Remote Sensing Applications: Society and Environment 36 (2024) 101336

Contents lists available at ScienceDirect

Remote Sensing Applications: Society and Environment

journal homepage: www.elsevier.com/locate/rsase

Building detection in VHR remote sensing images using a novel dual attention residual-based U-Net (DAttResU-Net): An application to generating building change maps

Ehsan Khankeshizadeh [a], Ali Mohammadzadeh [a,*], Amin Mohsenifar [a], Armin Moghimi [b], Saied Pirasteh [c], Sheng Feng [c], Keli Hu [c], Jonathan Li [d]

[a] Department of Photogrammetry and Remote Sensing, Geomatics Engineering Faculty, K. N. Toosi University of Technology, Tehran, 15433-19967, Iran
[b] Ludwig-Franzius-Institute for Hydraulic, Estuarine and Coastal Engineering, Leibniz University Hannover, Nienburger Str. 4, Hanover, 30167, Germany
[c] Institute of Artificial Intelligence, Shaoxing University, 508 West Huancheng Road, Yuecheng District, Shaoxing, Zhejiang Province, 312000, China
[d] Department of Geography and Environmental Management, University of Waterloo, Waterloo, ON, N2L 3G1, Canada

ARTICLE INFO

Keywords:
VHR-RSIs
Building change map
DAttResU-Net
Building detection

ABSTRACT

In today's era, increasing access to very high-resolution remote sensing images (VHR-RSIs) has enhanced building detection and change assessment capabilities. These applications provide accurate urban mapping, facilitate effective land management, and support disaster assessment by delivering detailed insights into building structures and their temporal changes. This study uses a two-stage process to present a pioneering approach for generating precise building maps (BMs) and subsequent building change maps (BCMs) from VHR-RSIs. The primary question addressed by the research is how to enhance the U-Net architecture to improve its sensitivity to both high-level semantic features (HLSF) and low-level spatial features (LLSF) in the building detection task. For this purpose, in the initial stage of the method, a novel deep learning model called dual attention residual-based U-Net (DAttResU-Net) is introduced. This model incorporates two significant modifications to the conventional U-Net, enhancing its capacity to yield bi-temporal BMs. Firstly, each standard convolutional block (CB) is replaced with an optimized CB incorporating a channel-spatial attention module attuned to the building objects' crucial HLSF. Secondly, an additional attention module is integrated into the encoder-decoder path of the model, heightening the sensitivity of U-Net to vital LLSF of buildings while disregarding extraneous background spatial information during the fusion of HLSF and LLSF. In the subsequent stage, the bi-temporal BMs generated by the DAttResU-Net are subjected to a box-based class-object change detection methodology to produce accurate BCMs. The effectiveness of the proposed architecture is rigorously evaluated against state-of-the-art models in both BM and BCM generation contexts, utilizing the well-established WHU dataset for experimentation. The experimental results indicated that the DAttResU-Net model, boasting an average of P_{FN}/ P_{FP} value of 2.33/1.34 (%) surpasses the performance of the state-of-the-art models in generating bi-temporal BMs. Furthermore, the building change detection outcomes demonstrated the proficient role of the bi-temporal BMs predicted by the proposed model in leading to the most optimal BCMs, exhibiting average P_{FN}/ P_{FP} value of 2.63/8.93 (%), outperforming comparative networks. Finally, we concluded that the

* Corresponding author.
E-mail addresses: Seyedehsan.Khankeshizadeh@email.kntu.ac.ir (E. Khankeshizadeh), a_mohammadzadeh@kntu.ac.ir (A. Mohammadzadeh), a.mohsenifar@email.kntu.ac.ir (A. Mohsenifar), moghimi@lufi.uni-hannover.de (A. Moghimi), sapirasteh1@usx.edu.cn (S. Pirasteh), fengsheng_13@usx.edu.cn (S. Feng), hukeli@usx.edu.cn (K. Hu), junli@uwaterloo.ca (J. Li).

https://doi.org/10.1016/j.rsase.2024.101336
Received 17 July 2024; Received in revised form 12 August 2024; Accepted 27 August 2024
Available online 2 September 2024

proposed DAttResU-Net architecture is a highly promising and applicable model for producing reliable BMs and BCMs.

1. Introduction

The rapid pace of urbanization brings many challenges globally, including accelerating the construction of buildings in regions where populations move rapidly (due to nomadic displacement or migration) (The World Bank, 2023). Hence, automatically creating building maps (BMs) and building change maps (BCMs) is necessary to demonstrate local environmental improvement, land management, disaster assessment, and other geospatial applications (Manno-Kovacs and Sziranyi, 2015; Feng et al., 2020). In this regard, remote sensing images (RSIs) have been broadly used to reduce the necessity for costly field survey operations and speed up the generation of BMs and BCMs. Particularly, very high-resolution RSIs (VHR-RSIs) contain detailed information about terrestrial phenomena and have been widely used in BM generation studies (Farnood Ahmadi et al., 2024), coupled with either traditional or modern image processing methods.

The traditional methods usually employ low-level building features like color, geometrical, and contextual features. In this context, many algorithms have been previously developed based on pixel (Sirmaçek and Ünsalan, 2008; Fuentes Reyes et al., 2023), spectrum (Zhang 1999; Sheng-Hua et al., 2008), length, edge (Yong and Huayi, 2007; Ferraioli, 2010; Pirasteh et al., 2019), texture (Awrangjeb et al., 2013; Li et al., 2021a,b), and shadow (Zhu et al., 2021) information. For example, Farhadi et al. (2023) developed an unsupervised method for extracting buildings using Sentinel-1&2 satellite imagery and SRTM DEM data. They introduced a novel radar index (NRI) for primary building extraction and combined it with spectral indices to enhance accuracy. However, these low-level features are highly affected by seasonal illumination and radiometric changes, reducing their capability in building detection (Ji and Wei, 2019; Chen et al., 2023). Therefore, traditional methods require a high level of feature engineering where the appropriate features should be manually designed based on the imaging and atmospheric conditions of the target area. Moreover, these algorithms are less generalizable, and their accuracy also depends more on the expert's expertise. In addition, the complexity of building objects and their similarity to the non-building classes may lead to many challenges in traditional methods.

Recent advances in computer hardware, algorithmic developments, computational capabilities, and access to big data have enabled deep learning (DL) approaches, particularly convolutional neural networks (CNNs), to address the challenges related to traditional approaches (Uzar et al., 2021; Kaya et al., 2023; Moghimi et al., 2024). Indeed, these techniques have great potential to automatically extract complicated features from spectral-spatial-temporal data (Khankeshizadeh et al., 2022, 2024) and have recently been proven efficient in building detection from VHR-RSIs. Notably, fully CNNs (FCNNs) have recently optimized CNNs by replacing fully connected layers with CLs layers throughout the entire network. Especially, newly developed Encoder-Decoder FCNN-based networks like SegNet (Badrinarayanan et al., 2017), DeconvNet (Noh et al., 2015), and U-Net (Ronneberger et al., 2015) models have attained exceptional outcomes in the BM generation field (Chen and Lu, 2019; Ivanovsky et al., 2019; Naanjam and Farnood Ahmadi, 2024). Of these, U-Net-based architectures have been broadly used to detect buildings due to their fast, precise response, rapid training phase, and inexpensive hardware requirements (Ronneberger et al., 2015). Hence, several studies have employed U-Net variants for building detection, leveraging its ability to handle complex and high-resolution satellite imagery. For instance, Alsabhan et al. (2022) used the U-Net model for automatic building detection from VHR satellite images. They found that U-Net, especially with a VGGNet backbone, achieved high accuracy (84.9%) even with limited training data, outperforming existing models for similar tasks. Xu et al. (2018) and Wang and Miao (2022) combined ResNet residual block with U-Net to advance the model in building extraction by avoiding the gradient vanishing issue in the learning stage. Kang et al. (2019) also proposed the EU-Net model, which can use a dense spatial pyramid integration to achieve better detailed building information at different scales.

In previous studies, various attention modules inspired by the human visual system, due to carefully subjecting specific details under scrutiny, have extensively been added to the typical U-Net model at two different structural levels of the model. At the first level, spatial, channel, or channel-spatial attention modules replaced the simple CLs of U-Net to intensify their attention and sensitivity to building features (Pan et al., 2019; Li et al., 2021a; Xu et al., 2021). At the second level, another attention module was added to the model's encoder-decoder (skip connection) path to emphasize critical spatial information while minimizing irrelevant background objects (Guo et al., 2020; Li et al., 2021b; Yu et al., 2022). Lei et al. (2024) introduced the double hybrid attention U-Net (DHAU-Net) to improve building detection in VHR remote sensing images. Using dual-parallel hybrid attention modules in their proposed DHAU-Net model enhances feature extraction while reducing interference.

Although the recently presented attention-coupled U-Net models provided satisfactory performance in the building detection domain, some substantial limitations still remain, which need to be well tackled. For instance, the existing BM generation studies solely use a single attention module at either the first or second level of the model. In detail, on the one hand, the use of an attention mechanism only sensitive to building high-level semantic features (HLSF) allows the irrelevant low-level spatial features (LLSF) to flow with no restrictions in the encoder-decoder path of the model and be aggregated with HLSF, which can probably degrade sensitivity to the building features. Furthermore, the existence of many CLs after adding the attention mechanisms to the U-Net model itself disrupts the gradient descent process, resulting in the gradient vanishing problem. On the other hand, solely relying on an LLSF-sensitive attention mechanism ignores the critical attention to the building HLSF when applying the CLs. Hence, employing attention modules at either the first or second levels neglects the potential benefits offered by both LLSF and HLSF. The risk of generating subtle differences between buildings and non-buildings arises by amalgamating these two sources of building information. In some cases, this approach could lead to the partial or complete omission of buildings within the BMs. This omission underscores the significance of a comprehensive approach encompassing the integration of LLSF and HLSF to ensure accurate and robust building detection and representa-

tion within the models. This strategy prevents the potential loss of vital building data and contributes to more precise and reliable building representations in the BMs.

To cope with the limitations of the attention-based U-Net models and to intensify their sensitivity to building LLSF and HLSF, a novel DL architecture, named dual attention residual-based U-Net (DAttResU-Net), was proposed in this study. The proposed DAttResU-Net first leveraged the potential of the synergic addition of residual blocks and attention mechanisms to the U-Net network to produce bi-temporal BMs. These bi-temporal maps were then compared through a box-based approach to reach a BCM. In summary, the main contributions of this paper are as follows.

1) To amplify the U-Net's concentration on satisfactory building HLSF, a new proposed channel-spatial attention residual convolutional block (csAttResConvB) replaced simple CLs in U-Net. Moreover, an attention module named attention gate (AttG) was incorporated into the encoder-decoder path of U-Net in order to increase the sensitivity to significant LLSF.
2) The proposed DAttResU-Net, due to exploiting the two powerful csAttResConvB and AttG modules, is the first U-Net-based model integrating beneficial both LLSF and HLSF while minimizing the negative impact of non-building information.
3) The current study is the first attempt to explore if the dependable bi-temporal BMs predicted by DAttResU-Net are applicable for detecting building changes.

