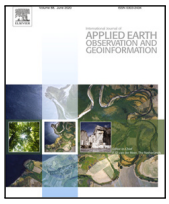
International Journal of Applied Earth Observation and Geoinformation 139 (2025) 104486

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



BANet: A bilateral attention network for extracting changed buildings between remote sensing imagery and cadastral maps

Qingyu Li [a], Lichao Mou [a], Yilei Shi [b], Xiao Xiang Zhu [a,c,*]

[a] Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany
[b] School of Engineering and Design, Technical University of Munich, Munich, 80333, Germany
[c] Munich Center for Machine Learning, Munich, 80333, Germany

ARTICLE INFO

Keywords:
Remote sensing imagery
Building change detection
Cadastral map
Deep learning
Bilateral attention

ABSTRACT

Up-to-date cadastral maps are vital to local governments in administrating real estate in cities. With its growing availability, remote sensing imagery is the cost-effective data for updating semantic contents on cadastral maps. In this study, we address the problem of updating buildings on cadastral maps, as city renewal is mainly characterized by new construction and demolition. While previous works focus on extracting all buildings from remote sensing images, we argue that these methods not only disregard preliminary information on cadastral maps but also fail to preserve building priors in unchanged areas on cadastral maps. Therefore, we focus on the task of extracting changed buildings (i.e., newly built and demolished buildings) from remote sensing images and cadastral maps. To address this task, we create an image-map building change detection (IMBCD) dataset, formed by around 27K pairs of remote sensing images and maps and their corresponding changed buildings in six distinct geographical areas across the globe. Accordingly, we propose a Bilateral Attention Network (BANet), introducing a novel attention mechanism: changed-first (CF) attention and non-changed-first (NCF) attention. This bilateral attention mechanism helps to refine the uncertain areas between changed and non-changed regions. Extensive experiments on our IMBCD dataset showcase the superior performance of BANet. Specifically, our BANet outperforms state-of-the-art models with F1 scores of 90.00% and 63.00% for the IMBCD-WHU and IMBCD-Inria datasets. This confirms that the leverage of bilateral attention blocks (BAB) can boost performance.

1. Introduction

Cadastral maps assist local authorities in land management, as they provide a comprehensive recording of the real estate (Henssen, 1995). Once changes occur in the physical world, the corresponding contents on cadastral maps must be updated regularly to ensure their validity (Revaud et al., 2019). By tracking changes on cadastral maps, urban planners can make informed decisions about infrastructure development, zoning regulations, and resource allocation. In the context of disaster management, up-to-date cadastral maps can assist in quickly evaluating the extent of damage to structures, prioritizing rescue and relief efforts, and planning reconstruction activities. Cities are dynamic, mostly revealed in the construction or demolition of buildings (Kraff et al., 2020). Hence, in this study, our focus is on the fundamental task of updating buildings on cadastral maps. Benefiting from the rising availability of remote sensing imagery, the cost of updating cadastral maps can be lowered by using computer vision to retrieve up-to-date information. Traditionally, this task relies on the manual interpretation of remote sensing images, but it is tedious and expensive. In this study, our goal is to make use of both remote sensing images and cadastral maps for updating buildings on cadastral maps. For this, we believe that cadastral maps can be leveraged as input for network learning, as they contain semantic information about buildings before the change.

This novel perspective on updating cadastral maps leads to the task addressed in this paper: extracting changed buildings (*i.e.*., newly constructed and demolished buildings) between remote sensing images and cadastral maps. Specifically, given an image-map pair, a model is expected to segment changed buildings. We create an image-map building change detection (IMBCD) dataset, a new dataset that draws the attention of researchers to this interesting task. (see Fig. 1). IMBCD contains around 27 K pairs of remote sensing images and cadastral maps which cover six distinct geographical areas across the globe. We further provide annotated masks of changed buildings. IMBCD allows for leveraging information on cadastral maps to segment changed buildings.

* Corresponding author.
E-mail addresses: qingyu.li@tum.de (Q. Li), lichao.mou@tum.de (L. Mou), yilei.shi@tum.de (Y. Shi), xiaoxiang.zhu@tum.de (X.X. Zhu).

https://doi.org/10.1016/j.jag.2025.104486
Received 7 January 2024; Received in revised form 3 March 2025; Accepted 15 March 2025

Fig. 1. Example data in the IMBCD dataset: (left) cadastral map, (middle) remote sensing image, and (right) mask for changed buildings. The first row refers to the newly constructed buildings, while the second row denotes the demolished buildings.

To lay a solid foundation for this task, we propose a Bilateral Attention Network (BANet), an attention-empowered encoder–decoder deep network relying on two components: changed-first (CF) attention and non-changed-first (NCF) attention. The CF attention is about the refinement of changed regions (CR), while the NCF attention refers to the recovery of useful information in non-changed regions (NCR). We hypothesize that the bilateral exploration of the changed and no-changed cues is the most informative for our task, as it enhances the information in uncertain regions. Specifically, the original features are first multiplied with CF and NCF attention maps, respectively. By fusing CF and NCF attention-enhanced features, not only the details of changed buildings can be refined, but also the potential cues in NCR can be leveraged. The bilateral attention mechanism in BANet enables extracting changed buildings with high accuracy, which can support timely updates to cadastral maps and facilitate disaster response efforts, and improve property record management.

In summary, our primary contributions can be outlined as follows:

- We create a new dataset, IMBCD, which contains around 27 K pairs of remote sensing images and cadastral maps, as well as mask annotations of changed buildings from six distinct geographical areas across different continents.
- We propose a novel network, BANet, based on a bilateral attention mechanism to tackle the challenge, (*i.e.*., extracting changed buildings between remote sensing images and cadastral maps).
- The proposed method obtains satisfactory performance on both two subsets with different spatial resolutions of the IMBCD dataset. Compared with other methods, our approach can significantly improve accuracy metrics. The IMBCD dataset and the code of BANet will be made publicly available in https://github.com/lqycrystal/BANet.

2. Related work

In the existing literature, two main strategies exist for updating buildings on cadastral maps. One solution is to extract buildings from the latest remote sensing data, then changed buildings can be identified by comparing the extracted building masks and the existing cadastral map (Li et al., 2020, 2022b). The other solution is based on change detection algorithms that extract changed buildings from multi-temporal or bi-temporal remote sensing data.

