A. Building Extraction

The task of building extraction revolves around predicting the shape and location of buildings in RS images. Mainstream methods address tasks in three key aspects, catering to varying levels of requirements.

Semantic segmentation methods: These methods generate a pixel-level classification map for each input image and divide it into building and nonbuilding areas [22], [23], [24], [25], [26], [27]. MAP-Net [25] learns spatial localization-preserved multiscale features through a novel multiple attending path strategy, alleviating the restriction on feature extraction from the receptive field.

Instance segmentation methods: In addition to predicting semantic maps, instance segmentation methods assign individual masks for each building [28], [29], [30], [31], [32]. Considering the offsets between buildings' footprints and roofs in off-nadir aerial imagery, LOFT [30] adopts a multihead framework and introduces an orientation-based feature augmentation for building footprint instance segmentation.

Polygon generation methods: These methods aim to extract polygon contours of buildings, providing more precise geometry properties for subsequent applications [33], [34], [35], [36]. PolyBuilding [36] introduces a polygon Transformer to generate the bounding boxes and polygons simultaneously for the building instances. Its corner classification head reduces the redundancy of the vertex and enhances the building polygon regularity.

The adaptation of Transformer-based architectures, such as the Vision Transformer (ViT) [37] makes significant success in visual perceptual tasks due to its strong representation learning ability to capture long-range dependencies in images. For the segmentation task, ViT can be adopted as a backbone for feature extraction. Moreover, the self-attention idea can also be specifically applied to the mask prediction process. Mask2Former [38] introduces masked attention for constraining cross-attention within predicted mask regions and proposes a unified framework for panoptic segmentation, instance segmentation, and semantic segmentation.

Building semantic segmentation identifies building regions, distinguishing them from trees or roads in the background. Instance segmentation predicts an index for each building with a unique identifier (ID), facilitating the quantification of buildings' distribution and density. Polygon extraction offers an exact delineation of building locations and geometric shapes, crucial for accurate map construction. However, existing methods for these tasks predominantly focus on static building extraction, with a limited exploration into the changing trends of buildings based on revisiting time-series imagery. In this article, we address this gap by extracting multitemporal features in a global attention manner and integrating them into a building segmentation framework. Our objective is to simultaneously extract buildings and monitor their dynamic changes over time.

B. Building Change Detection

By comparing two images captured at different times, building change detection can extract the changed regions in the scene. According to the extracted features for change detection, the existing methods can be mainly categorized into two streams. Some methods adopt handcrafted features [39], [40], [41], [42] and some others leverage the deep neural network (DNN) to represent and delineate the scenarios implicitly. With the advancement of deep learning technology, learning-based methods [43], [44], [45], [46], [47], [48], [49], [50] are demonstrated to achieve a state-of-the-art (SOTA) performance on building change detection. DTCDSCN module [43] adopts a multitask learning (MTL) framework, simultaneously accomplishing both change detection and semantic segmentation. This strategy can facilitate learning more discriminative object-level features. For the dense buildings in RS images, the boundaries are intricate for the DNN due to detail information loss in some down-sampling layers. To emphasize the edge regions of the buildings, EGRCNN [44] incorporated both discriminative information and edge structure prior in one framework, resulting in better preservation of the original structure in the predicted changed regions. To enhance the discriminative features close to building boundaries, Zhang et al. [46] presented a novel method based on contrastive learning to exploit the temporal-spatial correlation in the neighborhood of the edge.

In contrast to semantic segmentation, the task of change detection for buildings focuses solely on predicting the regions that have changed. The output of building change detection reflects where a building has been constructed or demolished, omitting information about the distribution of buildings. To address this limitation, some methods [43], [45], [51], [52] adopt an MTL framework to simultaneously perform semantic segmentation and change detection for buildings. However, most of them are primarily designed to handle bi-temporal image pairs, limiting their capability to monitor the changing trends of buildings and identify the specific time point of change. In addition, when using semantic segmentation as the associated task for change detection, obtaining the changing patterns of individual buildings becomes challenging. To overcome these challenges, our proposed method generates change patterns for building instances from a group of TSRSI. To optimize computational efficiency, we adopt a postprocessing step following building semantic segmentation, eschewing the conventional two-stage instance segmentation. Consequently, our method, BuildMon, is capable of predicting the locations, shapes, and changing states for building identities simultaneously.

C. Time Series Remote Sensing Image Analysis

Time series RS image analysis includes various tasks like classification, regression, clustering, and change detection. Traditional analysis methods often adopt handcrafted operators and extract information from the statistics of TSRSI [53], [54], [55], [56]. Ruiz et al. [57] proposed an algorithm based on cumulative sum and statistical analysis, which uses dense Sentinel-1 image time series to achieve continuous and near-real-time monitoring of forests. Csillik et al. [58] compared the effects of various object-based dynamic time warping [59] crop classification methods, which overcomes the shortage of normalized difference vegetation index [60] and is more flexible. By leveraging the intratemporal features of TSRSI, the Temporal Clustering Matching algorithm [56] considers the building change detection in a semiautomatic fashion. It requires manual building mask annotation of the changed building in the first temporal, and then using the statistical difference to determine whether and when the building has been constructed or demolished.

Application of DNN models in RS images is becoming increasingly mature and gradually replacing traditional rule-based methods in the last decades [61]. Based on the temporal self-attention mechanism, Russwurm et al. [62] proposed to directly extract and classify vegetation from raw TSRSI, without data preprocessing or manually designed features. This verifies that the self-attention mechanism can effectively select temporal and spectral bands beneficial to the task, and suppress noise and redundant information. Garnot et al. [19] designed a lightweight temporal self-attention network, which replaces the projection of the input data with a learnable parameter as the attention query. For the panoptic segmentation, Garnot et al. [20] mixed convolution and self-attention mechanism, where the convolution module extracted features for each image separately, and then the self-attention module sums up the features from different temporal phases. In addition, Tarasiou et al. [21] argued that temporal information is more important than spatial information in agricultural remote sensing, so they introduce a temporal-spatial vision Transformer, where the temporal attention and the spatial attention are applied to the input sequentially.

Nevertheless, existing methods tend to overlook changes in the temporal dimension, often aggregating multitemporal information solely for common semantic predictions. These approaches operate under the assumption that the distribution of ground objects and the category of a particular region (e.g., crops and forests) remain invariant over time. In contrast, our BuildMon tackles both the building extraction and change detection in TSRSI while preserving the temporal dimension in the output predictions. Moreover, the holistic consideration of the appearance discrepancy in unchanged regions renders the model more adaptable for a wide range of TSRSI applications.

III. Methodology

In this section, we first introduce the overall framework and the pipeline of our BuildMon, and then describe the module designs and the learning schedule in detail.

A. Overall Architecture

As illustrated in Fig. 1, we choose to first accomplish a semantic segmentation and then instantiate the buildings using a post-process method. An encoder-decoder structure is employed for the segmentation stage. The pipeline of this framework starts with the feature extraction of TSRSI by a backbone based on convolutional neural network (CNN). Then the proposed ST