The rest of the paper is structured as follows: Section 2 details the proposed DL-based DAttResU-Net used to generate BMs and the straightforward box-based comparison manner employed to reach the BCM. Section 3 describes the specifications of the datasets used to conduct the experiments. The numerical accuracy metrics for evaluating the results attained in this research are also mentioned in Section 4. Then, the analysis and discussion of the quantitative and qualitative outcomes for both building identification and building change detection are presented in Section 5. Finally, we conclude the key findings derived from this research in Section 6.

2. Methodology

The framework of the proposed method for BM and BCM generation is composed of two main steps, as illustrated in Fig. 1 (a). The input bi-temporal VHR images are segmented into binary BMs using a proposed attention-based U-Net building extraction network in the first step. In the second step, any polygon in each BM is bound with a bounding box, and the small noisy non-building boxes caused by imperfect building extraction are filtered. Finally, corresponding building boxes in the bi-temporal BMs are compared to attain a BCM.

Mathematically, we considered bi-temporal geometrically corrected VHR-RSIs, $\boldsymbol{X_{t1}} = \{x_{t1}(i,j,k) | 1 \le i \le r, 1 \le j \le c, 1 \le k \le b\}$ and $\boldsymbol{X_{t2}} = \{x_{t2}(i,j,k) | 1 \le i \le r, 1 \le j \le c, 1 \le k \le b\}$, with dimensions of $r \times c \times b$. They were collected over identical geographical regions at two different times, t_1 and t_2, respectively.

In the proposed method, two primary goals are pursued: (1) using the DAttResU-Net model to generate optimal bi-temporal BMs, $\boldsymbol{BM_{t1}} = \{bm_{t1}(i,j) | 1 \le i \le r, 1 \le j \le c\}$ and $\boldsymbol{BM_{t2}} = \{bm_{t2}(i,j,k) | 1 \le i \le r, 1 \le j \le c\}$; $bm_{t1/2}(i,j) \in \{0,1\}$ (i.e., 'no-building = 0' and 'building = 1'), (2) creating a reliable binary BCM $\boldsymbol{C_M} = \{c_m(i,j) | 1 \le i \le r, 1 \le j \le c\}$; $c_m(i,j) \in \{0,1,2\}$ (i.e., 'negative change = 0′, 'no change = 1′ and 'positive change = 2′) by correspondingly comparing the building boxes previously extracted in the $\boldsymbol{BM_{t1}}$ and $\boldsymbol{BM_{t2}}$.

2.1. BM generation using the proposed DAttResU-Net (step 1)

2.1.1. DAttResU-Net architecture

A general overview of the proposed DAttResU-Net architecture is shown in Fig. 1 (b). In this study, DAttResU-Net was designed after making two major improvements to the structure of the ordinary U-Net model. As the first improvement, a channel-spatial attention module alongside a residual unit replaces the simple CLs in the original U-Net structure's encoding and decoding parts. In this way, the network's attention and sensitivity to sparse buildings are intensified. As the second improvement, AttG is added to the skip-connection path of the original U-Net structure to balance the aggregation of the same-scale LLSF and HLSF maps. Consequently, this module suppresses irrelevant and unreliable non-building information. Each improvement made to the typical U-Net model is detailed in the following.

2.1.1.1. Channel-spatial attention residual convolutional block (csAttResConvB) architecture. In the original U-Net architecture, CLs responsible for multi-level feature extraction have a limited receptive field and cannot model complex correlations between buildings and non-building objects such as soils and roads. Thus, in order to effectively highlight the target building features, inspired by (Woo et al., 2018), a csAttResConvB including the channel attention module (cAM), spatial attention module (sAM), and the residual unit was designed according to Fig. 2 (a) and replaced simple CLs in U-Net.

As shown in Fig. 2 (a), suppose $x \in \mathbb{R}^{H \times W \times C}$ is an arbitrary input to the convolutional block (CB), $F^N \in \mathbb{R}^{H \times W \times N}$ are the feature maps extracted by CB, $M_c \in \mathbb{R}^{1 \times 1 \times N}$ is a 1D channel attention map, and $M_s \in \mathbb{R}^{1 \times H \times W}$ is a 2D spatial attention map. The values H,W, and N also respectively denote the number of rows, columns, and the number of channels or CB filters in a layer of interest. The overall process of csAttResConvB can be summarized as follows:

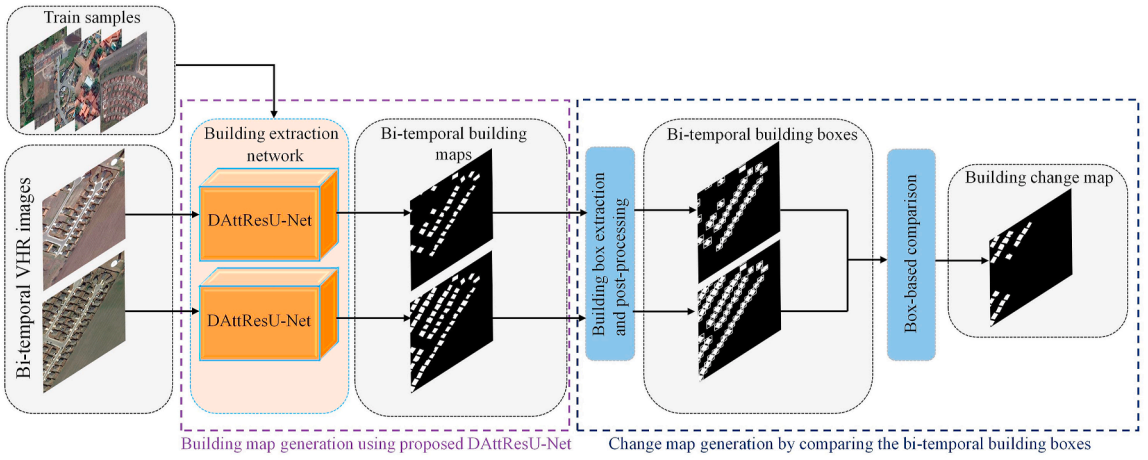

Fig. 1. (a) The workflow of the DAttResU-Net-based building detection and the box-based comparison change detection method; (b) The architecture of dual attention residual-based U-Net (DAttResU-Net).

$$\begin{aligned} F^{N} &= \left[x \times g1 + d + b + r\right]^{2}, \\ F^{N\prime} &= M_{c}\left(F^{N}\right) \otimes F^{N}, \\ F^{N\prime\prime} &= M_{s}\left(F^{N\prime}\right) \otimes F^{N\prime}, \\ x' &= \left[x \times g2 + b + r\right] + F^{N\prime\prime}. \end{aligned} \tag{1}$$

where, $g1$ and $g2$ are the convolutional kernels with different dimensions. Moreover, d, b, and r respectively signify the dropout (Drop), batch normalization (BN), and the rectified linear unit (ReLU) activation function layers. It is worth noting that, drop layer as a regularization technique is used to prevent overfitting by randomly deactivating neurons during training phase, promoting model generalization. In this study, a drop layer with a rate of 0.5 was applied. Also, $\times$ and $\otimes$ represent the convolution and element-wise multiplication operators, respectively. Furthermore, $F^{N\prime}$ and $F^{N\prime\prime}$ are the channel-enhanced and channel- and spatial-enhanced feature maps. Lastly, x' as the final output of csAttResConB is achieved after the features extracted using channel and spatial attention mechanisms are refined and combined with the input by a residual block. The proposed csAttResConvB structure is demonstrated in detail in the following steps.

(a) Channel attention module (cAM)

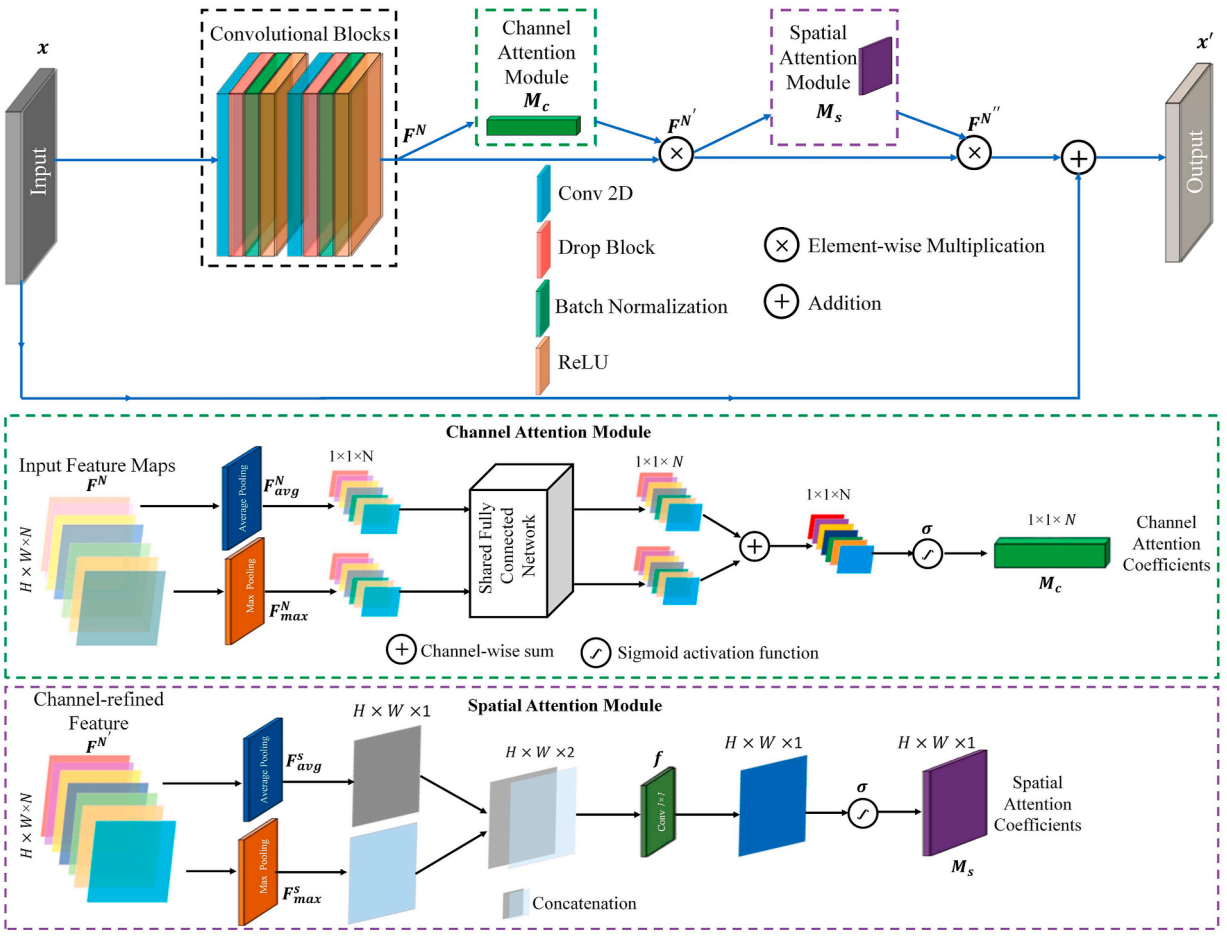

Fig. 2. (a) Diagram of the channel-spatial attention residual convolutional block (csAttResConvB); (b) The architecture of the attention gate (AttG) module.