The first solution relies heavily on building extraction methods to produce up-to-date building footprint maps. With the advent of deep learning methods, recent studies have achieved state-of-the-art results in building extraction by utilizing semantic segmentation networks. Their main goal is to tackle challenges related to pixel-level labeling. Specifically, these methods utilize semantic segmentation networks to assign a label to each pixel in the image, classifying it as either "building" or "nonbuilding". These techniques are typically developed to (1) correct inaccurate and irregular building boundaries (Li et al., 2021; Xu et al., 2022; Zorzi et al., 2022), (2) compensate for limited supervision information (Li et al., 2022a; Chen et al., 2023; Zheng et al., 2023), and (3) handle variations in building characteristics (Xu et al., 2023; Dai et al., 2023), such as size. Nevertheless, discrepancies between building extraction results and existing cadastral maps pose challenges in accurately identifying changed areas (Guo et al., 2021). Furthermore, this solution generates building footprints from scratch without considering the prior knowledge from existing cadastral maps. The challenge of integrating building predictions with existing building information remains to be addressed (Guo et al., 2021).

For the second solution, deep networks have been proven to be capable of automatically learning highly discriminating features and can achieve impressive results in building change detection. These methods address pixel-level labeling problems by assigning each pixel in the image a specific label, such as "changed" or "non-changed". To address the variability of buildings, many methods incorporate attention mechanisms that select the most distinguishable features (Zhang et al., 2022; Shu et al., 2022; Li et al., 2022c; Zhou et al., 2023; Wang et al., 2022; Feng et al., 2023). Some studies focus on regularizing boundary information of detected changes (Li et al., 2023), while other studies develop strategies to reduce the need for extensive pixel-level annotations (Wu et al., 2023). Despite these advancements, several critical gaps remain. First, most change detection methods rely on bi-temporal remote sensing imagery, which is often unavailable in practice. Pre-change remote sensing images corresponding to cadastral maps are particularly difficult to acquire (Li et al., 2020, 2022b; Liao et al., 2023). Second, existing methods are primarily designed for homogeneous data (e.g., remote sensing imagery) and do not effectively integrate heterogeneous data sources such as cadastral maps. This limits their applicability to real-world scenarios where cadastral maps provide valuable prior knowledge. Third, benchmark datasets such as WHU Building Change Detection Dataset (Ji et al., 2018), LEVIR CD (Chen and Shi, 2020), and S2Looking (Shen et al., 2021) are limited in scope, as they only provide bi-temporal remote sensing imagery and lack geographic diversity. These datasets are insufficient for evaluating methods that aim to integrate cadastral maps with remote sensing data.

Given these limitations, there is a clear need for a new approach that can effectively integrate cadastral maps with remote sensing imagery

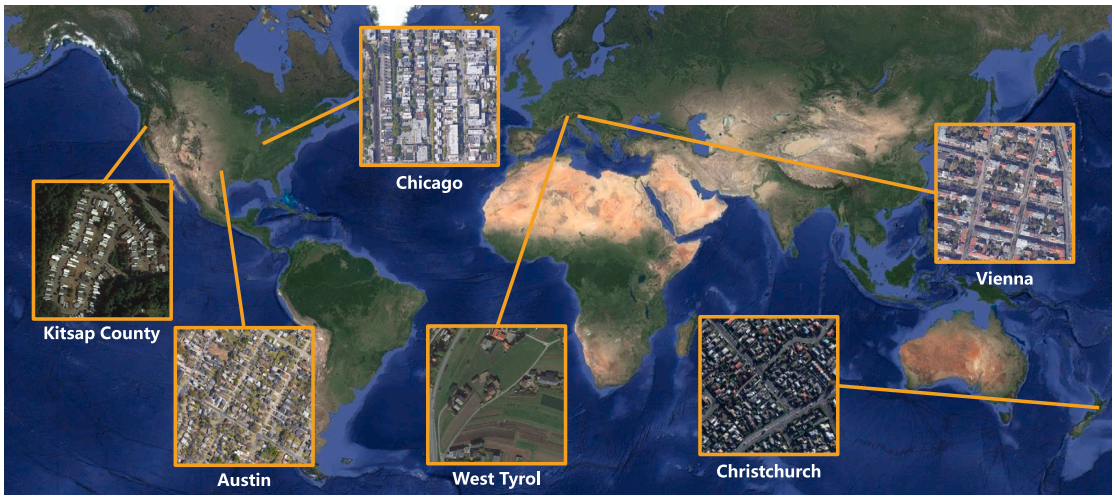


Fig. 2. Geospatial distributions and examples of remote sensing images in our IMBCD dataset. Buildings show great differences among varying geographical regions.

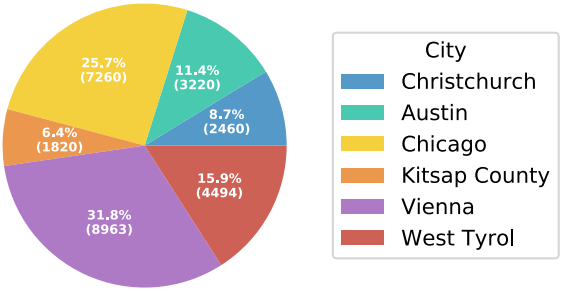


Fig. 3. Distribution of the number of changed buildings in the IMBCD dataset.

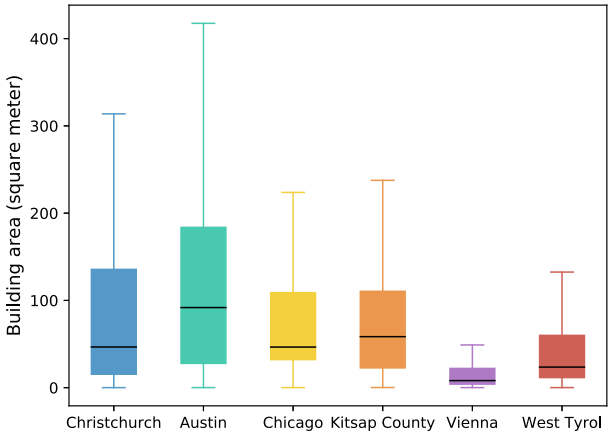


Fig. 4. The range of areas of changed buildings in the IMBCD dataset.

to detect building changes. Only one recent study (Liao et al., 2023) has attempted to address this challenge by providing the SI-BU and WHU-CD datasets. However, these datasets are limited in geographic diversity, as each covers only a single geographical area. This lack of diversity restricts their applicability to large-scale, real-world scenarios.

To address these gaps, we propose a novel approach that leverages both remote sensing imagery and cadastral maps as input data modalities. Unlike existing methods, our approach explicitly models the relationships between these two data sources to improve change detection accuracy. Furthermore, we introduce the IMBCD dataset, which includes six distinct geographical areas across the globe. This geographic diversity ensures that the IMBCD dataset can serve as a standardized benchmark for evaluating methods that integrate cadastral maps and remote sensing imagery.

Table 1
Data split in the IMBCD dataset.