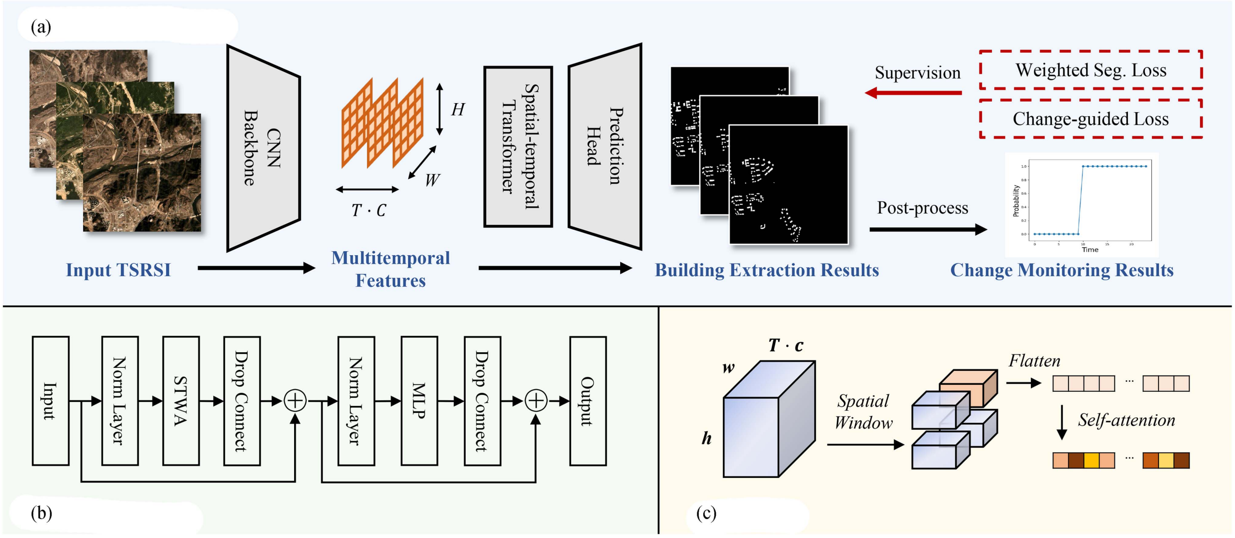


Fig. 1. (a) Displays the overall architecture of the proposed BuildMon. It first adopts an encoder-decoder framework for building semantic segmentation. Then the post-processing strategy generates the building instances from the segmentation results and assigns address IDs. Finally, the change trends of the buildings can be acquired from the TSRSI. (b) Illustrate the structure of the ST transformer, containing two residual connections. STWA module is embedded in the first connection. (c) Shows the steps in STWA. The input features are windowed in spatial dimension and calculated for self-attention after being flattened as vectors.

Transformer is employed as the neck to fuse the ST information and enhance the feature consistency of the same building across different temporal images. After that, the decode head predicts the building masks for each image separately. Note that the cross-temporal connections only exist within the ST Transformer. Finally, we apply the ST collapse [63] postprocessing algorithm to obtain the building instances and assign the address IDs for them at each temporal image. To effectively supervise the training of BuildMon and suppress the interimage appearance discrepancy, two customized loss functions are designed for the segmentation predictions.

B. Spatial-Temporal Transformer

In TSRSI, the unchanged regions in different temporal images usually have the same semantics, thus the context in the time dimension can be leveraged for a multitemporal feature aggregation. Moreover, the neighboring regions are also closely associated and their relationship should be under consideration. Inspired by video semantic segmentation methods [20], our method employs a ST Transformer as the neck to fully exploit the ST context in the time series images.

Given a group of co-registered TSRSI I_i $(i = 1, 2, \ldots, T)$, where T is the length of the series, a CNN-based backbone extracts the feature maps $F \in \mathbb{R}^{T\times C\times H\times W}$ in different levels. Note that in our method, the input can be a selected sub-series from the original TSRSI with an arbitrary length, so T is the input image series in the following paragraph. For the multilevel features, we align their scales with a bilinear upsampling operator and concatenate them in channel dimension, thus their heights and widths can be consistent.

Different from the moving targets in the videos, buildings in the case of co-registered TSRSI maintain the same location across different temporal images. Furthermore, since the semantic information dramatically varies in a long-distance scope, modeling the global spatial cross-attention is unnecessary for building extraction in a large scenario. On the contrary, due to the seasonality cycle and the randomness of imaging conditions, features at the same spatial positions in any two temporal images may exhibit a strong correlation. This correlation will not become much lower as the time interval increases. Thus we propose a window-spatial and temporal self-attention module in the ST Transformer. After a normalization layer, the feature maps are spatially windowed, i.e., chunked in the height and width dimension. Each 3D window contains a feature tensor with the shape of $[T, C, h, w]$, where h and w denote the shape of the spatial window. In the perspective of self-attention, there are $N(= T \times h \times w)$ tokens and each of them is a vector with a length of C. Then, the ST attention can be calculated as (1).

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{\text{model}}}}\right)\boldsymbol{V} \tag{1}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{d_{\text{model}}\times N}$ represent the query, key, and value sequences, respectively. $\boldsymbol{K}^T$ is the transpose of $\boldsymbol{K}$. d_{model} denotes the dimension of the key, and N represents the length of the token sequence. The scaling factor of attention weights, $\sqrt{d_{\text{model}}}$, is introduced to counteract the potential issues arising from large dimensions of query and key.

In self-attention, query, key, and value are obtained by linearly projecting the input features, as shown in the following equations:

$$\boldsymbol{Q} = \boldsymbol{W}^Q\boldsymbol{X} \tag{2}$$

$$\boldsymbol{K} = \boldsymbol{W}^K\boldsymbol{X} \tag{3}$$

$$\boldsymbol{V} = \boldsymbol{W}^V\boldsymbol{X} \tag{4}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^K$, and $\boldsymbol{W}^V \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are the projection matrices for query, key, and value, respectively, and $\boldsymbol{X} \in \mathbb{R}^{d_{\text{model}} \times N}$ represents the input sequence.

To handle complex attention patterns, we adopt the multihead attention strategy in the Transformer [64]. This mechanism allows for the projection of query, key, and value into different subspaces. The attention is then calculated separately in these subspaces, and the attention results from different heads are concatenated. This process can be expressed as follows:

$$\text{MultiHead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(h_1, \ldots, h_n)\boldsymbol{W}^O, \tag{5}$$

$$h_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q, \boldsymbol{K}\boldsymbol{W}_i^K, \boldsymbol{V}\boldsymbol{W}_i^V). \tag{6}$$

In (5) and (6), $\boldsymbol{W}_i^Q, \boldsymbol{W}_i^K$, and $\boldsymbol{W}_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$ represent the projection matrices for query, key, and value in the ith head, respectively. The parameter d_k denotes the dimension of each head while n is the number of heads. Finally, $\boldsymbol{W}^O \in \mathbb{R}^{hd_k \times d_{\text{model}}}$ represents the projection matrix after the concatenation of attention results from different heads.

The outputs of the attention module are passed through a normalization layer and a Multilayer Perceptron as the feed-forward network. Residual connections are applied in the ST Transformer for faster optimization, resulting in facilitating the training of the deeper network. The window-spatial and temporal attention can fully aggregate the long-term temporal context in TSRSI and consider the short-term spatial context simultaneously, enhancing building segmentation for each temporal image.

C. Instance Normalization

Due to the seasonal change and the variation of illumination conditions, there is often a significant appearance discrepancy between different temporal images in TSRSI, which interferes with the calculation of temporal self-attention. To address this issue, we propose a simple but effective modification in the Transformer. By employing data normalization methods, such as Batch Normalization (BN) [65], Layer Normalization (LN) [66], and Instance Normalization (IN) [67], the model's dependence on the data distribution can be alleviated. Among these methods, IN is particularly effective in normalizing statistics that carry image style information, which is widely used in image style transfer [67], [68], [69]. In our work, we replace BN in the backbone and LN in the ST Transformer with IN to enhance the model's integration of temporal information.

The calculation process for the normalization can be summarized using (7)

$$\boldsymbol{Y} = \frac{\boldsymbol{X} - \mu}{\sqrt{\sigma^2 + \epsilon}} \odot \gamma + \beta \tag{7}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the input and output features of the normalization layer, respectively. γ and β are the learnable scaling and offset parameters. The small constant ϵ is added to the denominator to prevent division by zero, and $\odot$ denotes element-wise multiplication. μ and σ are the mean and variance of the input features. IN calculates μ and σ separately for each sample and each channel. In this case, the mean and standard deviation statistics have dimensions of $N \times C \times 1 \times 1$.

Furthermore, replacing the IN layer may potentially disrupt the original network design. To preserve the harmonious design and retain most of the pre-trained weights, we only replace the first normalization layer in the backbone and the ST Transformer, while keeping the other normalization layers unchanged.

D. Optimization Schedule

Due to the category imbalance between changed and unchanged regions in the dataset, the model often treats the entire temporal sequence as a singular category. Consequently, the temporal attention mechanism experienced a degradation. Moreover, the unchanged regions in TSRSI may possess various appearances, leading to discrepant prediction probabilities for the same regions. To tackle these two issues, we devise a change-guided loss function to further enhance the ST Transformer. This design aims to accurately identify when a change occurs and keep the invariance of the unchanged regions among the temporal images.

The overall loss function comprises two components: a weighted building segmentation loss and the change-guided detection loss, as illustrated by (8).