After the initial feature maps (F^N) were extracted by the CB, the cAM process was here conducted to estimate some coefficients, called channel attention coefficient ($M_c\left(F^N\right)$), which determine how important initial maps are individually in terms of comprising beneficial building information (e.g., edge, texture, geometric, and shape). As seen in Fig. 2 (a), to reach the $M_c(F^N)$ in the cAM procedure, the pooling operations, including average-pooling and max-pooling were first used as information aggregators to specify the key information in each channel of the F^N and respectively generate $F^N_{avg} \in \mathbb{R}^{1\times1\times N}$ and $F^N_{max} \in \mathbb{R}^{1\times1\times N}$ according to Eq. (2):

$$\begin{aligned} F^N_{max} &= Max\ (\ F^n(i,j)) \\ F^N_{avg} &= \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} F^n(i,j) \\ 1<n<N, 1<i<H, 1<j<W \end{aligned} \tag{2}$$

where, *Max*(.) is the *'Maximum'* operator and $F^n(i,j)$ also indicates the pixel value at a position of interest (i,j) in the n th feature map. Secondly, the previously produced F^N_{avg} and F^N_{avg} were forwarded to a share-weighted multi-layer perceptron (MLP) network with two fully connected layers to refine the descriptors. A reduction ratio of 16 was selected for the shared MLP. This choice ensured that the CBAM implementation could efficiently enhance features at the channel level while keeping the number of parameters and computational demands within a manageable range.

The descriptors were then combined with the channel-wise summation operator to produce a new single descriptor (i.e., coefficient map), and the sigmoid activation function was used to rescale the values of the output coefficient map between 0 and 1 and reach the final coefficient map ($M_c(F^N)$). Briefly, the cAM is computed as follows:

$$M_c\left(F^N\right) = \sigma\left(W1\left(W0\left(F^N_{max}\right)\right) + W1\left(W0\left(F^N_{avg}\right)\right)\right) \tag{3}$$

where, σ denotes the sigmoid function and $W0$ and $W1$ represent the shared-weights of the fully connected network. Finally, $M_c(F^N)$ were applied to the F^N by the $\otimes$ operator to highlight the valuable building feature maps and suppress useless ones.

(b) Spatial attention module (sAM)

After the feature maps were independently recalibrated, the sAM process was performed to estimate a number of coefficients called spatial attention coefficients ($M_s(F^{N'})$), which was adopted to find some spatial positions of the $F^{N'}$ that contain important active building information. To do so, both average and maximum pooling operations were used here to determine the spatial importance level of the feature maps. However, the maximization or averaging calculations in this module were carried out at a spatial position of interest (i,j) along the depth of $F^{N'}$ (C) to achieve $F^s_{max} \in \mathbb{R}^{H\times W\times 1}$ and $F^s_{avg} \in \mathbb{R}^{H\times W\times 1}$ according to Eq. (4).

$$\begin{aligned} \mathrm{F}^{\mathrm{s}}_{\mathrm{max}}(\mathrm{i},\mathrm{j}) &= \mathrm{Max}\left(\mathrm{F}^{\mathrm{n}'}(\mathrm{i},\mathrm{j})\right) \\ \mathrm{F}^{\mathrm{s}}_{\mathrm{avg}}(\mathrm{i},\mathrm{j}) &= \frac{1}{\mathrm{N}'}\sum_{\mathrm{i}=1}^{\mathrm{H}}\sum_{\mathrm{j}=1}^{\mathrm{W}}\mathrm{F}^{\mathrm{n}'}(\mathrm{i},\mathrm{j}) \\ \mathrm{s}=1,\ 1<\mathrm{n}'<\mathrm{N}',\ 1<\mathrm{i}<\mathrm{H},\ 1<\mathrm{j}<\mathrm{W} & \end{aligned} \tag{4}$$

where, $F^{n'}(i,j)$ indicates the pixel value at the specified position (i,j) in the n' th $F^{N'}$.

Next, according to Fig. 2 (a) and Eq. (5), to produce an optimal descriptor, the two previously obtained single-channel descriptors (i.e., F^s_{max} and F^s_{avg}) were stacked, and a 1×1 CL was then applied to the stack to determine the spatial importance level of the features based on neighboring locations and also to generate a single descriptor from the F^s_{max} and F^s_{avg}. Using a 1×1 convolution operation greatly lowers computational complexity compared to a 7×7 operation (as used in (Woo et al., 2018)) while maintaining the effectiveness of the attention mechanism. The output was next rescaled between 0 and 1 using the sigmoid activation function. Ultimately, the estimated $M_s(F^{N'})$ were multiplicated to the $F^{N'}$ by the $\otimes$ operator to accentuate the critical spatial feature locations in feature maps and weaken ineffective ones.

$$M_s\left(F^{N'}\right) = \sigma\left(f\left[F^s_{max}; F^s_{avg}\right]\right) \tag{5}$$

where, the stacking procedure is signified with [(.); (.)] and f [.] is also the convolution operator.

(c) Residual block

While the cAM and sAM modules offer benefits in enhancing sensitivity to building features both spatially and in terms of channels, their incorporation into the first level of the standard U-Net architecture presents two noteworthy challenges in BM generation. The first issue is that the gradient vanishing problem arises due to the remarkable increment of CLs after adding the two aforementioned modules. In other words, the gradient values in the gradient descent process gradually damp during the backpropagation phase so that they reach almost zero. As a result, the network learning is not completely carried out. Moreover, the abundance of CLs introduces a second challenge: the potential loss of intricate building features, leading to the inadvertent propagation of inaccurate building information across the network during the training process. To cope with the two aforementioned challenges, the residual block as another part of the csAttResConvB was embedded after both cAM and sAM, as shown in Fig. 2 (a). This modification creates another independent gradient estimation path to the initial layers in order for the backpropagation stage to be fully completed. Additionally, the residual block avoids the loss of valuable building features and allows them to properly propagate in the proposed model.

2.1.1.2. Attention gate (AttG) architecture. The ordinary U-Net integrates the LLSF that contains useless non-building information extracted in the encoder part with the HLSF obtained in the decoder part. Integrating the LLSF and HLSF with no restriction in the model's encoder-decoder (skip-connection) path disrupts properly discriminating building objects from the heterogenous background. To minimize this limitation, an AttG inspired by Oktay et al. (2018) was designed according to Fig. 2 (b) and placed in the skip-connection path to simultaneously emphasize the satisfactory spatial information of LLSF while minimizing background non-building information.

To carry out this AttG-based refinement process, a gating signal ($g\epsilon\mathbb{R}^{h_g\times w_g\times n}$), incorporating beneficial HLSF was first attained. Afterwards, the attention coefficients ($\alpha\epsilon[0,1]$) was applied to n-layer LLSF ($x\epsilon\mathbb{R}^{h_x\times w_x\times n}$) using the $\otimes$ operator in order to compute the refined LLSF signified by x'. The entire AttG procedure can also be formulated as follows:

$$\begin{aligned} Q &= \phi\left(\sigma_1\left(W_g g + W_x x + b_g\right)\right) + b_\phi \\ \alpha &= \sigma_2(Q) \\ x' &= \alpha \otimes x \end{aligned} \tag{6}$$

where, σ_1 and σ_2 respectively denote the sigmoid and ReLU activation functions, and W_g and W_x signify the linear transformation weights of g and x, respectively. Further, ϕ is the convolutional operation and b_g and b_ϕ are also the bias parameters.

2.1.2. Data preparation, training parameter setting, and experimental environment

In the present study, to increase the amount of the available data and expedite the training processing, all the training/validation, testing, and their labels were partitioned into square patches of size 512×512 pixels without any overlaps. Additionally, 20% of the

training/validation patches were used for the model validation. Also, in order to increase the training data and prevent the overfitting problem, training patches were augmented with left/right, up/down, and angular rotations. Consequently, the shape of the training and testing datasets came out as (2852,512,512,3) and (20,512,512,3), respectively. To train the proposed DAttResU-Net with the prepared data, the adaptive moment estimation (ADAM) optimizer was utilized for gradient descent optimization.

Moreover, the class imbalance problem may arise in regions where the buildings typically cover only a small part of the image scene compared to the non-building background. To prohibit this issue, the combined loss function introduced in (Zhang et al., 2022), comprising both binary cross-entropy and Tversky loss (Salehi et al., 2017) functions, was employed to alleviate the effects of the imbalanced samples. Furthermore, an early stopping parameter for reducing the training time was used here to stop learning if no training improvement is observed after 5 consecutive epochs. The initial learning rate was also set as 0.001, and the ReduceLROnPlateau parameter was employed to reduce the learning rate by a factor of 0.33 when the performance improvement is negligible after 5 epochs. Finally, it is worth noting that all the training and testing procedures were performed using the Google Colaboratory Pro Python environment with 16 GB GPU (i.e., NVIDIA Tesla V100) and 25 GB of RAM.

2.2. Change map generation by comparing the bi-temporal building boxes (step 2)

In order to use the bi-temporal BMs derived from the proposed DAttResU-Net for the first time in the application of building change detection, this research generated BCMs through the widely used "class-object change detection (COCD)" (Chen et al., 2012) approach where radiometric differences in bi-temporal VHR-RSIs have no adverse impact on the produced BCMs. The COCD process in the present study was carried out in three main steps. In detail, in the first step, any building polygon in each predicted BM was bound with a bounding box in a way the box was considered to be the nearest neighbor of each polygon (Fig. 3 (a)). In the second step, the overlapping (intersection) area between the two matching building boxes in $\boldsymbol{BM_{t2}}$ and $\boldsymbol{BM_{t1}}$ was determined by calculating the intersection over the union (*IoU*) parameter according to Fig. 3 (b) and Eq. (7).

$$IoU = \frac{(Building\ Box\ Time1 \cap Building\ Box\ Time2)}{(Building\ Box\ Time1 \cup Building\ Box\ Time2)} = \frac{(area)}{(area1 + area2 - area)} \tag{7}$$

In the last step, in order to generate the BCM and identify the changed buildings, a specific threshold was applied to the IoU parameter. In this case, if the IoU value for two corresponding bounding boxes is greater than 60, the building is considered unchanged; otherwise, it is labeled as changed.

3. Dataset description

In order to assess the feasibility of the proposed DAttResU-Net in both BM and BCM generation, the WHU building change detection dataset (Ji et al., 2019) was used in this research.

These datasets contain bi-temporal aerial VHR images and their corresponding buildings/changed buildings labels with a spatial resolution of 0.2 m/pixel and a size of 15354 × 32507 acquired over Christchurch, New Zealand, in 2011 and 2016. The dataset itself was partitioned into four training and two testing sub-datasets.

In Fig. 4, the four training datasets chosen for 2011 and 2016, namely TVA-A, TVA-B, TVA-C, and TVA-D, are bolded in red, whereas the two testing areas signified as TA-A and TA-B are highlighted in yellow. The details of the sub-datasets are also listed in Table 1.