Subset	Geographical area	Train	Val	Test
IMBCD-WHU	Christchurch, New Zealand	1680	240	480
IMBCD-Inria	Austin, the United States	1652	236	472
	Chicago, the United States	5026	718	1436
	Kitsap County, the United States	602	86	172
	Vienna, Austria	5908	844	16 886
	West Tyrol, Austria	3878	554	1108

3. Dataset

3.1. Dataset overview

We create a geo-diverse dataset, the IMBCD dataset, which covers six distinct geographical areas (see Fig. 2) over the whole globe, including both cities and a county. Specifically, Christchurch is a city in New Zealand, Austin and Chicago are cities in the United States, Kitsap County is a county in the United States, Vienna is a city in Austria, and West Tyrol is a region (not a city but often treated as a unit for administrative and statistical purposes) in Austria.

The IMBCD dataset has 26,780 pairs of remote sensing images and maps as well as ground truth annotations. Each image has the size of 256 × 256 pixels. With a ratio of 7:1:2, We split image-map pairs into train, validation, and test data for each geographical area. Based on the absence or presence of changed buildings, the pairs can be grouped into two categories: changed and non-changed. To achieve an optimal balance, we ensure that samples of both types are distributed as evenly as possible within each geographical area. Next, Because of the variation in spatial resolution between aerial images from Christchurch and those from the other five regions, we divide the data into two sub-datasets: IMBCD-WHU and IMBCD-Inria. Table 1 shows the statistics of the IMBCD-WHU and IMBCD-Inria datasets. The number of samples of changed buildings in each geographical area is shown in Fig. 3. Differences also exist in the area of changed buildings among the six geographical regions studied (c.f. Fig. 4). This is because of the geographical variation of these regions.

3.2. Annotation details

The remote sensing image of Christchurch is collected from the WHU Building Change Detection Dataset (Ji et al., 2018) with a size

of 32507 × 15354 pixels at a spatial resolution of 0.2 m/pixel. For the other five geographical areas, the source of remote sensing imagery is the Inria Aerial Image Labeling Dataset (Maggiori et al., 2017), which comprises 180 aerial image tiles. Each geographical area is covered by 36 image tiles, with each tile having a size of 5000 × 5000 pixels at a spatial resolution of 0.3 m/pixel.

We collect cadastral maps from OpenStreetMap (OpenStreetMap contributors, 2017). It is a free and open geographic database. For each remote sensing image in our dataset, we retrieve and download the corresponding building masks on OpenStreetMap. We ensure that each pair of remote sensing images and its corresponding cadastral map cover the same geographical region and are of identical size. Specifically, we apply rigorous geometric correction and alignment procedures to ensure that the remote sensing images and cadastral maps are spatially aligned in the preprocessing stage.

Acquiring pixel-wise mask annotations for changed buildings is time-consuming and expensive. Instead, we design an efficient pipeline to generate annotation data. We first collect building masks that correspond to aerial imagery from the WHU Building Change Detection Dataset (Ji et al., 2018) and the Inria Aerial Image Labeling Dataset (Maggiori et al., 2017). After the collection, building masks are compared with cadastral maps in order to label changed buildings between images and maps. Here, newly constructed buildings are those that appear on a remote sensing image but are not present on the corresponding cadastral map. Demolished buildings are present on a cadastral map but not on the corresponding image. Further, we crop all data to patches with a size of 256 × 256 pixels. We also notice that some annotations are incorrect. For example, some changed buildings have not been labeled or some annotated changed buildings are incorrect. To avoid such noise, we manually check each image-map pair and its mask annotation, and the noisy data are discarded.

4. Method

4.1. Problem formulation

Given a remote sensing image $\boldsymbol{x}_{img} \in \mathbb{R}^{H\times W\times C}$ and a cadastral map $\boldsymbol{x}_{map} \in \mathbb{R}^{H\times W\times C}$ covering the same geographic region, we are interested in segmenting changed buildings between $\boldsymbol{x}_{img}$ and $\boldsymbol{x}_{map}$. H and W are height and width, and C is the number of channels. In this study, C is 3. Note that $\boldsymbol{x}_{map}$ is a binary map where 1 represents buildings and 0 denotes background. Moreover, we convert this single-channel binary cadastral map into a three-channel map ($H\times W\times 3$) to facilitate the joint processing of remote sensing images and cadastral maps. By replicating the single-channel cadastral map into three channels, we ensure the network can process both data types simultaneously and effectively. On the one hand, matching the channel count can prevent the network from disproportionately prioritizing the image (with richer initial channels) over the cadastral map. On the other hand, each replicated channel of the cadastral map could implicitly capture different contextual aspects (*e.g.*., edges, regions, or spatial hierarchies) through convolutional filters, enabling more effective fusion in deeper layers. The resulting segmentation mask is $\boldsymbol{y} \in \mathbb{R}^{H\times W}$, which is defined as binary labeling of all pixels for indicating whether buildings have changed. 1 denotes change, while 0 represents no change.

4.2. Network architecture

We can observe that the distribution of CR and NCR is quite different. Focusing more on changed buildings contributes to predicting high-confidence objects, which may lead to incomplete results. More complete objects can be identified if we pay more attention to NCR. However, unexpected noise may be introduced. Thus, we propose to jointly learn changed and non-changed cues. By doing so, not only can complete masks for changed buildings be predicted but the background noise can also be suppressed.

Fig. 5 illustrates the proposed network, which has three main modules: encoder, bilateral attention blocks (BAB), and decoder. Input remote sensing image $\boldsymbol{x}_{img}$ and cadastral map $\boldsymbol{x}_{map}$ are first concatenated and fed into the encoder to learn multi-level features $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_5\}$. The coarse prediction $\boldsymbol{P}_5$ for changed buildings is derived from $\boldsymbol{F}_5$, and $\{\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3, \boldsymbol{F}_4\}$ are prepared for BAB in order to further refine multi-level prediction maps. This is achieved by distinguishing uncertain areas as changed or non-changed through a top-down approach. The initial prediction obtained from high-level features is coarse but contains rich semantic information that is useful for predicting the initial position of the CR and NCR. Therefore, BAB can utilize a prediction map at a higher level and features at the current level to obtain the current-level prediction map and bilaterally enhanced features. Following the integration of multi-level bilaterally enhanced features into the decoder, segmentation mask $\boldsymbol{y}$ can be generated.