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda \mathcal{L}_{\text{cd}} \tag{8}$$

where λ represents the weight assigned to the change-guided loss $\mathcal{L}_{\text{cd}}$, which is defined as follows:

$$\mathcal{L}_{cd} = \begin{cases} \left\|\boldsymbol{P}_i^1 - \boldsymbol{P}_i^2\right\|_1 & \text{if } \boldsymbol{Y}_i^1 = \boldsymbol{Y}_i^2 \\ -\omega_{\text{cd}} \left\|\boldsymbol{P}_i^1 - \boldsymbol{P}_i^2\right\|_1 & \text{if } \boldsymbol{Y}_i^1 \neq \boldsymbol{Y}_i^2. \end{cases} \tag{9}$$

$\boldsymbol{P}_i^1$ and $\boldsymbol{P}_i^2$ represent the predicted probabilities of the ith pixel in two selected temporal images. ω_{cd} is the weight for changed area. This loss function can strengthen the robustness of the network's predictions for unchanged regions and enhance its attention on changing regions. Different from the bitemporal change detection loss, we calculate the change-guided loss for every two temporals in the TSRS. $\boldsymbol{Y}_i^1$ and $\boldsymbol{Y}_i^2$ represent the corresponding ground truth and the symbol $||\cdot||_1$ denotes the L_1 norm.

The building segmentation loss $\mathcal{L}_{\text{seg}}$ is a standard binary cross entropy loss, to highlight the significance of changed buildings, the segmentation loss also incorporates a weighted scheme for changed regions

$$\mathcal{L}_{\text{seg}} = -\sum_{i=1}^{HW} \boldsymbol{\omega}_i \sum_{c=1}^{C} \boldsymbol{Y}_i \log \boldsymbol{P}_i \tag{10}$$

where $\boldsymbol{Y}_i$ and $\boldsymbol{P}_i$ denote the label and predicted result for the position i, respectively. Furthermore, $\boldsymbol{\omega}_i$ denotes the weight assigned to the pixel at position i.

E. Postprocessing

The winning solution of the SpaceNet 7 competition [63] introduced a postprocessing technique, named spatiotemporal collapse, to convert converts building semantic segmentation results into building instance polygons by effectively leveraging observed change patterns in TSRSI. Building upon this, our approach incorporates the temporal information within the training stage into the generation of building instances.

The algorithm operates under two key assumptions: First, considering buildings as stable man-made structures, it assumes that they generally remain unchanged over short periods. Second, it assumes that different building instances in the images are separated by at least one pixel. The first assumption aligns with common real-world scenarios, simplifying the postprocessing task and ultimately enhancing result accuracy. The second assumption facilitates the transformation of the building instance segmentation problem into a semantic segmentation problem. With these assumptions in mind, the algorithm unfolds in two distinct steps: temporal collapse and spatial collapse.

Temporal Collapse: Consider a TSRSI with N temporals, the network outputs a predicted probability map $\boldsymbol{P}_t$ for *t*th temporal image. According to the first assumption, we can compress the temporal dimension and predict the position and shape of each building only once, then the building segmentation map $\boldsymbol{S}$ is obtained. Formally, we can express this temporal collapse process as (11).

$$\boldsymbol{S} = \frac{\sum_t \boldsymbol{P}_t \cdot \mathbb{I}(\boldsymbol{P}_t \geq \alpha)}{\max(\sum_t \mathbb{I}(\boldsymbol{P}_t \geq \alpha), \epsilon)} \tag{11}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which evaluates to 1 when the condition is true and equals to 0 otherwise. ϵ is introduced to prevent division by zero errors, and α represents the probability threshold.

This step combines the prediction results across the temporal dimension, compensating for potential inaccuracies in individual images and resulting in more accurate building boundaries. Once the building segmentation probability map for the entire time series image is obtained through (11), an improved watershed algorithm [70] is utilized to convert them into building instances.

Spatial Collapse: Once the building address IDs are obtained, the next step involves determining the time point when the buildings appear. To achieve this, we first calculate the pixel-average probability for each building instance, which can be expressed as

$$T_l^t = \frac{1}{|\{(i) \in \mathbb{G}_l\}|} \sum_{i \in \mathbb{G}_l} \boldsymbol{P}_i^t \tag{12}$$

where $\mathbb{G}_l$ represents the lth building instance, $|\cdot|$ denotes the cardinality of a set, and T_l^t denotes the probability of the lth building instance in the tth temporal.

Then a moving average difference method [71] is employed for correction, which involves the following steps: First, the forward and backward moving average sequences are computed as (13).

$$\overset{\leftarrow}{T_l^t} = \frac{1}{t} \sum_{k=1}^{t} T_l^k, \quad \vec{T}_l^t = \frac{1}{N_t - t + 1} \sum_{k=t}^{N_t} T_l^k. \tag{13}$$

If $\max_t(\vec{T}_l^{t+1} - T_l^{t\leftarrow}) < \gamma_d$, there is no change in the building state within the time series. In this case, we assign 0 to indicate the absence of the building and 1 to indicate its presence.

This dual-step process ensures a robust conversion from semantic segmentation to building instance masks, capitalizing on the stability of buildings and the spatial separation between distinct instances. Through this postprocessing method, we can obtain the address IDs of all buildings, as well as the changing state of building instances in multitemporal images.

IV. Experimental Results and Analysis

A. Dataset

In this article, we choose to conduct experiments on the SpaceNet 7 dataset [7] to evaluate the effectiveness of BuildMon. The proposed approach addresses two primary subtasks: 1) building extraction; and 2) change monitoring in TSRIS. While there exist several large-scale public datasets dedicated to building extraction, such as WHU building [72], Inria [73], and SpaceNet series [7], [74], [75], and some solid datasets for building change detection, such as LEVIR-CD [76] and BANDON [77], datasets specifically designed for both building extraction and change monitoring are scarce. Consequently, the SpaceNet 7 dataset [7] is chosen since it comprehensively fulfills the requirements of our task.

The SpaceNet 7 dataset comprises 101 TSRSI collected by Planet Labs' Dove constellation between 2017 and 2020, which are preprocessed with an orthorectification. A time series consists of 18–26 temporal images, with a monthly imaging frequency. The size of each image is 1024×1024 pixels and the resolution is about 4 m, covering an approximate real geographic area of 18 km^2. The dataset contains more than 11 million building instances in total, and the imaging period is in line with the typical time scale of urban development. It captures geographic regions with diverse characteristics and exhibits urbanization changes over two years.

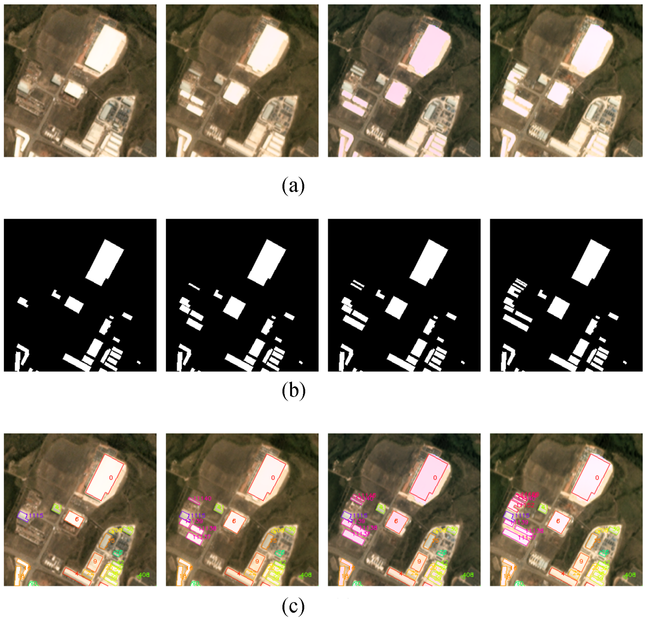
Each building instance in the SpaceNet 7 dataset is annotated with a polygon describing its shape and location, as shown in Fig. 2, along with a unique ID. The same building retains the same identifier across different temporal images, enabling cross-temporal tracking and change analysis of buildings within the dataset.