4. Evaluation metrics

To assess the proficiency of the proposed approach in generating both BM and BCM, a rigorous evaluation was conducted. This involved a thorough comparison between the polygons outlining the predicted buildings or changed buildings and those representing the corresponding reference buildings or changed buildings as documented in the reference BMs and BCMs. This meticulous examination aimed to quantify the method's effectiveness across both BM and BCM generation aspects. In this regard, the false negative rate (P_{FN}), false positive rate (P_{FP}), overall accuracy (OA), precision, recall, and f-score (F_s) measures were calculated as follows:

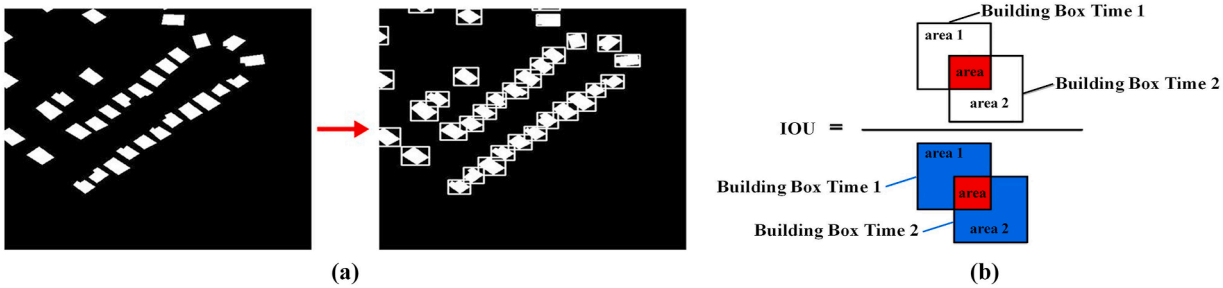

Fig. 3. (a) An example of creating bounding boxes for a binary BM; (b) Intersection over Union (IoU) calculation for corresponding bounding boxes.

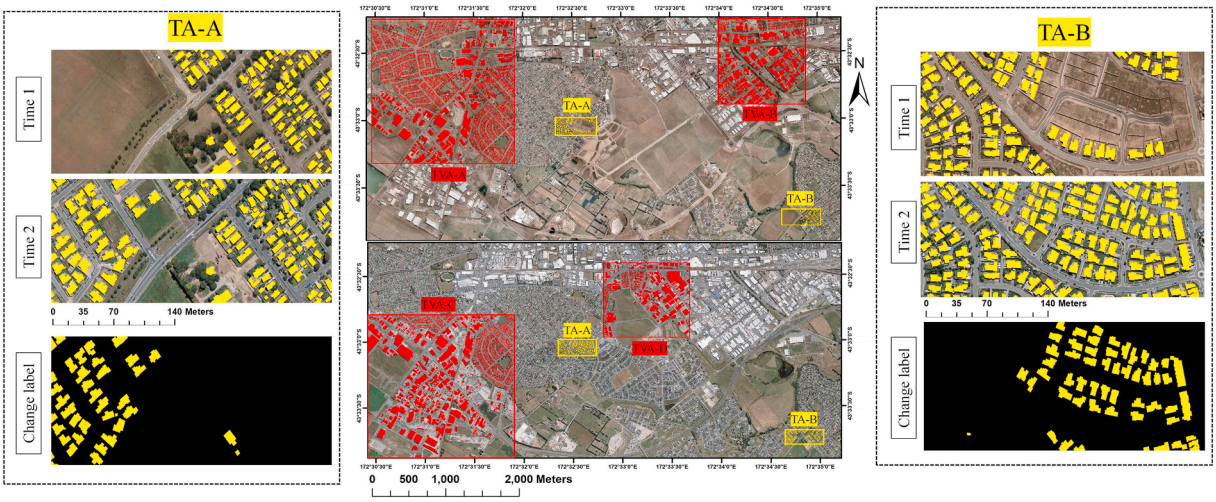

Fig. 4. The six sub-datasets selected from the WHU building dataset. The red boxes signify the training/validation sub-datasets, and the yellow boxes show the test sub-datasets.

Table 1
The details of the selected sub-datasets used in this research.

Datasets	GSD (m)	Area (km^2)	Tiles	Pixels	Box Color (Fig. 8)
TVA-A	0.2	20	361	512 × 512	Red
TVA-B	0.2	7.2	121	512 × 512	Red
TVA-C	0.2	20	361	512 × 512	Red
TVA-D	0.2	5.9	99	512 × 512	Red
TA-A	0.2	0.5	10	512 × 512	Yellow
TA-B	0.2	0.5	10	512 × 512	Yellow

$$P_{FN} = \left(\frac{N_{FN}}{N_{FN} + N_{TP}}\right) \times 100\ \% \tag{8}$$

$$P_{FP} = \left(\frac{N_{FP}}{N_{FP} + N_{TP}}\right) \times 100\ \% \tag{9}$$

$$OA = \left(\frac{N_{TP}}{N_{TP} + N_{FP} + N_{FN}}\right) \times 100\ \% \tag{10}$$

$$\text{Precision} = \left(\frac{N_{TP}}{N_{TP} + N_{FP}}\right) \times 100\ \% \tag{11}$$

$$\text{Recall} = \left(\frac{N_{TP}}{N_{TP} + N_{FN}}\right) \times 100\ \% \tag{12}$$

$$F_s = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{13}$$

where, true positives (N_{TP}), false negatives (N_{FN}), and false positives (N_{FP}) signify the number of correctly detected buildings/changed buildings, the number of missed buildings/changed buildings, and the number of false buildings/changed buildings, respectively. These three metrics are achieved with a confusion matrix defined in Table 2 in which the 'Actual labels' refer to the labels of the ground-truth BM/BCMs, and 'Predicted labels' characterize the labels of the generated BM/BCMs.

Table 2
The confusion matrix.

		Predicted labels	
Actual labels	Category	Building/changed	Non-building/unchanged
	Building/changed	N_{TP}	N_{FN}
	Non-building/unchanged	N_{FP}	N_{TN}

5. Experimental results and discussion

In the present study, in order to reveal how robust proposed DAttResU-Net is in both BM and BCM extraction tasks, its performance was quantitively and qualitatively compared with five state-of-the-art semantic segmentation architectures, including traditional U-Net (Ronneberger et al., 2015), residual U-Net (ResU-Net) (Zhang et al., 2018), and LinkNet (Chaurasia and Culurciello, 2017). Furthermore, three ablation experiment scenarios were designed to establish how the attention mechanisms and the residual block align contribute to improving the typical U-Net model. In the first scenario, adding AttG to the encoder-decoder path (i.e., second level) of the U-Net structure, signified by encoder-decoder attention U-Net (EDAttU-Net), was evaluated to see if it could improve U-Net. In the second scenario, using cAM and sAM along with the residual block rather than the simple CLs in the U-Net structure (i.e., first level), indicated by residual convolutional block attention U-Net (ResCBAttU-Net), was inspected. In the last scenario, the performance of simultaneously using the first and second scenarios as the proposed architecture (i.e., DAttResU-Net) was investigated in BM generation.

The outcomes of this research work are presented and discussed in three subsections. As for the first subsection, all the compared networks were evaluated in the training phase for a number of pre-defined training parameters. In the second subsection, the binary bi-temporal BMs predicted by the compared networks were analyzed and discussed qualitatively and quantitatively. Eventually, the BCMs produced by the compared networks were evaluated qualitatively and quantitatively in the last subsection to also compare how important the qualities of bi-temporal BMs can be to reach an optimal BCM.

5.1. Training performance evaluation of the compared models

In order to peruse whether the overfitting issue occurs during the training stage and also to determine the optimal training epoch for each DL network, the training and validation accuracy graphs of the model were plotted for different training epochs, as shown in Fig. 5. Moreover, to ensure the results of all the networks are impartially compared, the same structure and hyperparameters (see Table 3) were defined for the models.

From Fig. 5, in contrast to the training accuracy graphs exhibiting almost similar behavior against different epochs, the validation accuracy graph, which is the actual indicator of a DL model's reliability, revealed the least variation for the proposed DAttResU-Net compared to U-Net, ResU-Net, EDAttU-Net, ResCBAttU-Net, and LinkNet. This indicates the high stability and generalizability of DAttResU-Net on new unseen test datasets.

5.2. Building extraction

To better visualize the BMs produced by the comparative networks, i.e., DAttResU-Net, ResCBAttU-Net, EDAttU-Net, ResU-Net, U-Net, and LinkNet, the binary maps obtained from the models for the first and second study areas are respectively superimposed on the original VHR images as seen in Figs. 7 and 8, in which the best and worst qualitative results for each study area are also presented in

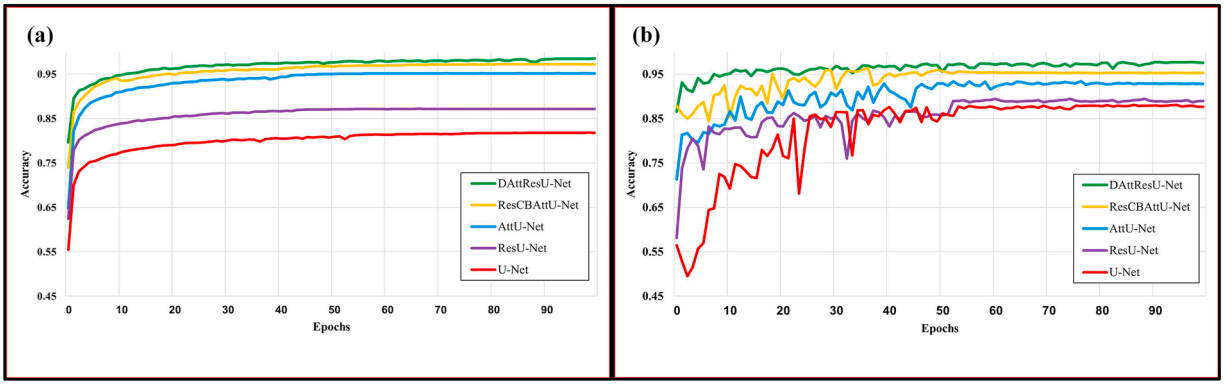

Fig. 5. The (a) training and (b) validation data-based accuracy graphs of the proposed DAttResU-Net against ResCBAttU-Net, EDAttU-Net, ResU-Net, U-Net, and LinkNet.

Table 3
The regularization parameters set for the comparative DL models.

Model	Initial Learning rate	Batch size	Maximum epochs	Best Epochs	No. of Parameters ($\times 10^6$)
LinkNet	0.0001	16	100	82	20.33
U-Net	0.0001	16	100	42	2.165
ResU-Net	0.0001	16	100	87	2.255
EDAttU-Net	0.0001	16	100	39	2.342
ResCBAttU-Net	0.0001	16	100	37	2.284
DAttResU-Net	0.0001	16	100	92	2.658

addition to the superimposed images. Furthermore, Table 4 and Fig. 6 demonstrate the quantitative results derived by comparing the obtained binary maps to the reference BMs.