As the network depth increases, high-level features (*e.g.*., $\boldsymbol{F}_5$) become more effective in capturing global context but may lose details of changed buildings. When directly upsampling high-level predictions, such as masks for changed buildings (*e.g.*., $\boldsymbol{P}_5$) will be blurred. Thus, we devise BAB to enhance the difference between CR and NCR. In our BAB, the higher-level prediction functions as a CF attention map. In contrast, the reversed prediction acts as an NCF attention map, facilitating the integration of bilateral attention on both CR and NCR. As can be seen in Fig. 6, after the sigmoid activation is applied, the upsampled predictions from the high level will be used as CF maps $\{\boldsymbol{S}^A\}_{i=1}^5$. NCF maps $\{\boldsymbol{S}^B\}_{i=1}^5$ are produced by subtracting CF maps from $\boldsymbol{E}$ that is a matrix with all elements as 1.

$$\begin{cases} \boldsymbol{S}_i^A = \boldsymbol{\delta}(\boldsymbol{U}(\boldsymbol{P}_{i+1})), \\ \boldsymbol{S}_i^B = \boldsymbol{E} - \boldsymbol{S}_i^A, \end{cases} \quad i \in \{1,2,3,4,5\}, \tag{1}$$

where $\boldsymbol{U}(\cdot)$ is an upsampling operation and $\boldsymbol{\delta}(\cdot)$ is a sigmoid activation function. Afterward, we apply CF and NCF attention maps to multi-level features, respectively, and further acquire enhanced features and refined predictions.

$$\begin{cases} \boldsymbol{F}_i' = \boldsymbol{\gamma}([\boldsymbol{\beta}(\boldsymbol{F}_i \odot \boldsymbol{S}_i^A), \boldsymbol{\beta}(\boldsymbol{F}_i \odot \boldsymbol{S}_i^B)]), \\ \boldsymbol{P}_i = \boldsymbol{\alpha}(\boldsymbol{F}_i'), \end{cases} \quad i \in \{1,2,3,4\}, \tag{2}$$

where $\odot$ is an element-wise multiplication operation, $[\cdot,\cdot]$ denotes concatenation. $\boldsymbol{\beta}$ comprises a convolutional layer with the kernel size of 3 × 3 and a ReLU layer. $\boldsymbol{\gamma}$ is a convolution with the kernel size of 1 × 1 that reduces the number of feature channels. The goal of $\boldsymbol{\alpha}$ is to generate a probability map by using two convolutional layers and a ReLU layer. The first convolutional layer is with a 3 × 3 kernel, while the second convolutional layer outputs a single channel map via a 1 × 1 kernel.

The comparison between original features and enhanced features from BAB (see Fig. 6) provides a better understanding of the bilateral attention mechanism. The original features are divided into two branches, with one branch being multiplied by the CF attention map and the other by the NCF attention map. It is observable that the CF branch redirects attention towards the predicted CR from its higher level, allowing for the exploration of cues related to changed buildings. Complementarily, the NCF branch focuses on the NCR to look for possible changed buildings within them. With joint learning of CF and NCF features, the bilaterally enhanced features show sharp contrast on the edges of changed buildings. This suggests BAB cannot only suppress the NCR but also enhance the discriminative information of changed buildings.

4.3. Loss functions

For the supervised learning of BANet, an overall objective function $\mathcal{L}$ is devised:

$$\mathcal{L} = \mathcal{L}_{seg1} + \lambda \mathcal{L}_{seg2}\,, \tag{3}$$

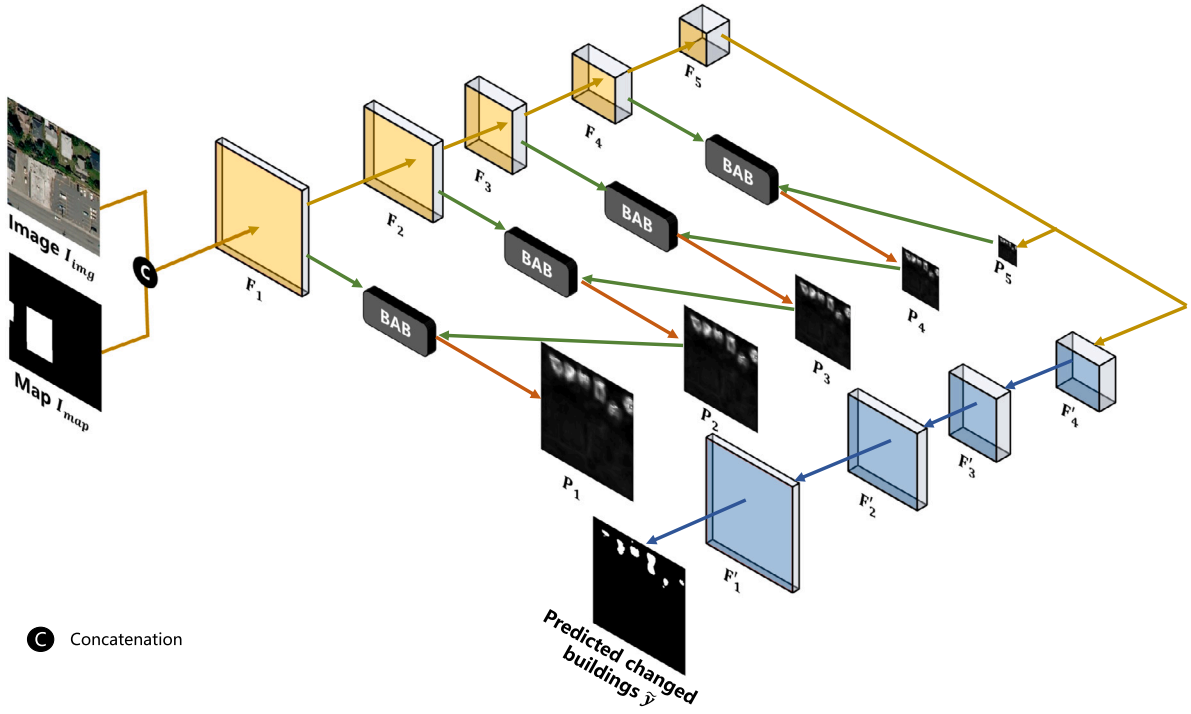


Fig. 5. Flowchart of our BANet. BANet contains three modules: encoder, bilateral attention blocks (BAB), and decoder. Specifically, a remote sensing image x_{img} and a cadastral map $\boldsymbol{x}_{map}$ are first concatenated and fed into the encoder to learn multi-level features $\{\boldsymbol{F}\}_{i=1}^{5}$. The highest level features $\boldsymbol{F}_5$ is used to predict a coarse map $\boldsymbol{P}_5$ for changed buildings. The higher-level prediction map $\boldsymbol{P}_{i+1}$ and current-level features $\boldsymbol{F}_i$ are fed into BAB where the outputs are bilaterally enhanced features $\boldsymbol{F}'_i$ and new prediction map $\boldsymbol{P}_i$. By leveraging BAB in a top-down manner, we are able to obtain the final segmentation mask $\tilde{\boldsymbol{y}}$.