B. Evaluation Metrics

To provide a comprehensive assessment of the proposed method in building extraction and change monitoring, we employ both pixel-level evaluation metrics, commonly used in semantic segmentation, and instance-level tracking metrics [7]. The precision, the recall, and the Intersection over Union (IoU) are utilized for pixel-level evaluation. In addition to overall accuracy, we also consider adopting the Boundary IoU (BIoU) metric [78] to evaluate boundary accuracy. It is defined as the following equation:

$$\text{BIoU} = \frac{|(G_d \cap G) \cap (P_d \cap P)|}{|(G_d \cap G) \cup (P_d \cap P)|} \tag{14}$$

where G denotes the ground truth, P denotes the prediction, and G_d and P_d represent the boundary regions. To define the boundaries, we follow the recommended practice in [78] to expand the



(a)

(b)

(c)

Fig. 2. Sample of SpaceNet 7 dataset [7]. A group of TSRSI is captured in an identical region at different times. The annotations of each image contain the positions and shapes of all the building instances. (a) Time series remote sensing images. (b) Building mask ground truth. (c) Overlayed image and instance annotation.

regions outward by d pixels from the original boundary. In this case, d is set to 5% of the image diagonal length.

For building instance tracking and change monitoring, the SpaceNet Change and Object Tracking (SCOT) metric [7] is introduced as a comprehensive evaluation metric. It consists of two components: 1) the tracking term; and 2) the change detection term. The tracking term assesses the model's ability to accurately track the same building across different temporal images. It takes into account the number of matching items and nonmatching items, and this metric can be formulated as follows:

$$F_{\text{track}} = \frac{TP_s}{TP_s + \frac{1}{2}(FP_s + FN_s)} \tag{15}$$

where TP_s represents the number of matching pairs, FN_s is the number of unmatched instances in annotation, and FP_s is the number of unmatched instances in prediction. The change detection term focuses on evaluating the model's ability to correctly identify newly appearing buildings. The change detection term only considers newly appearing instances which never present in the previous temporal images, calculated by an F1-score as (16).

$$F_{\text{change}} = \frac{TP_{\text{new}}}{TP_{\text{new}} + \frac{1}{2}(FP_{\text{new}} + FN_{\text{new}})} \tag{16}$$

where the subscript new denotes newly appearing instances.

The SCOT score is computed as a weighted harmonic average (β score) of the tracking term score and the change detection term score, as shown in (17)

$$F_{\text{scot}} = \frac{(1+\beta^2) \cdot F_{\text{change}} \cdot F_{\text{track}}}{\beta^2 \cdot F_{\text{change}} + F_{\text{track}}}. \tag{17}$$

C. Implementation Details

Our experiments are all implemented based on the PyTorch framework on four NVIDIA TITAN V GPUs. The training set comprises 50 time series, while the testing set consists of ten time series. To get a tradeoff between the performance and the computational efficiency, we employ HRNetV2p-W18 [79] with ImageNet pretrained weights as the backbone. For the training process, the following parameter settings are adopted: 40 000 iterations training, the batch size is set to 4, clip length for a batch is 4, AdamW optimizer [80] is employed, the learning rate is 0.005, and decay is set to 0 according to the polynomial learning rate policy. The other optimizer parameters are set to their default values. Since the Transformer and CNN have different optimal learning rates, we set the learning rate of the Transformer to 0.1 times the global learning rate. Uniform cropping is performed to obtain image patches with the size of 512 × 512.

D. Compared With the State-of-the-Art

Given that few studies focus on simultaneously handling building extraction and change monitoring in TSRSI, we compare the proposed BuildMon with existing approaches in the following three categories.

1) *Segmentation methods:* We employ three methods for comparison, i.e., U-net [22], DeepLabv3 [23], and SegFormer [24]. The segmentation predictions of them are postprocessed to calculate the SCOT-related metrics.
2) *Multitask-learning methods:* MTL-U-Net [51] and TBFFNet [52]. They have two task heads that enable joint learning for building extraction and change detection. Given they cannot provide the instance mask for each building, the prediction results also need to be postprocessed.
3) *Video semantic segmentation methods:* CFFM [13] and MRCFA [14] are adopted since they can process the multitemporal image series which is similar to the video frames. To ensure fair comparisons, we employ the same postprocessing algorithm in conjunction with the compared methods.

Table I presents the quantitative comparison results with other state-of-the-art methods. Metrics in three aspects are employed for the evaluation: efficiency, pixel-level accuracy, and instance-level accuracy. We include the model scale, i.e., number of weight parameters (M), computational complexity FLOPs (G), and inference time (ms). Besides the traditional IoU-related metrics, we considered the SCOT-related metrics like SCOT, tracking score, and change score. The proposed BuildMon method demonstrates significant superiority over other methods in terms of pixel-level and instance-level accuracy. When it works without SWTA neck, BuildMon displays precision at 73.8 and recall at 37.2, coupled with an IoU at 33.21, a mIoU at 64.40, and a SCOT at 57.48 in the table, underscoring its efficacy in detecting structural changes. When embedded with the transformer-based module as the network neck, our method achieves a remarkable recall at 49.1 and pushes the IoU to an impressive 40.14, the mIoU to 67.90, the SCOT to 39.73, and the tracking score to 59.68. These metrics display BuildMon's

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON SPACENET 7 DATASET

Method	Paras(M)	Flops(G)	infer time(ms)	prec.	recall	IoU	mIoU	SCOT	Tracking Score	Change Score
U-net [22]	28.99	202.56	29.1	72.27	25.80	23.46	59.32	27.08	40.42	14.29
DeepLabv3 [23]	65.74	269.64	52.1	45.96	70.98	38.69	66.03	16.41	24.15	8.85
SegFormer [24]	3.72	6.36	26.1	67.72	22.98	20.71	57.83	17.87	30.04	8.52
MTL-U-Net [51]	9.43	374.88	83.7	72.49	29.57	26.59	60.94	30.83	44.04	17.78
TBFFNet [52]	26.02	1604.20	68.2	68.81	35.78	30.79	62.06	36.89	53.62	19.9
CFFM [13]	4.56	29.18	94.2	74.34	37.83	33.46	64.55	34.71	51.18	18.98
MRCFA [14]	5.13	16.91	83.4	74.08	36.11	32.06	63.81	33.91	49.60	18.61
BuildMon (w/o neck)	9.64	73.73	31.4	73.8	37.2	33.21	64.40	39.30	57.48	**21.40**
BuildMon	10.51	133.1	35.3	67.1	**49.1**	**40.14**	**67.90**	**39.73**	**59.68**	**21.34**

The bold values means the best performance.

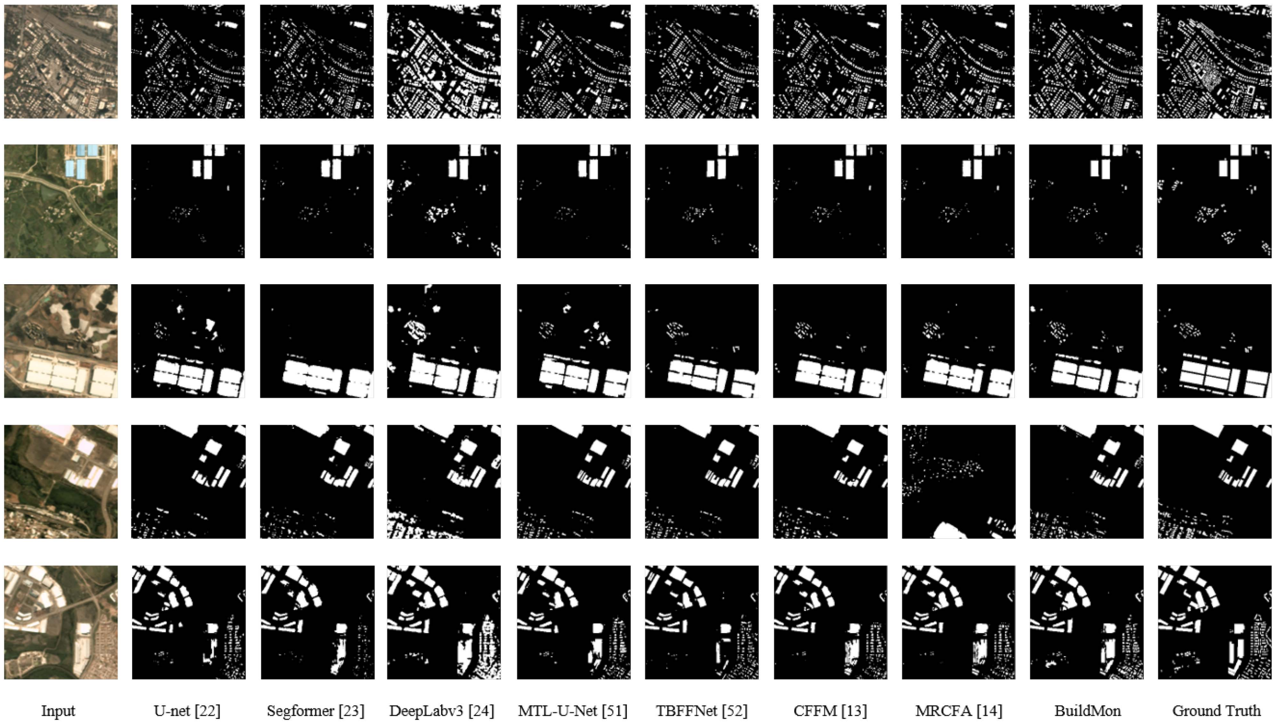


Fig. 3. Qualitative comparison of building extraction results. The first column is the input image, and the last column is the building mask annotation. The results suggest that BuildMon can identify more buildings, and the predicted building masks are more accurate.

superior performance in terms of segmentation and change monitoring.