From the quantitative outcomes listed in Tables 4 and it can be perceived that the proposed DAttResU-Net yielded the most satisfactory BM compared with the other U-Net-based models for both study areas. In detail, for the first study area, at time 1, DAttResU-

Table 4
Quantitative performance evaluation of the comparative DL models in BM generation (the numbers in bold represent the best value in each column).

study area	Time	Total Building (No.)	Model	Evaluation Metrics						
				TP (No.)	FN (No.)	FP (No.)	OA (%)	Precision (%)	Recall (%)	F_s (%)
#1	#1	94	U-Net	67	27	1	70.53	98.53	71.28	82.72
			ResU-Net	85	9	3	87.63	96.59	90.43	93.41
			EDAttU-Net	86	8	4	87.76	95.56	91.49	93.48
			ResCBAttU-Net	87	7	1	91.58	98.86	92.55	95.60
			LinkNet	90	4	2	93.75	97.83	95.74	96.77
			DAttResU-Net	**92**	**2**	**1**	**96.84**	**98.92**	**97.87**	**98.40**
	#2	130	U-Net	79	51	3	59.40	96.34	60.77	74.53
			ResU-Net	119	11	3	89.47	97.54	91.54	94.44
			EDAttU-Net	119	11	**1**	90.84	**99.17**	91.54	95.20
			ResCBAttU-Net	125	5	4	93.28	96.90	96.15	96.53
			LinkNet	128	2	3	96.24	97.71	98.46	98.08
			DAttResU-Net	**128**	**2**	2	**96.97**	98.46	**98.46**	**98.46**
#2	#1	82	U-Net	44	38	6	50.00	88.00	53.66	66.67
			ResU-Net	66	16	4	76.74	94.29	80.49	86.84
			EDAttU-Net	71	11	4	82.56	94.67	86.59	90.45
			ResCBAttU-Net	73	9	4	84.88	94.81	89.02	91.82
			LinkNet	77	5	3	90.59	96.25	93.90	95.06
			DAttResU-Net	**78**	**4**	**1**	**93.98**	**98.73**	**95.12**	**96.89**
	#2	134	U-Net	90	44	1	66.67	98.90	67.16	80.00
			ResU-Net	102	32	4	73.91	96.23	76.12	85.00
			EDAttU-Net	110	24	2	80.88	98.21	82.09	89.43
			ResCBAttU-Net	110	24	2	91.49	94.85	96.27	95.56
			LinkNet	131	3	3	95.62	97.76	97.76	97.76
			DAttResU-Net	**133**	**1**	**2**	**97.79**	**98.52**	**99.25**	**98.88**

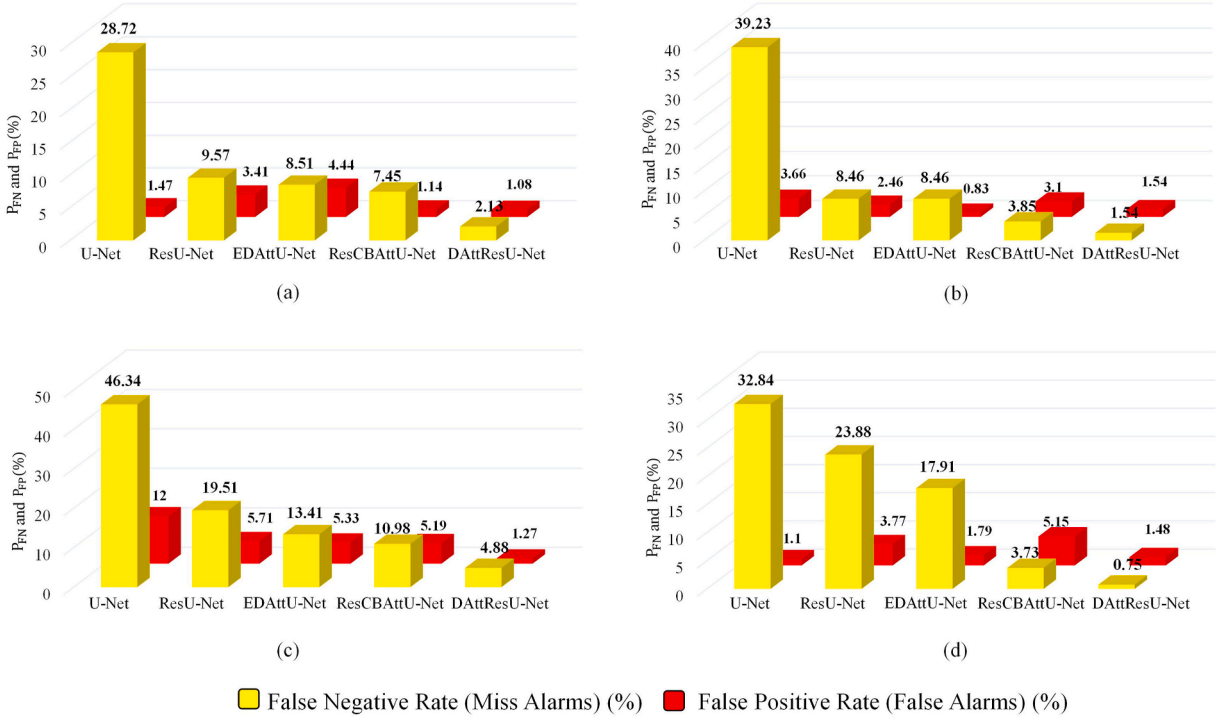

Fig. 6. The false negative rate and false positive rate bar graphs for the bi-temporal BMs attained through various DL models: (a) and (b) the results of the first and second time for the first study area, (c) and (d) the results of the first and second time for the second study area.

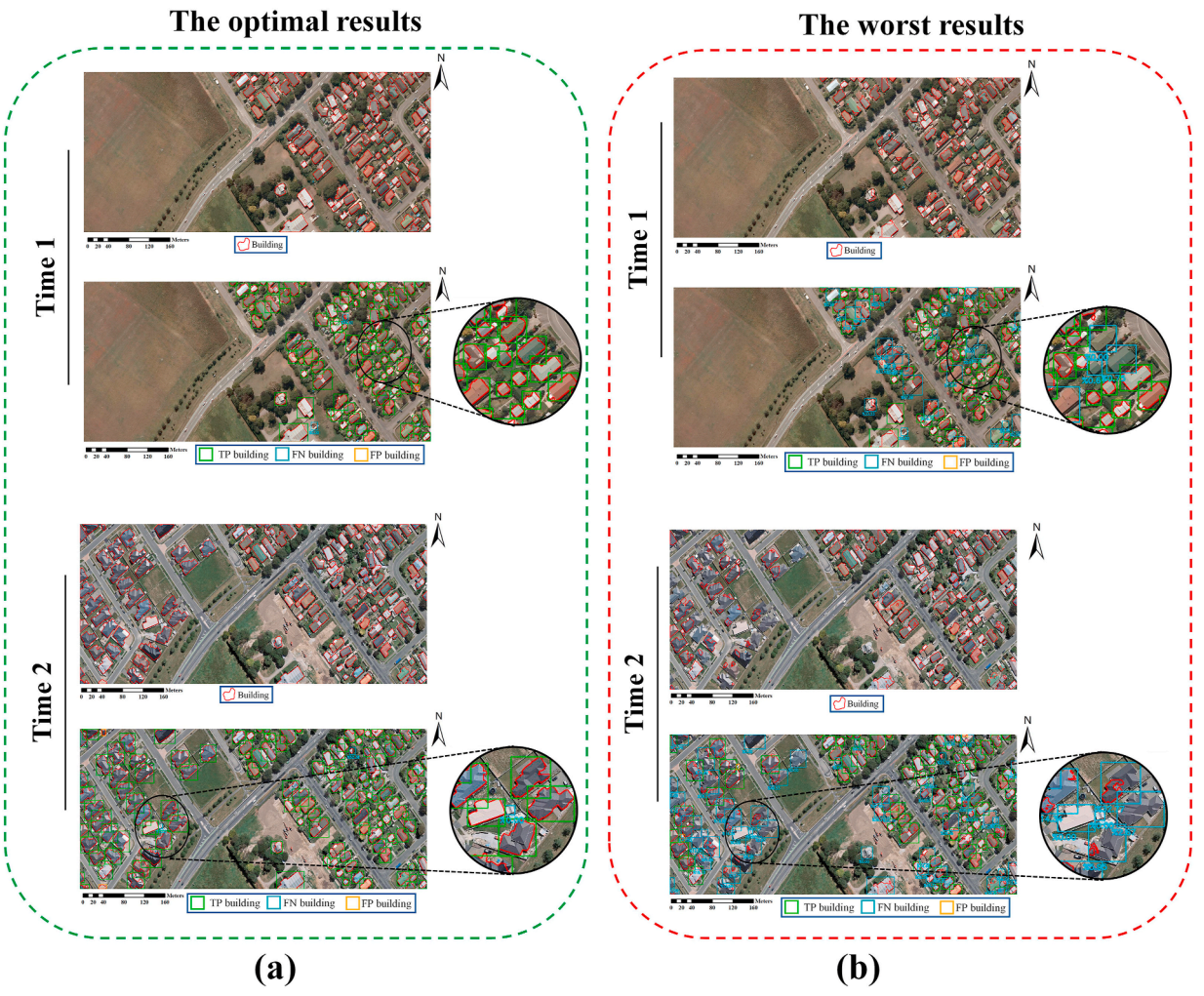

Fig. 7. The best and worst bi-temporal BMs produced for the first study area: (a) DAttResU-Net and (b) U-Net, where the first line for each time represents the binary BM superimposed on the original image and the second line signify the analyzed and evaluated BM.

Net revealed the highest OA/ F_s values of 96.84/98.40 (%) with improvements of 26.31/15.68, 9.21/4.99, 9.08/4.92, 5.26/2.80, and 3.09/1.63 (%) compared to the U-Net, ResU-Net, EDAttU-Net, ResCBAttU-Net, and LinkNet respectively. For the second study area, at time 1, the proposed DAttResU-Net with a Recall value of 95.12% could enhance the Recall values of the U-Net, ResU-Net, EDAttU-Net, ResCBAttU-Net, and LinkNet by 41.46, 14.63, 8.53, 6.10, and 1.22 (%), respectively.

Furthermore, according to Fig. 6, ablation results for the second scenario (i.e., ResCBAttU-Net) showed that the modifications made to the typical U-Net resulted in degrading P_{FN} values at any time of the two study areas, indicating the combined use of the residual unit, cAM, and sAM successfully managed to both preserve and accentuate the building objects. As for the ablation outcomes for the first scenario (i.e., EDAttU-Net), similar to the second scenario, the AttG module improved the P_{FN} in all cases compared to typical U-Net, and although the module reduced the P_{FP} values only in the second/first time of the first/second study areas, the integration of the AttG module with the residual unit, cAM, and sAM within the proposed DAttResU-Net totally outperformed both ResCBAttU-Net and EDAttU-Net models. Extensively, the DAttResU-Net with an average P_{FN}/ P_{FP} value of 2.33/1.34 (%), over the first and second times of the two study areas diminished the average P_{FN}/ P_{FP} values of ResCBAttU-Net and EDAttU-Net by 4.17/2.31 (%) and 9.74/1.76 (%), respectively. Moreover, the least difference value between P_{FN} and P_{FP} for the proposed network asserts that DAttResU-Net can make the best trade-off between the two aforementioned metrics compared to both ResCBAttU-Net and EDAttU-Net, indicating the proposed structure has the highest accuracy, consistency, and reliability when separating the building objects from the background. At last, the superiority of the proposed DL architecture over the compared structures is that the model exploits two effective attention modules alongside the residual unit to properly identify the building objects.

In more detail, attention modules embedded at the first and second levels of the U-Net, respectively called csAttResConvB and AttG, simultaneously not only raise the network's attention to the building information such as edge, texture, shape, and color but also restrict the non-building features circulation during the training phase. These cases lead to optimal BMs containing maximum TP and minimum FP, as acknowledged by the quantitative results reported in Table 4.