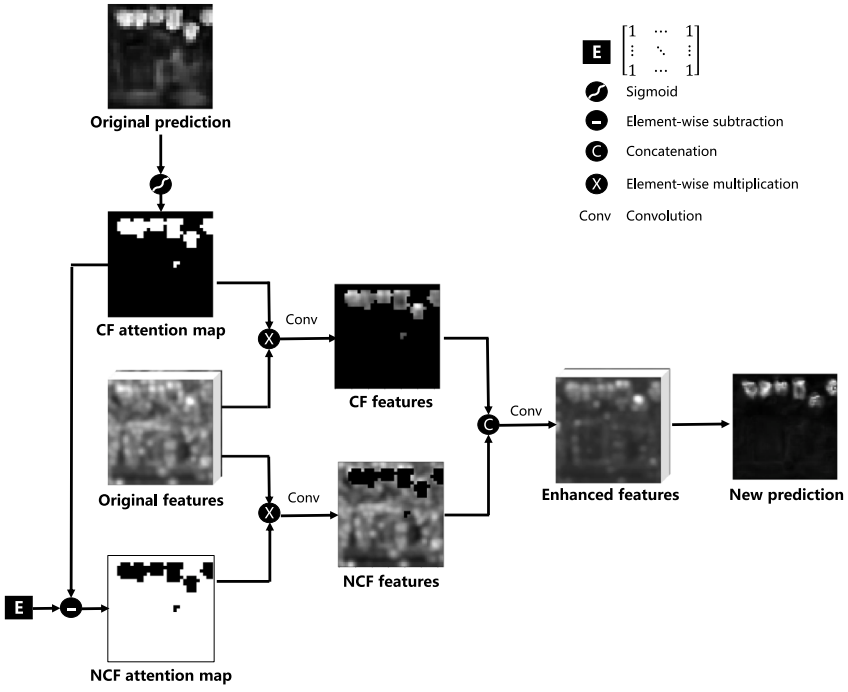


Fig. 6. Flowchart of the bilateral attention block (BAB). 'CF' means changed-first, while 'NCF' denotes non-changed-first.

Table 2
Numerical results of various methods on the IMBCD dataset.

Method	IMBCD-WHU		IMBCD-Inria	
	IoU	F1 score	IoU	F1 score
ChangeNet (Varghese et al., 2018)	46.14	63.15	20.34	33.81
DR-TANet (Chen et al., 2021)	20.55	34.09	5.36	10.17
ADHR-CDNet (Zhang et al., 2022)	68.77	81.49	27.94	43.68
SUNet (Shao et al., 2021)	71.22	83.19	32.27	48.79
USSFC-Net (Lei et al., 2023)	60.98	75.76	42.32	59.25
BCE-Net (Liao et al., 2023)	71.77	83.56	29.91	46.05
CMNeXt (Zhang et al., 2023)	66.95	80.21	32.10	48.60
PGDENet (Zhou et al., 2022)	76.19	86.49	40.15	57.30
DeepLab v3+ (Chen et al., 2018)	59.85	74.89	22.31	36.48
HRNet (Yuan et al., 2020)	70.70	82.84	32.12	48.62
SegFormer (Xie et al., 2021)	58.21	73.59	25.32	40.41
BANet	**81.83**	**90.00**	**45.99**	**63.00**

where λ is a hyperparameter controlling the weight of each loss term. $\mathcal{L}_{seg1}$ is cross-entropy loss function between the predicted masks $\widetilde{\boldsymbol{y}}$ and ground truth $\boldsymbol{y}$. $\mathcal{L}_{seg2}$ is employed to minimize the cross-entropy loss between the multi-level predictions $\{\boldsymbol{P}\}_{i=1}^{5}$ and ground truth $\boldsymbol{y}$. Note that we implemented the operation of bilinear upsampling on $\{\boldsymbol{P}\}_{i=1}^{5}$ to retain the same size as $\widetilde{\boldsymbol{y}}$ before calculating $\mathcal{L}_{seg2}$.

5. Experiments

5.1. Implementation details

The encoder in BANet consists of four blocks, each with one max pooling layer and two convolutional layers. For the decoder, we make use of four blocks with two convolutional layers and one upsampling layer in each block. The parameter λ in Eq. (3) is set as 0.01 empirically. At training time, stochastic gradient descent (SGD) is the optimizer with a learning rate of 0.0001. The model is trained for 200 epochs and the training batch size is 4. All experiments are performed using PyTorch on an NVIDIA Tesla P100 GPU equipped with 16 GB of memory. The performance of models is assessed according to two classical segmentation metrics: F1 score and intersection over union (IoU). Note that these metrics are calculated based on building pixels rather than building objects.

5.2. Comparison with state-of-the-art methods

IMBCD contains images and cadastral maps that belong to different modalities. This prompts the inquiry of how to more effectively harness multimodal data to improve segmentation performance. To evaluate the effectiveness of BANet, we select two types of approaches as competitors: change detection methods and semantic segmentation models. For the former, we apply the following change detection algorithms to our task: ChangeNet (Varghese et al., 2018), DR-TANet (Chen et al., 2021), ADHR-CDNet (Zhang et al., 2022), SUNet (Shao et al., 2021), USSFC-Net (Lei et al., 2023), and BCE-Net (Liao et al., 2023). For the latter, we deem this task a semantic segmentation problem by feeding paired remote sensing images and cadastral maps into a semantic segmentation network. Two types of semantic segmentation networks are selected in this study. One type is algorithms devised for multimodal semantic segmentation, including CMNeXt (Zhang et al., 2023) and PGDENet (Zhou et al., 2022). The other type is concatenating images and maps together and feeding them into a single segmentation network. Previous research (Zhao et al., 2020; Jiang et al., 2022) has demonstrated that this concatenation is effective in multimodal learning. More specifically, the following networks are considered: DeepLab v3+ (Chen et al., 2018), HRNet (Yuan et al., 2020), and SegFormer (Xie et al., 2021).

Table 2 reports numerical results, and BANet outperforms all competitors significantly. It can be seen from the results that BANet obtains improvements of at least 5.64% and 3.67% in IoU on IMBCD-WHU and IMBCD-Inria datasets, respectively. Regarding selected change detection methods, ChangeNet and DR-TANet have relatively lower accuracy. Since they use two separate encoders for extracting features from the image and map separately, this structure seems less effective for multimodal change detection tasks. As cadastral maps are binary, extracting useful information for later fusion with image features based on this structure is very difficult. However, with a specially designed feature fusion strategy, the results can be greatly improved with this structure, (*e.g.*., multiscale spatial feature attention module in ADHR-CDNet and spatial–spectral feature cooperation module in USSFC-Net). The benefit of a specially designed fusion strategy is also demonstrated by a multimodal semantic segmentation network with two separate encoders, (*i.e.*., the progressive complementary fusion module in PGDENet). Like SUNet, DeepLab v3+, HRNet, and SegFormer, our BANet concatenates images and maps before feeding them into the encoder. This allows the model to learn joint representations that capture each modality's complementary and supplementary information. Despite the same data input strategy, the differences in network architectures (*e.g.*., transformer and convolution neural network (CNN)) also lead to variances in different methods. For instance, the transformer-based architecture (SegFormer and CMNeXt) performs relatively worse. Furthermore, the buildings can be easily confused with other land cover classes (*e.g.*., paved roads and bare land). BAB make BANet a more robust model for such a challenging task, and the bilateral attention mechanism can collaboratively leverage the cues from CR and NCR to extract changed buildings. We also notice that metrics are lower on the IMBCD-Inria dataset which is more complex in terms of geographical variation (such as the color, shape, and size of buildings) than the IMBCD-WHU dataset.