Fig. 3 presents a qualitative evaluation of building extraction performance across various methods, highlighting the effectiveness of our proposed BuildMon approach. The building masks produced by BuildMon demonstrate superior completeness and smoother edges compared to other methods. Alternative methods exhibit deficiencies, with noticeable omissions of building pixels, instances being fused, and rougher edges. Notably, BuildMon's results show significant improvements in addressing these issues. Some other methods struggle to differentiate buildings from the background, resulting in fragmented segmentation results and numerous omissions. In contrast, BuildMon effectively distinguishes buildings, yielding more accurate segmentation. When handling the small buildings in the image, other methods produce results with distorted edges or fail to distinguish adjacent buildings accurately. On the contrary, BuildMon's outputs display smoother building contours and precise differentiation between adjacent buildings.

Given that traditional change detection is not aiming at predicting building masks, we have provided a supplemental comparison using the mIoU and F1-score metrics for changed regions, as Table II and Fig. 4. This serves as an auxiliary comparison, allowing us to present a comprehensive view of how our approach performs on the specific change detection task compared to other SOTA methods.

E. Ablation Study

1) Structure of ST Transformer: Architectural configuration of the Transformer plays a pivotal role in influencing model performance. In order to assess the impact of specific factors on our

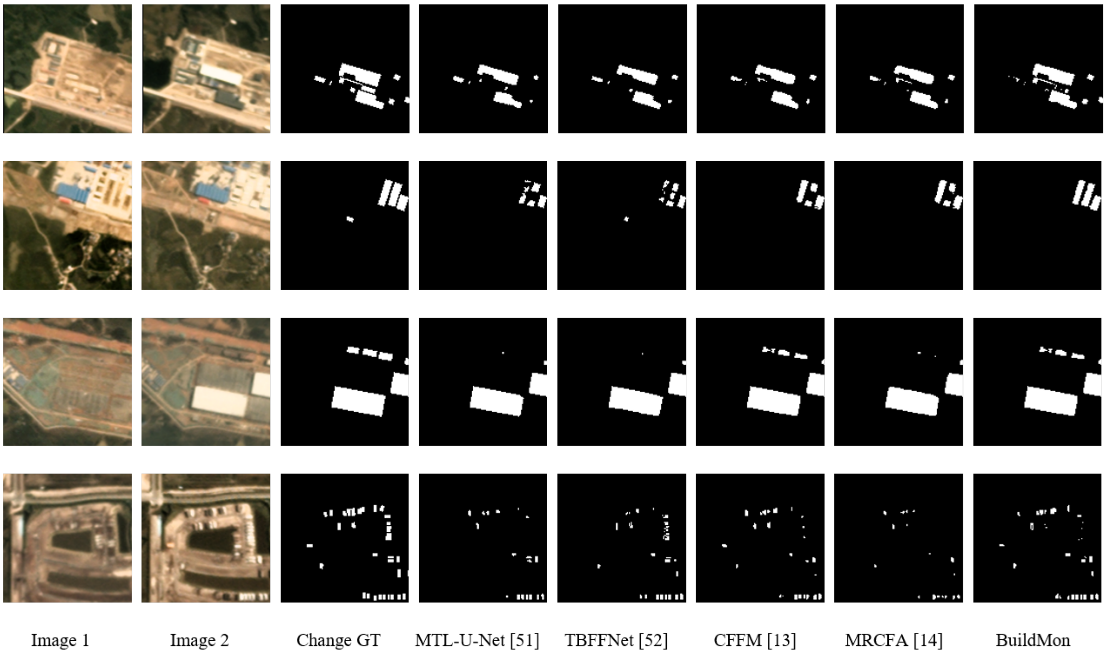


Fig. 4. Qualitative comparison of change detection results. BuildMon performs better in monitoring changed buildings. The first and second columns are the images of the two input temporal, and the third column is the ground truth annotation of the changed buildings. The experiment shows that the proposed method also has better performance on the change detection task.

TABLE II
CHANGE COMPARISON WITH SOTA METHODS ON THE SPACENET 7 DATASET

Method	Change mIoU	Change F1-score
MRCFA [14]	53.09	14.76
CFFM [13]	53.24	15.23
MTL-U-Net [51]	52.51	12.78
TBFFNet [52]	52.06	11.92
BuildMon	**57.52**	**28.13**

The bold values means the best performance.

TABLE III
ABLATION STUDY FOR HYPERPARAMETERS IN THE STRUCTURE OF TRANSFORMER NECK

#layers	#heads	prec.	recall	IoU	mIoU	SCOT	BIoU
1	1	**72.73**	43.05	37.07	66.41	**39.17**	36.06
1	2	70.33	43.92	37.06	66.38	37.64	36.09
2	1	71.02	**44.27**	**37.50**	**66.61**	37.71	**36.68**

The bold values means the best performance.

TABLE IV
ABLATION STUDY FOR DIFFERENT SPATIAL WINDOW SIZES IN THE STWA MODULE

Window Size	Prec.	Recall	IoU	mIoU	SCOT	BIoU
7	**72.73**	43.05	37.07	66.41	**39.17**	36.06
5	72.17	43.57	37.30	66.53	40.04	36.38
3	70.92	**44.18**	**37.40**	**66.56**	**40.07**	36.3
1	71.97	43.52	37.22	66.48	39.91	**36.41**

The bold values means the best performance.

model's performance, we undertake ablation experiments, focusing on the number of layers and heads within the Transformer. It is important to note that we initiate these experiments from the baseline model, incorporating solely the ST Transformer neck and utilizing the vanilla unweighted cross-entropy loss. For consistency, the spatial window size is set to 7, following the parameter setting in [81]. The results of these experiments are detailed in Table III.

Remarkably, the experimental results reveal that the variation in the number of layers and heads within the Transformer exerts minimal impact on the overall performance. This observation can be attributed to the relatively short length of each sequence, which is limited to 4. As a result, the temporal attention patterns within the TSRSI may not necessitate intricate coverage by a single Transformer head. Consequently, the experimental results indicate that a ST Transformer with one layer and one head is sufficient for extracting and fusing the temporal attention. In light of these insights, we opt for a simplified Transformer structure for subsequent experiments. This strategic simplification facilitates computational efficiency without compromising the model's ability to capture and leverage ST attention.

The size of the spatial window plays a crucial role in determining the field of spatial attention. To identify the optimal spatial window size for building monitoring in TSRSI, we conduct an ablation experiment comparing different settings while keeping the temporal length fixed at 4. The results are presented in Table IV, which reveal that the network achieves the best performance on both tasks when the window size is set to 3 × 3. This suggests that overly large or small spatial sizes have a detrimental impact on the performance of building extraction and change monitoring.

2) Comparison With 3-D Convolution Layer: Traditional 3-D convolution layer can also be adopted for feature extraction in TSRSI. To explore the difference between spatial-temporal window attention (STWA) and the 3-D convolution layer, a

TABLE V
COMPARISON THE STWA MODULE WITH 3-D CONVOLUTION LAYER WITHIN THE NECK OF THE NETWORK

Method	Presc.	Recall	IoU	mIoU	F1-Score
w/o neck	73.8	37.2	33.21	64.40	49.48
3D Conv	68.6	46.2	38.1	66.90	55.2
STWA	67.1	**49.1**	**40.14**	**67.90**	**57.2**

The bold values means the best performance.

TABLE VI
ABLATION STUDY FOR DIFFERENT BACKBONE IN THE STWA MODULE

BackBone	Paras(M)	Prec.	Recall	mIoU	SCOT
ResNet18 [82]	12.4	71.4	39.4	64.8	39.4
ResNeXt50 [83]	25.5	66.4	44.3	65.8	35.6
HRNet18 [79]	10.5	67.1	**49.1**	**67.9**	**39.7**

The bold values means the best performance.