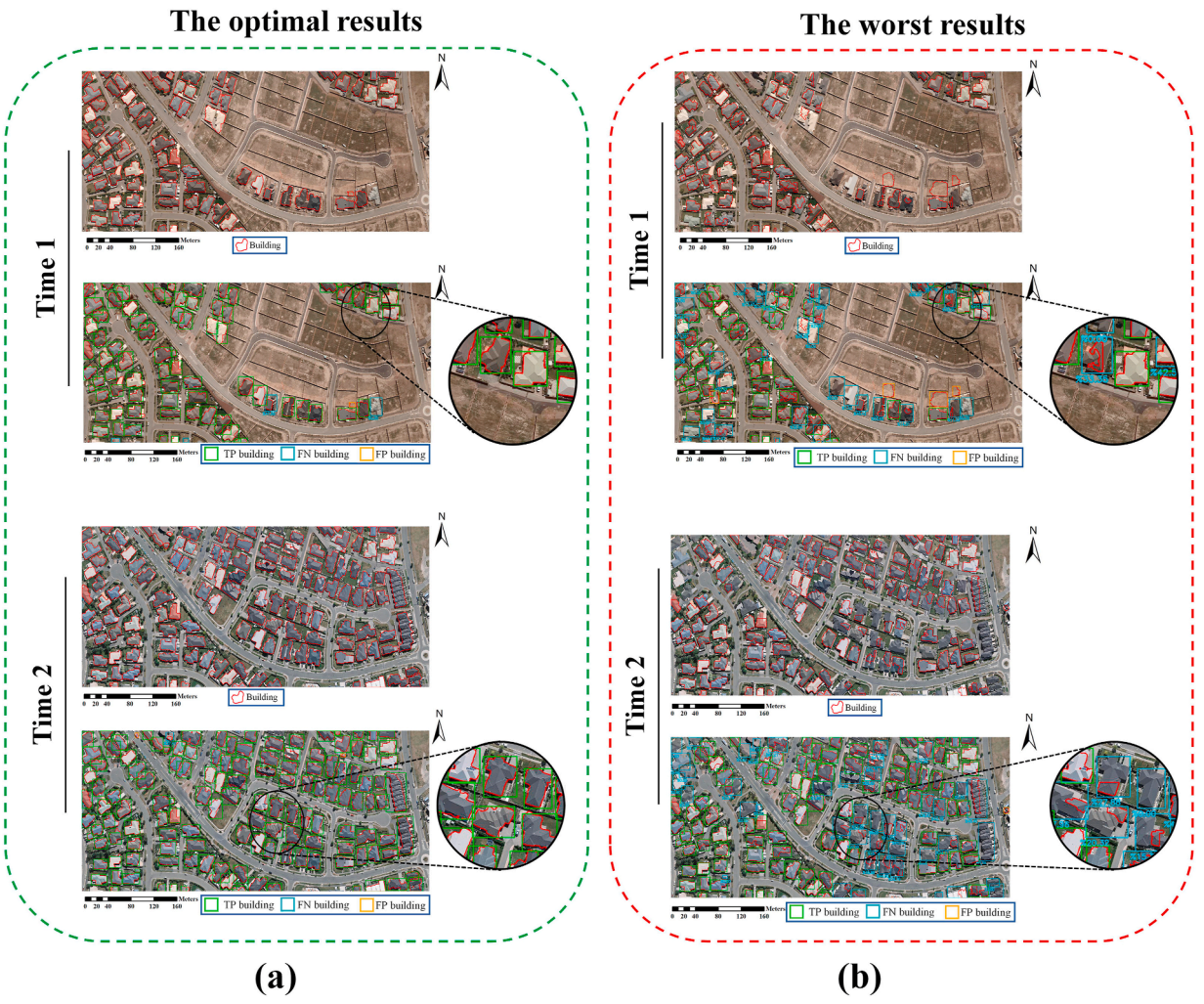

Fig. 8. The best and worst bi-temporal BMs produced for the second study area: (a) DAttResU-Net and (b) U-Net, where the first line for each time represents the binary BM superimposed on the original image and the second line signify the analyzed and evaluated BM.

As depicted in Figs. 7 and 8, the traditional U-Net produced the worst BMs for the two study areas, where many buildings were lost. Especially according to the zoomed areas, the model either fully missed some buildings or partially extracted them. Additionally, as can be seen, the first time for the second study area (Fig. 8b, yellow boxes), the U-Net architecture falsely identified some non-building features as building objects. These issues are mainly because the U-Net lacks effective mechanisms at the different levels of its architecture. On the one hand, the CBs in U-Net have no attention module to highly emphasize the building features surrounded by non-building ones.

On the other hand, there is no attention module at the second level (i.e., skip-connection path) of the U-Net structure to concentrate on important LLSF and filter irrelevant useless ones. Contrary to the U-Net model, the novel DAttResU-Net structure coincidently exploits powerful proposed csAttResConvB and AttG at both structural levels. Firstly, the csAttResConvB module used at the first level puts the detailed important building HLSF under scrutiny, making the model identify the building objects more accurately, as can be observed in the zoomed regions. Secondly, incorporating the AttG mechanism at the second level prevents the inclusion of extraneous non-building LLSF alongside the crucial HLSF. This strategic integration minimizes non-building artifacts' presence and significantly reduces their propagation throughout the network. This enhancement substantially contributes to the production of BMs with a notably higher degree of accuracy and fidelity.

5.3. Building change detection

In order to better visualize the COCD-based BCMs related to various DL models, both the total and binary BCMs for the first and second study areas are superimposed on the VHR image acquired at time 2 as represented in Figs. 10 and 11. The total BCMs show the three types of changes, including the negative and positive changes as well as the unchanged buildings, whereas the binary BCMs solely demonstrate two changed and unchanged categories. Obviously, since the bi-temporal change detection procedure for BCM

generation is achieved by directly comparing the two BMs previously predicted by a DL model, the quality of the bi-temporal BMs as inputs to this comparison is of high importance to appropriately characterize the changed and unchanged buildings.

As accordingly expected, the proposed DAttResU-Net based on quantitative results reported in Table 5 yielded the most accurate BCM in comparison with the other DL models. In detail, in the first/second study area, the BCM derived from the proposed model led to the highest recall/precision values of 94.74/89.83 (%) enhancing the recall/precision values of the U-Net, ResU-Net, EDAttU-Net, ResCBAttU-Net, and LinkNet by 60.53/33.53 (%), 15.79/31.65 (%), 18.42/7.61 (%), 2.63/17.37, and 7.24/1.9 (%), respectively.

Moreover, as depicted in Fig. 9, the BCM attained by the proposed model performed best in detecting the changed buildings with a minimum noise over both study areas compared to the other models. For example, in the first study area, DAttResU-Net-derived BCM contained the least P_{FN}/ P_{FP} values of 5.26/7.69 (%), substantially improving those of U-Net, ResU-Net, EDAttU-Net, ResCBAttU-Net, and LinkNet by 60.53/48.98 (%), 15.79/22.54 (%), 18.42/21.58 (%), 2.63/12.76, and 7.24/0.2 (%). As noticed from analyzing these quantitative outcomes, the quality of BCMs is drastically influenced by the P_{FN} and P_{FP} values of the bi-temporal BMs previously predicted by the DL models, indicating how effective a BM generation network is in the bi-temporal change detection application. Hence, due to exploiting strong building-sensitive mechanisms, the DAttResU-Net structure proposed in this study contained minimum errors in bi-temporal BMs and, subsequently, BCM generation.

From Figs. 10 and 11, notably in the zoomed areas, the BCMs indirectly obtained from the ordinary U-Net had some artifacts and failed to identify many changed buildings. This is due to the buildings partially or fully overlooked in previously U-Net-derived bi-temporal BMs for the first and second study areas, as the model inherently contains low sensitivity to the building objects. In contrast to the simple U-Net, since the proposed DAttResU-Net structure leverages two powerful attention modules highly sensitive to buildings and thereby leads to accurate bi-temporal BMs, the BCM related to the model well discriminated the changed buildings from the unchanged ones, revealing the proposed structure can be an efficient choice to produce reliable BCMs.

6. Conclusion

The present study proposed a novel DL U-Net-based approach, namely dual attention residual-based U-Net (DAttResU-Net), to generate reliable bi-temporal BMs and peruse their applicability to produce a BCM. The network simultaneously exploits two new

Table 5
Quantitative performance evaluation of the comparative DL models in BCM generation (the numbers in bold represent the best value in each column).

study area	Ground Truth and Models	No Change (No.)	Pos. Change (No.)	Neg. Change (No.)	Evaluation Metrics						
					TP (No.)	FN (No.)	FP (No.)	OA (%)	Precision (%)	Recall (%)	F_s (%)
#1	**GT**	92	36	2							
	U-Net	64	56	11	13	25	17	23.64	43.33	34.21	38.24
	ResU-Net	78	49	6	30	8	13	58.82	69.77	78.95	74.07
	EDAttU-Net	79	44	7	29	9	12	58.00	70.73	76.32	73.42
	ResCBAU-Net	78	42	3	35	3	9	74.47	79.55	92.11	85.37
	LinkNet	93	30	1	35	5	3	81.40	92.11	87.50	89.74
	DAttResU-Net	92	36	2	**36**	**2**	**3**	**87.80**	**92.31**	**94.74**	**93.51**
#2	**GT**	81	53	0							
	U-Net	37	54	13	27	26	22	36.00	56.10	50.94	52.94
	ResU-Net	55	85	9	32	21	23	42.11	58.18	60.38	59.26
	EDAttU-Net	68	65	5	37	16	8	60.66	82.22	69.81	75.51
	ResCBAU-Net	65	59	10	50	3	19	69.44	72.46	94.34	81.97
	LinkNet	75	54	3	51	2	7	85.00	87.93	96.23	91.89
	DAttResU-Net	74	58	1	**53**	**0**	**6**	**89.83**	**89.83**	**100.00**	**94.64**

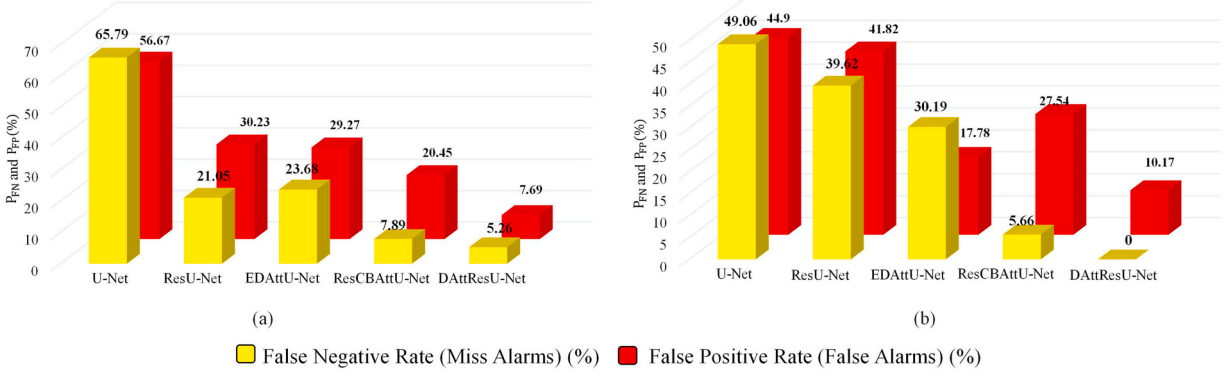

Fig. 9. The false negative rate and false positive rate bar graphs for the BCMs attained through various DL models: (a) first study area and (b) second study area.