Example results using various approaches are shown in Figs. 7 and 8. BANet performs better in extracting changed buildings. Although some predicted masks are not perfect (*e.g.*., some buildings have blob-like shapes), it does not prevent our method from predicting the right ones.

The number of parameters and FLOPs for all methods are shown in Table 3. The higher FLOPs of our method are a direct result of its design, which prioritizes feature richness and fusion effectiveness over computational efficiency. The architecture is optimized to extract and integrate multi-modal features (remote sensing images and cadastral maps) effectively, leading to significant gains in segmentation accuracy. This trade-off is often necessary for complex tasks like semantic segmentation, where accuracy is critical. Moreover, the parallelizable nature of modern GPUs ensures that our method remains feasible for large-scale applications. Additionally, techniques such as model pruning, quantization, or distillation can be applied to further reduce computational costs without significantly compromising accuracy. Despite the higher FLOPs, our method has a relatively small number of parameters (fifth-smallest among 12 networks). This indicates that the model is memory-efficient and less prone to overfitting, making it feasible for deployment in resource-constrained environments.

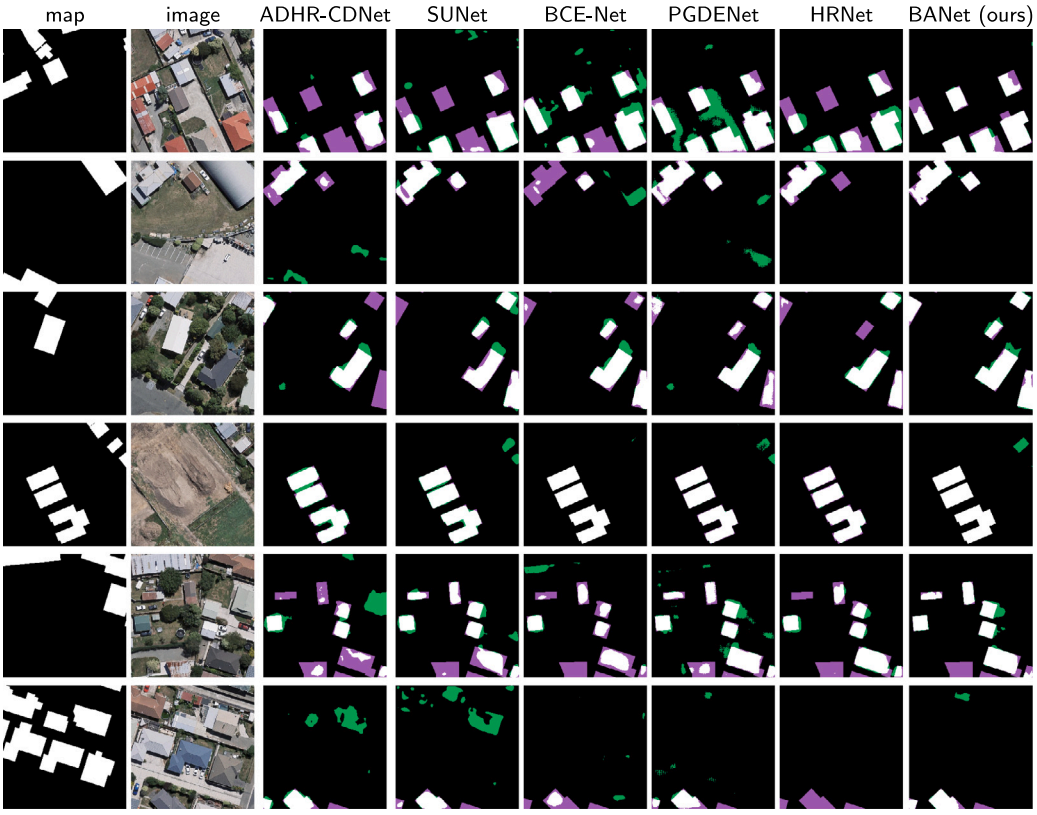


Fig. 7. Prediction results on the IMBCD-WHU dataset. Pixel-based true positives, false positives, and false negatives are marked in white, green, and purple, respectively.