TABLE VII
ABLATION FOR NORMALIZATION LAYER

IN Position		IoU	mIoU	SCOT	BIoU
Backbone	Neck				
		37.22	66.48	**39.91**	36.41
✓		38.53	67.17	39.14	37.26
	✓	38.59	67.14	39.18	36.77
✓	✓	**39.68**	**67.73**	38.84	**38.00**

The bold values means the best performance.

TABLE VIII
QUANTITATIVE COMPARISON OF DIFFERENT NORMALIZATION LAYER

Normalization Method	ED-RGB(↓)	SSIM(↑)	PSNR(↑)
BN	0.193	0.631	13.468
LN	0.213	0.609	14.153
IN	**0.155**	**0.669**	**14.342**

The bold values means the best performance.

comparative analysis is conducted with a focus on the neck module in our network architecture. Table V demonstrates the performance achieved using various implementations on a SpaceNet 7 dataset [7]. In the absence of an auxiliary network module (denoted as "w/o neck"), the results exhibit a mIoU of 64.40% and an F1-Score of 49.48%. The introduction of the 3-D convolution layer results in an enhancement, marked by an improved IoU of 38.1%, a higher mIoU of 66.90%, and F1-Score reaching 55.2%. Most notably, the STWA method manifests the most substantial performance enhancements, obtaining the IoU of 40.14%, the highest mIoU of 67.90%, and an F1-Score of 57.2%, which signifies the effectiveness of this approach.

3-D convolution only expands the dimension based on 2-D convolution and it still only extracts the features in a local receptive field, thus it cannot work well in modeling long-distance relationships in the TSRSI. Moreover, if the temporal receptive field is expanded to the whole input image group, 3-D convolution layers often struggle due to their fixed kernel size, thus the number of the inputs (or the channels) is unchangeable. At the same time, the 3-D convolution layer will bring a huge increase in the parameter number, which aggravates the computation burden. On the contrary, the ability of STWA to dynamically adapt to different data scales allows higher flexibility, which is lacking in conventional 3-D convolution methods. These advantages of STWA can be demonstrated with quantitative evidence in Table V.

3) Alternative Backbone: To investigate the impact of different backbones on the performance of building extraction and change detection, we selected three distinct feature extractors (i.e., ResNet [82], ResNeXt [83], and HRNet [79]) for comparison. The results, as depicted in Table VI, reveal that the HRNet [79] architecture demonstrates superior results across the board. Specifically, HRNet achieves a Recall of 49.1%, which is notably higher than the 39.4% and 41.4% attained by ResNet [82] and ResNeXt [83], respectively. Furthermore, HRNet outperforms the other architectures in terms of mIoU and SCOT, with scores of 67.9% and 39.73%, respectively. These metrics are critical for evaluating the accuracy of building extraction and change monitoring, which suggests that HRNet can retain more feature details and provide a more superior representation of the buildings for our tasks.

4) Instance Normalization Layer: The intricate appearance variations among temporal images in TSRSI establish an obstacle for modeling temporal attention. In response, we propose to replace the normalization layer in the Transformer with an IN layer. To evaluate the effectiveness of this strategy, we conduct an ablation study to determine the optimal position for replacing the IN layer within the network structure. Table VII presents the results of the ablation experiment, showing the performance of the network with the IN layer at different positions. The results suggest that when the IN layer is replaced in both the backbone network and the ST Transformer, the network achieves the best accuracy. The building IoU metric increases by 2.46 points and the BIoU metric is enhanced by 1.59 points.

To analyze the impact of the normalization methods on image appearance, we conducted an ablation experiment of normalization methods and selected Euclidean Distance in RGB Space (ED-RGB) [84], Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) as evaluation indexes. We apply these normalization methods to the whole set of TSRSI and analyze the similarity between the neighbor images. The metrics involved cover the similarity of color space and brightness. Experimental results reported in Table VIII show that IN distinctly outperforms the other methods, achieving the lowest ED-RGB score at 0.155, the highest SSIM and PSNR at 0.669 and 14.342. Fig. 5 presents a visual comparison of different normalization methods, showcasing the diverse effects on the image series. BN introduces information from other samples within the batch, which disrupts the normalization process of the current sample, resulting in insufficient appearance uniformity across different temporal images. LN shares the same mean and variance across all channels, leading to a shifted color tone. Considering the image appearances, IN is the most suitable normalization method for our specific task. It effectively eliminates image appearance and enhances the model's capacity to integrate temporal information.

TABLE IX
ABLATION FOR CHANGE-GUIDED LOSS

Change detection loss		prec.	recall	IoU	mIoU	SCOT			BIoU
Paired	Global					overall	tracking	change	
		70.40	47.62	39.68	67.73	38.84	**62.24**	19.45	38.00
✓		**70.71**	46.53	39.01	67.39	39.75	61.79	20.3	37.12
	✓	68.03	48.97	39.81	67.75	39.04	59.4	20.02	38.19
✓	✓	67.12	**49.97**	**40.14**	**67.90**	**39.73**	59.68	**21.34**	**38.48**
Weighted for building*		59.97	61.64	43.67	69.51	31.81	55.04	14.81	40.67

*This setting assigns a higher weight for building regions no matter if there is a change.
The bold values means the best performance.

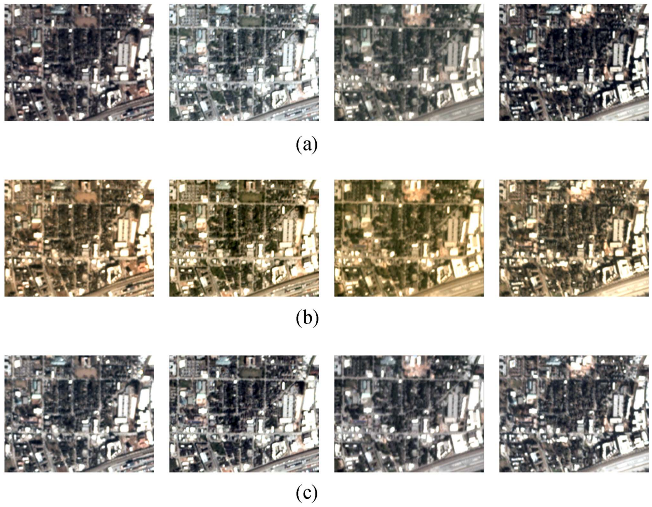


Fig. 5. Visual comparison of different normalization methods. (a) Displays the impact of the commonly used BN. There is an illumination variety among the images. (b) Shows a part of the image group processed by LN, where a shifted color tone exists. (c) The results of IN, keeping a consistency in appearance.

TABLE X
ABLATION STUDY FOR WEIGHT FACTOR λ IN LOSS FUNCTION

λ	IoU	mIoU	SCOT	Tracking Score	Change Score
0.5	38.56	67.07	37.4	57.68	19.36
0.7	38.1	66.92	38.61	59.09	20.57
1	40.14	67.90	39.73	59.68	21.34
1.5	38.97	67.34	39.54	60.36	20.29
3	38.17	66.87	38.53	58.98	19.95

5) Loss Functions: To effectively leverage valuable temporal change information in TSRSI, we design a change-guided loss to guide the ST Transformer to focus on change patterns. The results of ablation experiments for loss functions are detailed in Table IX, demonstrating that the proposed change-guided losses significantly enhance the model's performance. The *Paired* represents weighting the regions changed in two selected temporal images while *Global* corresponds to weighted the regions changed in the whole image series. For clarity, we record the scores of building tracking and change detection, as detailed in Table IX. The experimental results indicate a notably higher score for the tracking term compared to the change detection term in the original model. This emphasizes the effectiveness of the proposed change-guided losses in enhancing the model by directing more attention to the changed regions in TSRSI. Furthermore, the combination of the two losses achieves a substantial improvement, yielding a 67.90 mIoU and a 39.73 SCOT.

In addition, we experiment with weights exclusively applied to building regions to isolate the impact of the change-guided loss. The results are reported in the last row of Table IX. While assigning a higher weight for losses on building regions significantly enhances pixel-level accuracy, i.e., IoU and Boundary IoU, all three SCOT metrics experience a notable decline. This decline is attributed to the increase in loss weight for the building class, encouraging the model to prioritize building predictions. Consequently, there is a significant increase in the recall of buildings at the expense of more false positives, ultimately diminishing precision and SCOT.