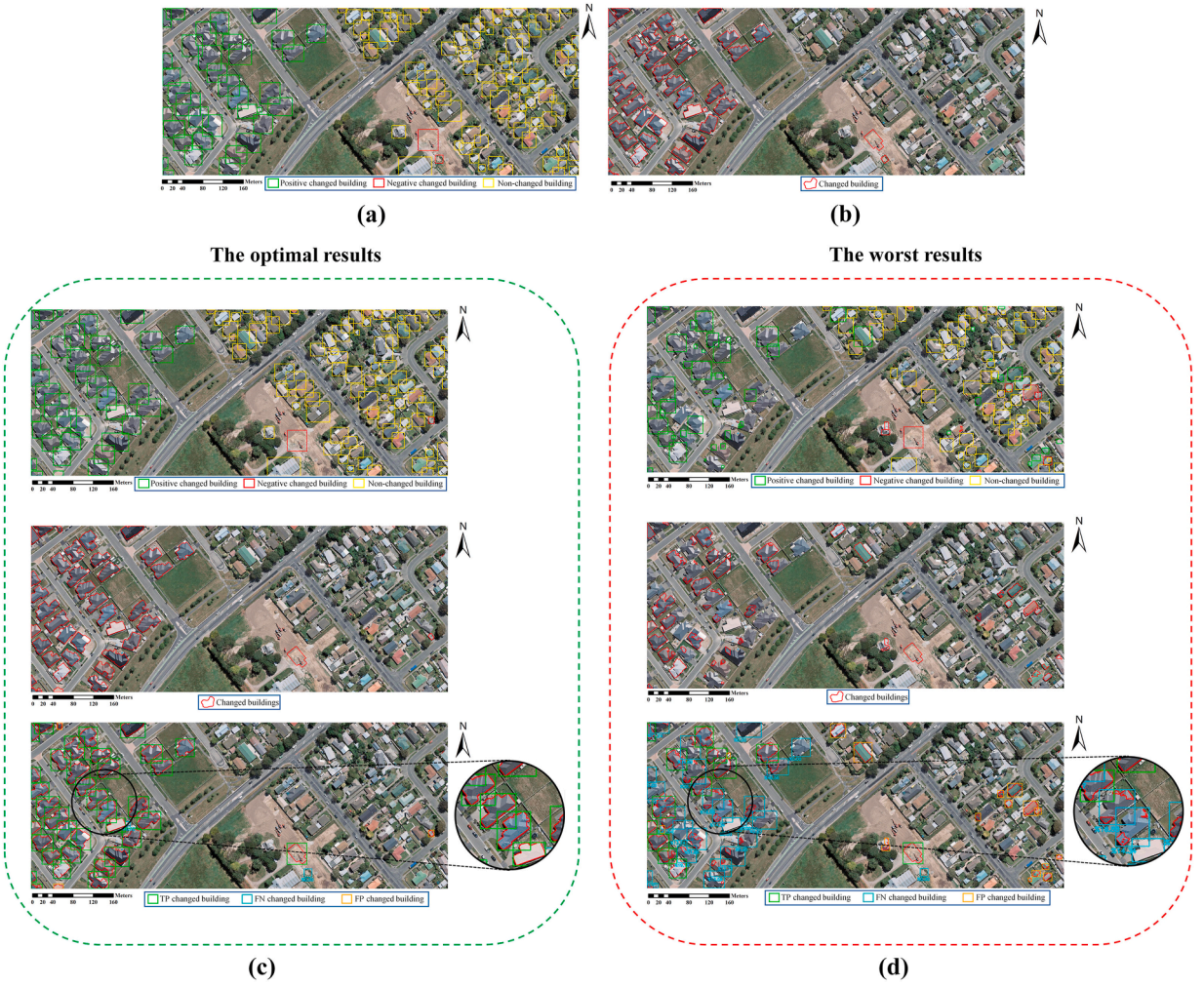

Fig. 10. The best and worst bi-temporal BCMs derived from the comparative DL models for the first study area: (a) the total ground-truth BCM, (b) the binary ground-truth BCM, (c) the optimal results obtained from DAttResU-Net, (d) the worst results output by U-Net, where the total BCM, binary BCM, and evaluated BCMs all superimposed on the original image are respectively represented in the first, second, and third lines.

strong semantic- and spatial-sensitive attention mechanisms at two different structural levels of the U-Net. In the first level, in order to increase the attention of the proposed model to the building high-level semantic information while preserving them, a channel-spatial attention residual convolutional block (csAttResConvB) replaced the simple convolutional blocks (CBs) of U-Net. In the second level, to raise the model's concentration on the valuable building low-level spatial features and filter irrelevant useless ones, another attention module, called attention gate (AttG), was embedded in the skip-connection path of U-Net. After designing the proposed DAttResU-Net architecture as an application for building change detection, the bi-temporal BMs produced by the model were given to the broadly used class-object change detection approach to create a BCM for a region of interest. To this aim, building boxes for buildings in any of the bi-temporal BMs were first extracted, and the BCM was then created by correspondingly comparing the bi-temporal building boxes. In this paper, the benchmark WHU building change detection dataset was also used to quantitatively and qualitatively assess the performance of the DAttResU-Net model in the BM and BCM generation domains. As for producing BMs, the experimental results showed that the proposed architecture, due to leveraging powerful dual csAttResConvB and AttG, sensed the building objects accurately with a minimum amount of background features. Consequently, the DAttResU-Net could produce the most high-quality bi-temporal BMs in comparison with the state-of-the-art DL structures. The BCM derived from the proposed architecture resulted in the most precise exceptional change detection results in both study areas, followed by predicting reliable BMs. Accordingly, the novel DAttResU-Net revealed a high capability to detect building objects and applicably identify the changed buildings during the time and is also suggested to be evaluated in other remote sensing applications. While the proposed model shows promising results, several limitations could be addressed for future research. For instance, the integration of dual attention mechanisms and residual blocks increases the computational load and training time. Exploring optimization techniques or model simplifications could make the model more efficient. Moreover, investigating the model's performance with other types of remote sensing data, such as synthetic aperture radar (SAR) or multi-spectral images, could enhance its applicability and utility in different remote sensing scenarios.

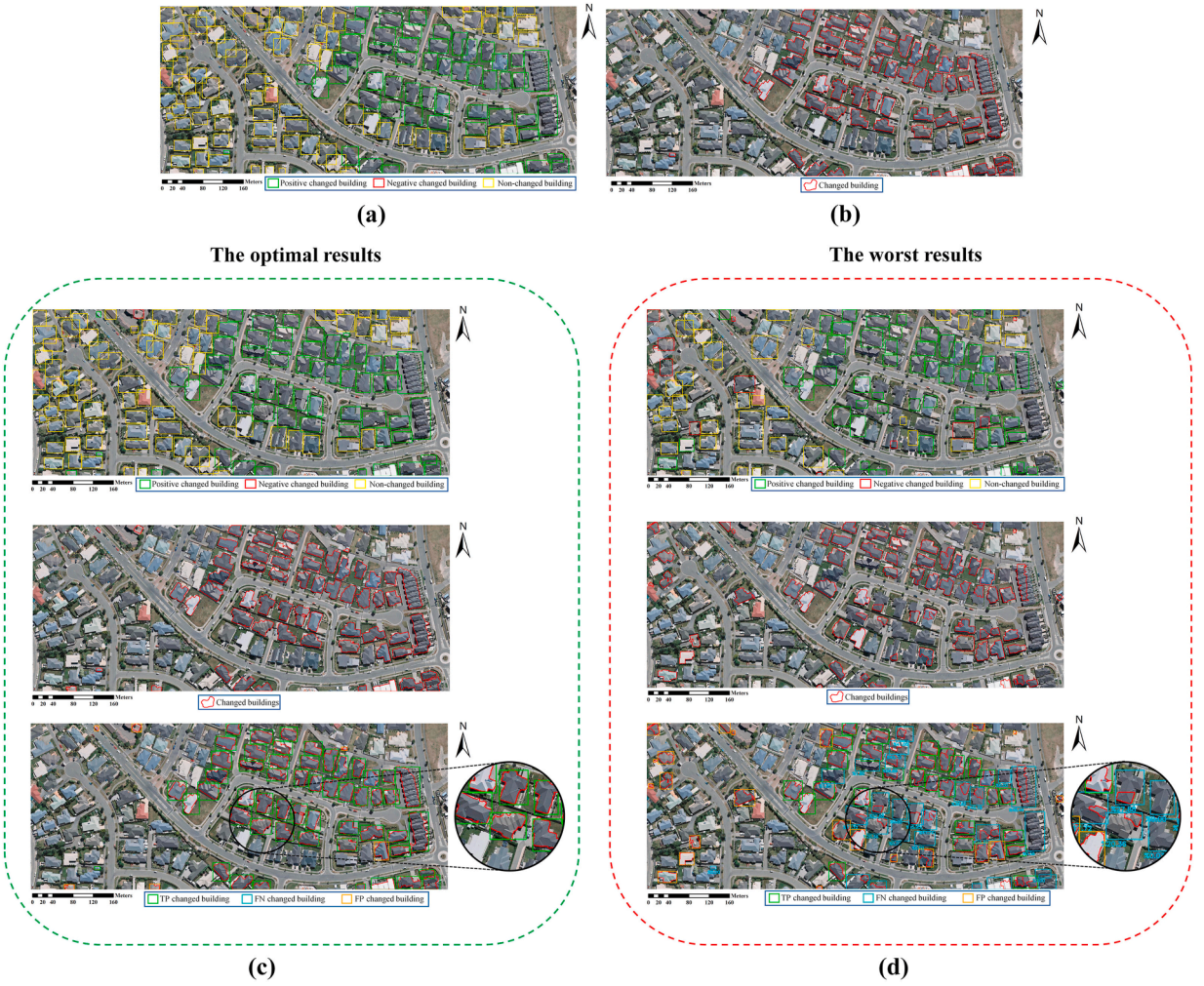

Fig. 11. The best and worst bi-temporal BCMs derived from the comparative DL models for the second study area: (a) the total ground-truth BCM, (b) the binary ground-truth BCM, (c) the optimal results obtained from DAttResU-Net, (d) the worst results output by U-Net, where the total BCM, binary BCM, and evaluated BCMs all superimposed on the original image are respectively represented in the first, second, and third lines.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial.

Data availability

The WHU benchmark building detection datasets analyzed during the current study are available on https://paperswithcode.com/dataset/whu-building-dataset (accessed on August 10, 2024)

Disclosure statement

The authors reported no potential conflict of interest.

Ethical Statement for Solid State Ionics

Hereby, We/Ehsan Khankeshizadeh, Ali Mohammadzadeh*, Amin Mohsenifar, Armin Moghimi, Saied Pirasteh, Sheng Feng, Keli Hu, Jonathan Li/consciously assure that for the manuscript/Building Detection in VHR Remote Sensing Images using a Novel Dual Attention Residual-based U-Net (DAttResU-Net): An Application to Generating Building Change Maps/the following is fulfilled.

1) This material is the authors' own original work, which has not been previously published elsewhere.
2) The paper is not currently being considered for publication elsewhere.
3) The paper reflects the authors' own research and analysis in a truthful and complete manner.
4) The paper properly credits the meaningful contributions of co-authors and co-researchers.