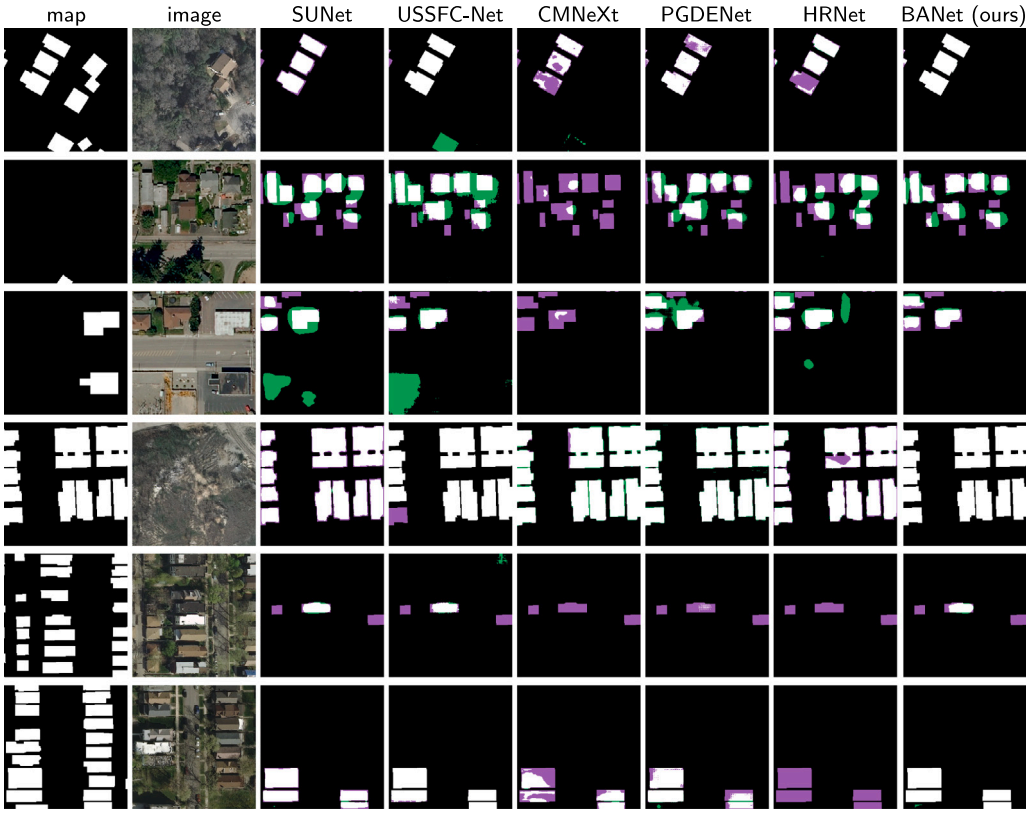


Fig. 8. Prediction results on the IMBCD-Inria dataset. Pixel-based true positives, false positives, and false negatives are marked in white, green, and purple, respectively.

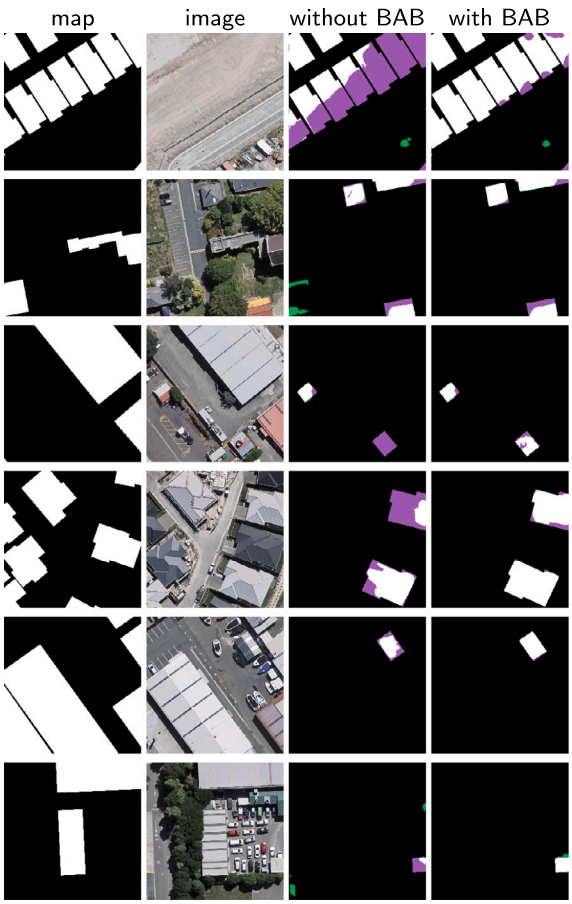


Fig. 9. Prediction results on the IMBCD-WHU dataset. Pixel-based true positives, false positives, and false negatives are marked in white, green, and purple, respectively. BAB represents bilateral attention blocks.

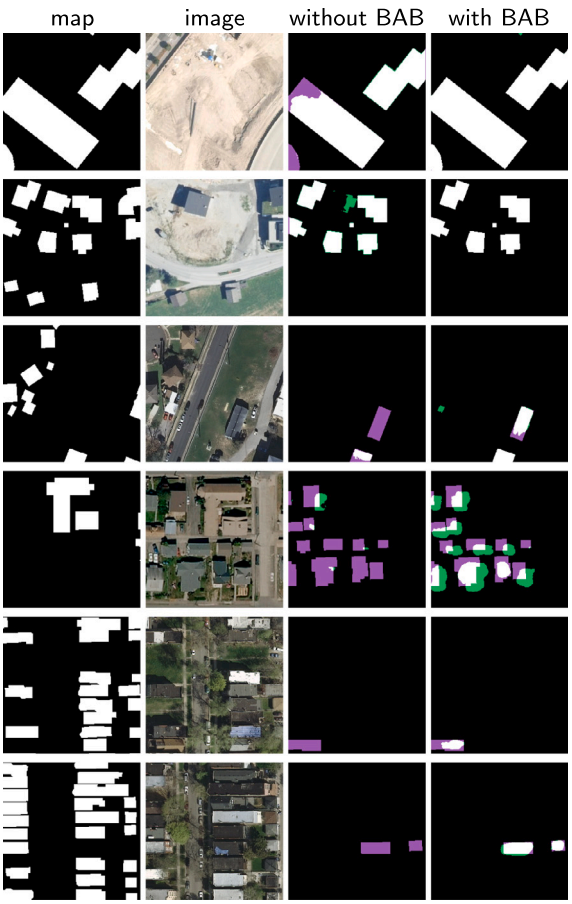


Fig. 10. Prediction results on the IMBCD-Inria dataset. Pixel-based true positives, false positives, and false negatives are marked in white, green, and purple, respectively. BAB represents bilateral attention blocks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.3. Ablation study

To further confirm the efficacy of the bilateral attention blocks implemented in BANet, we conduct an ablation study on both IMBCD-WHU and IMBCD-Inria datasets. We obtain the results of BANet where bilateral attention networks are removed in the objective function. That is to say, $\mathcal{L}_{seg1}$ is the only loss term for learning changed buildings. From Table 4, the leverage of BAB boosts the performance on both datasets. More specifically, it brings 1.19% gain in IoU on IMBCD-WHU dataset and 12.89% boost in IoU on IMBCD-Inria dataset. Figs. 9 and 10 present the corresponding visual results. The segmentation masks obtained from the method with BAB closely align with the ground reference. Hence, we can infer that BANet contributes to the addressed task.

In what follows, we discuss in detail the underlying factors contributing to our model's success for this task. The most important factor is that our model leverages bilateral attention mechanisms, which include complimentary CF and NCF attention mechanisms. CF attention mechanism ensures that the model captures and emphasizes crucial alterations, enhancing the detection of meaningful changes. NCF attention mechanism provides a stable reference, reducing the risk of false positives and helping the model to differentiate subtle changes from noise. On the one hand, the bilateral attention mechanism enables a richer and more nuanced feature representation by leveraging both CF and NCF attention. On the other hand, the complementary CF and NCF attention mechanisms provide a balanced and holistic analysis of the scene, enhancing the model's ability to discern true changes from noise and irrelevant variations. Second, our model concatenates maps and images before feeding them into the encoder, allowing the model to attend to and integrate information from different modalities simultaneously. This also contributes to a more comprehensive understanding of the scene. Third, BANet distinguishes the uncertain regions as changed or non-changed in a top-down manner to further refine upsampled change maps. The model builds a more robust feature representation by preserving low-level details and capturing high-level semantic information. Furthermore, the progressive integration of features at different levels is beneficial in capturing intricate details and subtle changes.