The weight λ in (8) balances the attention on building extraction and change detection tasks, respectively. To explore how our model performance is affected by this variation, we have conducted an ablation study and its results are shown in Table X. It is observed that a value of 1 achieves the premier accuracy of these metrics, with the IoU peaking at 40.14, the mIoU at 67.90, the SCOT at 39.73, together with the highest tracking score of 59.68 and the change score of 21.34. This specific value of the weight factor seems to not only enhance model performance in terms of robustness and accuracy, but also optimize the tracking and detection of temporal changes of building instances.

V. DISCUSSION

Focusing on the buildings' properties in multitemporal RS image series, we simultaneously handle both tasks of building extraction and change monitoring, thus the network can predict the shapes, locations, tracked address IDs, and changes for each building instance in a pipeline. Regarding the definition of our task, change monitoring includes two subtasks, e.g., building instance tracking and construction status discovery, instead of solely predicting the change regions between two temporal-neighbor images like traditional change detection. Based on the experimental findings from the SpaceNet 7 Dataset [7], our proposed method, BuildMon, achieves state-of-the-art performance, effectively addressing both building extraction and change monitoring tasks simultaneously. Ablation studies reveal the effectiveness of all proposed designs and modifications in enhancing the network's performance.

Our BuildMon is not a simple combination of semantic segmentation, instance segmentation, and change detection. The windowed ST attention module aggregates temporal context to associate the building instance across different temporal images. Notably, our design differs from traditional 3D Transformers by employing small-scale windows on the spatial dimension, focusing solely on the local context for building extraction. This design fits the property of the TSRSI well since the semantic correlation in the whole temporal dimension is inherently stronger than that across the long-distance spatial scopes. Besides, in our framework, the change monitoring results are generated from the building extraction results and then the ground truth change patterns are employed as a building segmentation weight, merged in the supervised signals, resulting in a joint learning for both tasks.

BuildMon follows a sequential pipeline from segmentation to change detection, rather than adopting a multitask learning framework with two downstream branches. This approach allows the supervision information from the change monitoring task to guide the preparatory task of building extraction, facilitating feature representation learning to serve both tasks simultaneously. Despite the tasks' similarities, conflicts arise within them, as evidenced by ablation results for loss functions (see Table IX). Assigning weights solely to building regions results in a notable decrease in SCOT, indicating that the network tends to classify more regions as buildings, adversely affecting instance-level tracking and change monitoring. In addition, configurations without change-guided loss exhibit higher tracking scores but lower change scores. This discrepancy stems from two factors: First, in the tracking task, buildings missed in earlier detections may be identified in subsequent temporal instances, allowing for recovery. Conversely, in change detection, missed changes at specific time points are challenging to rectify. Second, models without change-guided loss have limited capacities to represent change patterns, leading to lower change detection scores.

Our method provides a new perspective for building extraction and change patterns analysis in multitemporal image series, while it is not without limitations at the same time. In our approach, the targets do not undergo displacement and do not experience drastic "appear-disappear" changes. Therefore, our method struggles to extend to rapidly moving targets such as ships or vehicles, or vegetation and forests with frequent periodic changes. However, the concept we proposed for temporal-spatial context-aware modeling in TSRSI can be applied to other long-sequence target analysis problems. Note that it is necessary to construct specific pattern extraction networks based on the characteristics of the target changes. The two-step strategy of our method for building extraction and change monitoring may limit the simplicity and flexibility of the network. Moreover, considering computational complexity, we opt to first extract buildings through semantic segmentation, followed by acquiring building instances via postprocessing methods. However, this redundant pipeline may hinder end-to-end optimization and exacerbate training difficulty. Therefore, our future work will explore novel building delineation approaches and efficient instance segmentation networks with simpler structures to handle both tasks concurrently.

VI. Conclusion

In this article, we have presented BuildMon, a novel framework aimed at addressing the challenges of building extraction and change monitoring within TSRSI. BuildMon incorporates a ST Transformer, enabling the capture of ST attention across sequential image data, thereby facilitating the modeling of relationships among neighboring regions and cross-image patches. To mitigate appearance discrepancies within time series images, we introduce the utilization of IN within the Transformer architecture. Furthermore, we present two customized loss functions tailored for change detection, effectively addressing class imbalance and emphasizing regions undergoing significant changes. Extensive experiments conducted on the SpaceNet 7 dataset demonstrate the effectiveness of the BuildMon framework.

References

[1] X. Huang, D. Zhu, F. Zhang, T. Liu, X. Li, and L. Zou, "Sensing population distribution from satellite imagery via deep learning: Model selection, neighboring effects, and systematic biases," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 5137–5151, 2021.

[2] G. Schrotter and C. Hürzeler, "The digital twin of the city of Zurich for urban planning," *J. Photogrammetry, Remote Sens. Geoinf. Sci.*, vol. 88, no. 1, pp. 99–112, 2020.

[3] M.-D. Yang, K.-S. Huang, J. Wan, H. P. Tsai, and L.-M. Lin, "Timely and quantitative damage assessment of oyster racks using UAV images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2862–2868, Aug. 2018.

[4] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 115, pp. 22–36, May 2016.

[5] M. Decuyper et al., "Continuous monitoring of forest change dynamics with satellite time series," *Remote Sens. Environ.*, vol. 269, 2022, Art. no. 112829.

[6] M. Zhao, S. Li, S. Xuan, L. Kou, S. Gong, and Z. Zhou, "SatSOT: A benchmark dataset for satellite video single object tracking," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[7] A. V. Etten, D. Hogan, J. M. Manso, J. Shermeyer, N. Weir, and R. Lewis, "The multi-temporal urban development Spacenet dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6394–6403.

[8] Z. Lv, J. Liu, W. Sun, T. Lei, J. A. Benediktsson, and X. Jia, "Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[9] Z. Lv, H. Huang, W. Sun, T. Lei, J. A. Benediktsson, and J. Li, "Novel enhanced UNet for change detection using multimodal remote sensing image," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[10] Z. Lv, M. Zhang, W. Sun, J. A. Benediktsson, T. Lei, and N. Falco, "Spatial-contextual information utilization framework for land cover change detection with hyperspectral remote sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.

[11] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, pp. 1–47, 2020.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Mask2former for video instance segmentation," 2021, *arXiv:2112.10764*.

[13] G. Sun, Y. Liu, H. Ding, T. Probst, and L. V. Gool, "Coarse-to-fine feature mining for video semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3116–3127.

[14] G. Sun, Y. Liu, H. Tang, A. Chhatkuli, L. Zhang, and L. V. Gool, "Mining relations among cross-frame affinities for video semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 522–539.

[15] T. Zhou, F. Porikli, D. J. Crandall, L. V. Gool, and W. Wang, "A survey on deep learning technique for video segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7099–7122, Jun. 2023.

[16] Y. Liu, C. Shen, C. Yu, and J. Wang, "Efficient semantic video segmentation with per-frame inference," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 352–368.

[17] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Time-space tradeoff in deep learning models for crop classification on satellite multi-spectral image time series," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 6247–6250.

[18] V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata, "Satellite image time series classification with pixel-set encoders and temporal self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12322–12331.

[19] V. S. F. Garnot and L. Landrieu, "Lightweight temporal self-attention for classifying satellite images time series," in *Proc. Adv. Analytics Learn. Temporal Data*, 2020, pp. 171–181.

[20] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4852–4861.

[21] M. Tarasiou, E. Chavez, and S. Zafeiriou, "ViTs for SITS: Vision transformers for satellite image time series," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervention*, 2015, pp. 234–241.

[23] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[24] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[25] Q. Zhu, C. Liao, H. Hu, X. Mei, and H. Li, "MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6169–6181, Jul. 2021.

[26] M. Luo, S. Ji, and S. Wei, "A diverse large-scale building dataset and a novel plug-and-play domain generalization method for building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4122–4138, 2023.

[27] J. Liu, H. Huang, H. Sun, Z. Wu, and R. Luo, "LRAD-Net: An improved lightweight network for building extraction from remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 675–687, 2023.

[28] X. Liu et al., "Building instance extraction method based on improved hybrid task cascade," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] Q. Li et al., "Instance segmentation of buildings using keypoints," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1452–1455.