5) The results are appropriately placed in the context of prior and existing research.
6) All sources used are properly disclosed (correct citation). Literally copying of text must be indicated as such by using quotation marks and giving proper reference.
7) All authors have been personally and actively involved in substantial work leading to the paper, and will take public responsibility for its content.

The violation of the Ethical Statement rules may result in severe consequences.

To verify originality, your article may be checked by the originality detection software iThenticate. See also http://www.elsevier.com/editors/plagdetect.

I agree with the above statements and declare that this submission follows the policies of Solid State Ionics as outlined in the Guide for Authors and in the Ethical Statement.

CRediT authorship contribution statement

Ehsan Khankeshizadeh: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ali Mohammadzadeh:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Amin Mohsenifar:** Writing – review & editing, Methodology, Conceptualization. **Armin Moghimi:** Writing – review & editing, Methodology, Conceptualization. **Saied Pirasteh:** Writing – review & editing. **Sheng Feng:** Writing – review & editing. **Keli Hu:** Writing – review & editing. **Jonathan Li:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Alsabhan, W., Alotaiby, T., Dudin, B., 2022. Detecting buildings and nonbuildings from satellite images using U-net. Comput. Intell. Neurosci. 2022. https://doi.org/10.1155/2022/4831223.

Awrangjeb, M., Zhang, C., Fraser, C.S., 2013. Improved building detection using texture information. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. XXXVIII-3/W22. https://doi.org/10.5194/isprsarchives-xxxviii-3-w22-143-2011.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12). https://doi.org/10.1109/TPAMI.2016.2644615.

Chaurasia, A., Culurciello, E., 2017. LinkNet: exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing. VCIP 2017. https://doi.org/10.1109/VCIP.2017.8305148 2018-January.

Chen, G., Hay, G.J., Carvalho, L.M.T., Wulder, M.A., 2012. Object-based change detection. In: International Journal of Remote Sensing, 33. Taylor and Francis Ltd, pp. 4434–4457. https://doi.org/10.1080/01431161.2011.648285 14.

Chen, H., Lu, S., 2019. Building extraction from remote sensing images using SegNet. In: 2019 IEEE 4th International Conference on Image, Vision and Computing. ICIVC 2019. https://doi.org/10.1109/ICIVC47709.2019.8981046.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogrammetry Remote Sens. 195. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Farhadi, H., Ebadi, H., Kiani, A., 2023. F2BFE: development of feature-based building footprint extraction by remote sensing data and GEE. Int. J. Rem. Sens. 44 (19). https://doi.org/10.1080/01431161.2023.2255351.

Farnood Ahmadi, F., Naanjam, R., Salimi, A., 2024. Developing an automatic training technique based on integration of radar and optical remotely sensed images for building extraction. Earth Science Informatics 17 (1). https://doi.org/10.1007/s12145-023-01154-w.

Feng, W., Sui, H., Hua, L., Xu, C., Ma, G., Huang, W., 2020. Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map. Int. J. Rem. Sens. 41 (17). https://doi.org/10.1080/01431161.2020.1742944.

Ferraioli, G., 2010. Multichannel InSAR building edge detection. IEEE Trans. Geosci. Rem. Sens. 48 (3 PART 1). https://doi.org/10.1109/TGRS.2009.2029338.

Fuentes Reyes, M., Xie, Y., Yuan, X., d'Angelo, P., Kurz, F., Cerra, D., Tian, J., 2023. A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection. ISPRS J. Photogramm. Rem. Sens. 205, 74–97. https://doi.org/10.1016/J.ISPRSJPRS.2023.09.013.

Guo, M., Liu, H., Xu, Y., Huang, Y., 2020. Building extraction based on U-net with an attention block and multiple losses. Rem. Sens. 12 (9). https://doi.org/10.3390/RS12091400.

Ivanovsky, L., Khryashchev, V., Pavlov, V., Ostrovskaya, A., 2019. Building detection on aerial images using U-NET neural networks. Conference of Open Innovation Association, FRUCT, 2019-April. https://doi.org/10.23919/FRUCT.2019.8711930.

Ji, S., Wei, S., 2019. Building extraction via convolutional neural networks from an open remote sensing building dataset. Cehui Xuebao/Acta Geodaetica et Cartographica Sinica 48 (4). https://doi.org/10.11947/j.AGCS.2019.20180206.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Rem. Sens. 57 (1), 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Kang, W., Xiang, Y., Wang, F., You, H., 2019. EU-Net: an efficient fully convolutional network for building extraction from optical remote sensing images. Rem. Sens. 11 (23). https://doi.org/10.3390/rs11232813.

Kaya, Y., Şenol, H.İ., Yiğit, A.Y., Yakar, M., 2023. Car detection from very high-resolution UAV images using deep learning algorithms. Photogramm. Eng. Rem. Sens. 89 (2). https://doi.org/10.14358/PERS.22-00101R2.

Khankeshizadeh, E., Mohammadzadeh, A., Arefi, H., Mohsenifar, A., Pirasteh, S., Fan, E., Li, H., Li, J., 2024. A novel weighted ensemble transferred U-net based model (WETUM) for post-earthquake building damage assessment from uav data: a comparison of deep learning- and machine learning-based approaches. IEEE Trans. Geosci. Rem. Sens. https://doi.org/10.1109/TGRS.2024.3354737.

Khankeshizadeh, E., Mohammadzadeh, A., Moghimi, A., Mohsenifar, A., 2022. FCD-R2U-net: forest change detection in bi-temporal satellite images using the recurrent residual-based U-net. Earth Science Informatics 15 (4). https://doi.org/10.1007/s12145-022-00885-6.

Lei, J., Liu, X., Yang, H., Zeng, Z., Feng, J., 2024. Dual hybrid attention mechanism-based U-net for building segmentation in remote sensing images. Appl. Sci. 14 (3). https://doi.org/10.3390/app14031293.

Li, C., Fu, L., Zhu, Q., Zhu, J., Fang, Z., Xie, Y., Guo, Y., Gong, Y., 2021a. Attention enhanced u-net for building extraction from farmland based on google and

worldview-2 remote sensing images. Rem. Sens. 13 (21). https://doi.org/10.3390/rs13214411.
Li, C., Liu, Y., Yin, H., Li, Y., Guo, Q., Zhang, L., Du, P., 2021b. Attention residual U-net for building segmentation in aerial images. In: International Geoscience and Remote Sensing Symposium (IGARSS). https://doi.org/10.1109/IGARSS47720.2021.9554058.
Manno-Kovacs, A., Sziranyi, T., 2015. Orientation-selective building detection in aerial images. ISPRS J. Photogramm. Rem. Sens. 108. https://doi.org/10.1016/j.isprsjprs.2015.06.007.
Moghimi, A., Welzel, M., Celik, T., Schlurmann, T., 2024. A comparative performance analysis of popular deep learning models and segment anything model (SAM) for river water segmentation in close-range remote sensing imagery. IEEE Access 12, 52067–52085. https://doi.org/10.1109/ACCESS.2024.3385425.
Naanjam, R., Farnood Ahmadi, F., 2024. An improved self-training network for building and road extraction in urban areas by integrating optical and radar remotely sensed data. Earth Science Informatics 17 (3). https://doi.org/10.1007/s12145-024-01270-1.
Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 1520–1528. https://doi.org/10.1109/ICCV.2015.178.
Oktay, O., Schlemper, J., Folgoc, L. Le, Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-net: learning where to look for the pancreas. http://arxiv.org/abs/1804.03999.
Pan, X., Yang, F., Gao, L., Chen, Z., Zhang, B., Fan, H., Ren, J., 2019. Building extraction from high-resolution aerial imagery using a generative adversarial network with spatial and channel attention mechanisms. Rem. Sens. 11 (8). https://doi.org/10.3390/rs11080966.
Pirasteh, S., Rashidi, P., Rastiveis, H., Huang, S., Zhu, Q., Liu, G., Li, Y., Li, J., Seydipour, E., 2019. Developing an algorithm for buildings extraction and determining changes from airborne LiDAR, and comparing with R-CNN method from drone images. Rem. Sens. 11 (11). https://doi.org/10.3390/rs11111272.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. Lect. Notes Comput. Sci. 9351. https://doi.org/10.1007/978-3-319-24574-4_28.
Salehi, S.S.M., Erdogmus, D., Gholipour, A., 2017. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10541 LNCS. https://doi.org/10.1007/978-3-319-67389-9_44.
Sheng-Hua, Z., Jian-Jun, H., Wei-Xin, X., 2008. A new method of building detection from a single aerial photograph. In: International Conference on Signal Processing Proceedings. ICSP. https://doi.org/10.1109/ICOSP.2008.4697350.
Sirmaçek, B., Ünsalan, C., 2008. Building detection from aerial images using invariant color features and shadow information. In: 2008 23rd International Symposium on Computer and Information Sciences. ISCIS 2008. https://doi.org/10.1109/ISCIS.2008.4717854.
The World Bank, 2023. Urban Development. https://www.worldbank.org/en/topic/urbandevelopment/overview.
Uzar, M., Öztürk, Ş., Bayrak, O.C., Arda, T., Öcalan, N.T., 2021. Performance analysis of YOLO versions for automatic vehicle detection from UAV images. Advanced Remote Sensing 1 (1).
Wang, H., Miao, F., 2022. Building extraction from remote sensing images using deep residual U-Net. European Journal of Remote Sensing 55 (1). https://doi.org/10.1080/22797254.2021.2018944.
Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: convolutional block attention module. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11211 LNCS. https://doi.org/10.1007/978-3-030-01234-2_1.
Xu, L., Liu, Y., Yang, P., Chen, H., Zhang, H., Wang, D., Zhang, X., 2021. HA U-net: improved model for building extraction from high resolution remote sensing imagery. IEEE Access 9. https://doi.org/10.1109/ACCESS.2021.3097630.
Xu, Y., Wu, L., Xie, Z., Chen, Z., 2018. Building extraction in very high resolution remote sensing imagery using deep learning and guided filters. Rem. Sens. 10 (1). https://doi.org/10.3390/rs10010144.
Yong, L.I., Huayi, W.U., 2007. Adaptive building edge detection by combining lidar data and aerial images. Int. Arch. Photogram. Rem. Sens. Spatial Inf. Sci. XXXVII (Part B1).
Yu, M., Chen, X., Zhang, W., Liu, Y., 2022. AGs-unet: building extraction model for high resolution remote sensing images based on attention gates U network. Sensors 22 (8). https://doi.org/10.3390/s22082932.
Zhang, X., He, L., Qin, K., Dang, Q., Si, H., Tang, X., Jiao, L., 2022. SMD-net: siamese multi-scale difference-enhancement network for change detection in remote sensing. Rem. Sens. 14 (7). https://doi.org/10.3390/rs14071580.
Zhang, Y., 1999. Optimisation of building detection in satellite images by combining multispectral classification and texture filtering. ISPRS J. Photogrammetry Remote Sens. 54 (1). https://doi.org/10.1016/S0924-2716(98)00027-6.
Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual U-net. Geosci. Rem. Sens. Lett. IEEE 15 (5). https://doi.org/10.1109/LGRS.2018.2802944.
Zhu, Y., Liang, Z., Yan, J., Chen, G., Wang, X., 2021. E-D-Net: automatic building extraction from high-resolution aerial images with boundary information. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 14. https://doi.org/10.1109/JSTARS.2021.3073994.