[30] J. Wang, L. Meng, W. Li, W. Yang, L. Yu, and G.-S. Xia, "Learning to extract building footprints from off-nadir aerial images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1294–1301, Jan. 2023.

[31] Y. Liu, D. Chen, A. Ma, Y. Zhong, F. Fang, and K. Xu, "Multiscale U-shaped CNN building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6106–6120, Jul. 2021.

[32] S. Zhang, Y. Cao, and B. Sui, "DF-Mask R-CNN: Direction field-based optimized instance segmentation network for building instance extraction," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[33] J. Yin, F. Wu, and Y. Qi, "Vector mapping method for buildings in remote sensing images based on joint semantic-geometric learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9067–9076, 2023.

[34] S. Zorzi, S. Bazrafkan, S. Habenschuss, and F. Fraundorfer, "PolyWorld: Polygonal building extraction with graph neural networks in satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1938–1947.

[35] B. Xu, J. Xu, N. Xue, and G.-S. Xia, "HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 284–296, 2023.

[36] Y. Hu, Z. Wang, Z. Huang, and Y. Liu, "Polybuilding: Polygon transformer for building extraction," *ISPRS J. Photogrammetry Remote Sens.*, vol. 199, pp. 15–27, 2023.

[37] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.

[38] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1280–1289.

[39] X. Huang, L. Zhang, and T. Zhu, "Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 1, pp. 105–115, Jan. 2014.

[40] D. Wen, X. Huang, L. Zhang, and J. A. Benediktsson, "A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 609–625, Jan. 2016.

[41] A. Lefebvre and T. Corpetti, "Monitoring the morphological transformation of Beijing old city using remote sensing texture analysis," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 539–548, Feb. 2017.

[42] X. Huang, Y. Cao, and J. Li, "An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images," *Remote Sens. Environ.*, vol. 244, 2020, Art. no. 111802.

[43] 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.

[44] B. Bai, W. Fu, T. Lu, and S. Li, "Edge-guided recurrent convolutional neural network for multitemporal remote sensing image building change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022.

[45] Y. Sun, X. Zhang, J. Huang, H. Wang, and Q. Xin, "Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[46] M. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Edge neighborhood contrastive learning for building change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[47] M. Li, X. Liu, X. Wang, and P. Xiao, "Detecting building changes using multimodal Siamese multitask networks from very-high-resolution satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–22, 2023.

[48] R. Hang, P. Yang, F. Zhou, and Q. Liu, "Multiscale progressive segmentation network for high-resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[49] F. Zhou, C. Xu, R. Hang, R. Zhang, and Q. Liu, "Mining joint intra-and inter-image context for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.

[50] R. Hang, S. Xu, P. Yuan, and Q. Liu, "AANet: An ambiguity-aware network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[51] 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.

[52] Z. Li, X. Wang, S. Fang, J. Zhao, S. Yang, and W. Li, "A decoder-focused multitask network for semantic change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024.

[53] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1997, pp. 731–737.

[54] J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor, "Detecting trend and seasonal changes in satellite image time series," *Remote Sens. Environ.*, vol. 114, no. 1, pp. 106–115, 2010.

[55] X. Su, C.-A. Deledalle, F. Tupin, and H. Sun, "NORCAMA: Change analysis in SAR time series by likelihood ratio change matrix clustering," *ISPRS J. Photogrammetry Remote Sens.*, vol. 101, pp. 247–261, 2015.

[56] C. Robinson, A. Ortiz, J. M. L. Ferres, B. Anderson, and D. E. Ho, "Temporal cluster matching for change detection of structures from satellite imagery," in *Proc. ACM SIGCAS Conf. Comput. Sustain. Societies*, 2021, pp. 138–146.

[57] J. Ruiz-Ramos, A. Marino, C. Boardman, and J. Suarez, "Continuous forest monitoring using cumulative sums of sentinel-1 timeseries," *Remote Sens.*, vol. 12, no. 18, pp. 3061–3084, 2020.

[58] O. Csillik, M. Belgiu, G. P. Asner, and M. Kelly, "Object-based time-constrained dynamic time warping classification of crops using Sentinel-2," *Remote Sens.*, vol. 11, no. 10, pp. 1257–1283, 2019.

[59] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.

[60] N. Pettorelli, *The Normalized Difference Vegetation Index*. New York, NY, USA: Oxford Univ. Press, 2013.

[61] Q. Yuan et al., "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, pp. 111716–111739, 2020.

[62] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 421–435, 2020.

[63] A. V. Etten and D. Hogan, "The SpaceNet multi-temporal urban development challenge," in *Proc. NeurIPS Competition Demonstration Track*, (Series Proceedings of Machine Learning Research), 2021, pp. 216–232.

[64] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[67] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1510–1519.

[68] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.

[69] Y. Li, N. Wang, J. Liu, and X. Hou, "Demystifying neural style transfer," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2230–2236.

[70] F. Meyer and S. Beucher, "Morphological segmentation," *J. Vis. Commun. Image Representation*, vol. 1, no. 1, pp. 21–46, 1990.

[71] W. Bao, D. Shen, P. Ni, J. Zhou, and Y. Sun, "Proposition and certification of moving mean difference method for detecting abrupt change points," *Acta Geographica Sinica*, vol. 73, no. 11, pp. 2075–2085, 2018.

[72] 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.

[73] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.

[74] A. V. Etten, D. Lindenbaum, and T. M. Bacastow, "SpaceNet: A remote sensing dataset and challenge series," 2018, *arXiv:1807.01232*.

[75] N. Weir et al., "SpaceNet MVOI: A multi-view overhead imagery dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 992–1001.

[76] 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, pp. 1662–1695, May 2020.

[77] C. Pang, J. Wu, J. Ding, C. Song, and G.-S. Xia, "Detecting building changes with off-nadir aerial images," *Sci. China Inf. Sci.*, vol. 66, no. 4, Mar. 2023, Art. no. 140306.

[78] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15329–15337.

[79] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[80] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019.

[81] Z. Liu et al., "Video Swin Transformer," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2022, pp. 3202–3211.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[83] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5987–5995.

[84] R. Fisher, "Change detection in color images," in *Proc. IEEE 7th Conf. Comput. Vis. Pattern*, 1999.

Yuxuan Wang (Graduate Student Member, IEEE) received the B.S. degree in measurement and control technology and instrumentation, in 2021, from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D. degree in communication and information system.

His research interests include computer vision and image processing, especially in cross-modality learning, semantic segmentation, and instance segmentation.

Shuailin Chen received the B.S. degree in electronic information engineering and the M.S. degree in electronic and communication engineering from Wuhan University, Wuhan, China, in 2020 and 2023, respectively.

His research interests include remote sensing image processing and SAR semantic segmentation.

Ruixiang Zhang (Student Member, IEEE) received the B.S. degree in electronic engineering, in 2019, from Wuhan University, Wuhan, China, where he is currently working toward the Ph.D degree in communication and information system.

His research interests include remote sensing image processing, label-efficient object detection, and cross-modal object detection.

Fang Xu received the B.S. degree in electronic and information engineering and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 2018 and 2023, respectively.

She is a Postdoctoral Researcher with the School of Computer Science, Wuhan University. Her research interests include remote sensing image processing, multimodal data matching, and fusion.

Shuo Liang received the B.S. degree in computer science and technology in 2014 from Northwestern Polytechnical University, and the M.S. degree in signal and information processing in 2017 from the Institute of Communication Measurement and Control Technology.

His research interests include aerospace information applications and remote sensing image processing.

Yujing Wang received the Ph.D degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2020.

Her research interests include remote sensing intelligent interpretation and spatial data mining.

Wen Yang (Senior Member, IEEE) received the B.S. degree in electronic apparatus and surveying technology, the M.S. degree in computer application technology, and the Ph.D. degree in communication and information system from Wuhan University, Wuhan, China, in 1998, 2001, and 2004, respectively.

From 2008 to 2009, he was a Visiting Scholar with the Apprentissage et Interfaces (AI) Team, Laboratoire Jean Kuntzmann, Grenoble, France. From 2010 to 2013, he was a Postdoctoral Researcher with the State Key Laboratory of Information Engineering, Surveying, Mapping and Remote Sensing, Wuhan University. Since then, he has been a Full Professor with the School of Electronic Information, Wuhan University. His research interests include object detection and recognition, multisensor information fusion, and remote sensing image processing.