5.4. Hyperparameter analysis

As described in Eq. (3), a hyperparameter λ is exploited to regulate the weighting of each loss term. We study the impact of λ on quantitative results on both IMBCD-WHU and IMBCD-Inria datasets. The results with different values of λ are illustrated in Table 5. The changes in λ can affect the model performance. We select λ as 0.01 for IMBCD-WHU and IMBCD-Inria datasets, respectively. This is because their corresponding statistical metrics are optimal.

In our approach, the depth of the convolutional layers representing the number of downsampling operations applied in the encoder is also a critical hyperparameter. To investigate the impact of depth, we set it as three different numbers, (*i.e.*., 4, 5, and 6), to explore its effect on final results. The experiments are carried out on IMBCD-WHU and

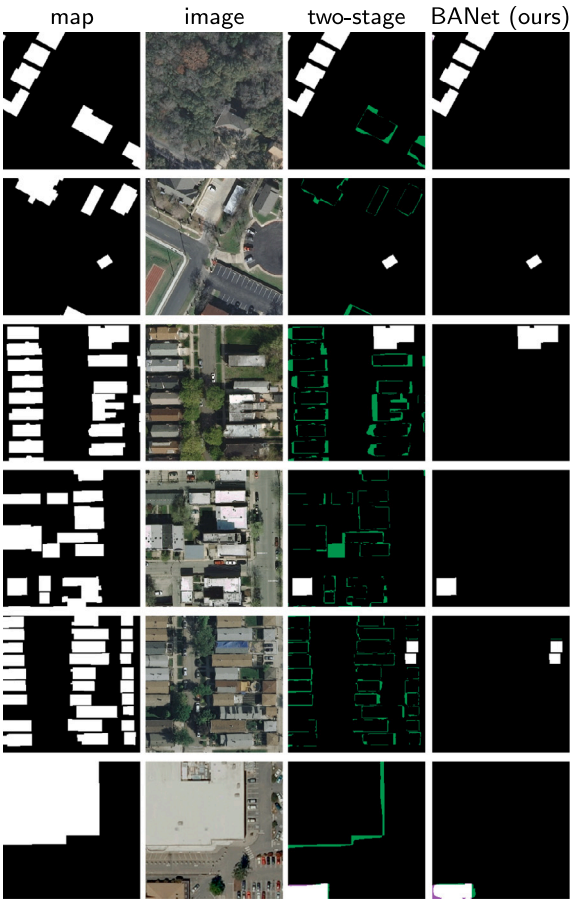


Fig. 11. Prediction results on the IMBCD-Inria dataset. Pixel-based true positives, false positives, and false negatives are marked in white, green, and purple, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of network structures in different methods.

Method	Parameters (M)	FLOPs (G)
ChangeNet (Varghese et al., 2018)	51.31	10.87
DR-TANet (Chen et al., 2021)	33.39	7.07
ADHR-CDNet (Zhang et al., 2022)	12.90	119.47
SUNet (Shao et al., 2021)	15.56	39.69
USSFC-Net (Lei et al., 2023)	1.52	3.23
BCE-Net (Liao et al., 2023)	31.67	11.28
CMNeXt (Zhang et al., 2023)	58.68	14.56
PGDENet (Zhou et al., 2022)	107.39	34.80
DeepLab v3+ (Chen et al., 2018)	54.70	20.70
HRNet (Yuan et al., 2020)	29.53	22.72
SegFormer (Xie et al., 2021)	7.72	3.34
BANet	20.79	55.03

Table 4
Ablation study results on our IMBCD dataset. BAB represents bilateral attention blocks.

Method	IMBCD-WHU		IMBCD-Inria	
	IoU	F1 score	IoU	F1 score
Without BAB	80.64	89.28	33.10	49.74
With BAB	**81.83**	**90.00**	**45.99**	**63.00**

IMBCD-Inria datasets, respectively. As can be seen in accuracy metrics (c.f. Table 6) on both datasets, the best result is obtained when depth is 5.

Table 5
Hyperparameter analysis for λ on the IMBCD dataset.

λ	IMBCD-WHU		IMBCD-Inria	
	IoU	F1 score	IoU	F1 score
0.001	79.86	88.80	41.95	59.11
0.01	**81.83**	**90.00**	**45.99**	**63.00**
0.1	81.67	89.91	44.71	61.79
1	72.04	83.75	35.29	52.28

Table 6
Hyperparameter analysis for network depth on the IMBCD dataset.

Depth	IMBCD-WHU		IMBCD-Inria	
	IoU	F1 score	IoU	F1 score
4	81.52	89.82	39.02	56.14
5	**81.83**	**90.00**	**45.99**	**63.00**
6	78.87	88.18	42.22	59.37

Table 7
Comparison with two-stage method on the IMBCD-Inria dataset.

Method	IoU	F1 score
Two-stage	29.38	45.41
BANet	**45.99**	**63.00**

5.5. Comparison with two-stage method

In our approach, the image and map are taken as network input, and the output is a change mask. In contrast to our direct change detection method, the existing two-stage method first identifies buildings and then discerns changes through comparison with cadastral maps. To validate the effectiveness of direct change detection methods for this task, we perform a comparative study with another competitor, (*i.e.*., a two-stage method). That is to say, a semantic segmentation network (Li et al., 2021) pre-trained on the Inria dataset is first utilized to extract buildings from remote sensing images. Then, changed buildings can be identified by comparing the extracted building masks and the existing cadastral map. The comparative analysis is carried out on the IMBCD-Inria dataset. Numerical results are shown in Table 7, while visual results are illustrated in Fig. 11. In particular, the IoU achieved by our approach is increased by more than 16.61% compared to the two-stage method. Besides, it can be seen that our method has fewer false alarms (c.f. Fig. 11), while the two-stage method fails to preserve building priors in unchanged areas on cadastral maps. This suggests that BANet can better leverage information from existing cadastral maps to improve results.

5.6. Semantic building change detection

In addition to binary labels (*i.e.*., changed and non-changed) of change masks, we also include more informative labels for the IMBCD dataset. Specifically, we provide details of the types of change (*i.e.*., newly built and demolished buildings). Three classes exist in this detailed change mask, denoted as background, newly built buildings, and removed buildings, corresponding to values 0, 1, and 2, respectively. Fig. 12 provides visual examples of these labels, where the colors black, red, and blue are employed to represent the values mentioned earlier. The ratios of newly built and demolished buildings in each geographical area are shown in Fig. 13. Because of geographical differences, the ratios vary across different geographical regions. For example, in Kitsap County, the ratio of newly constructed buildings is higher than that of demolished buildings, while in the other five geographical areas, vice versa. Fig. 14 shows the result of applying BANet at a larger scale.

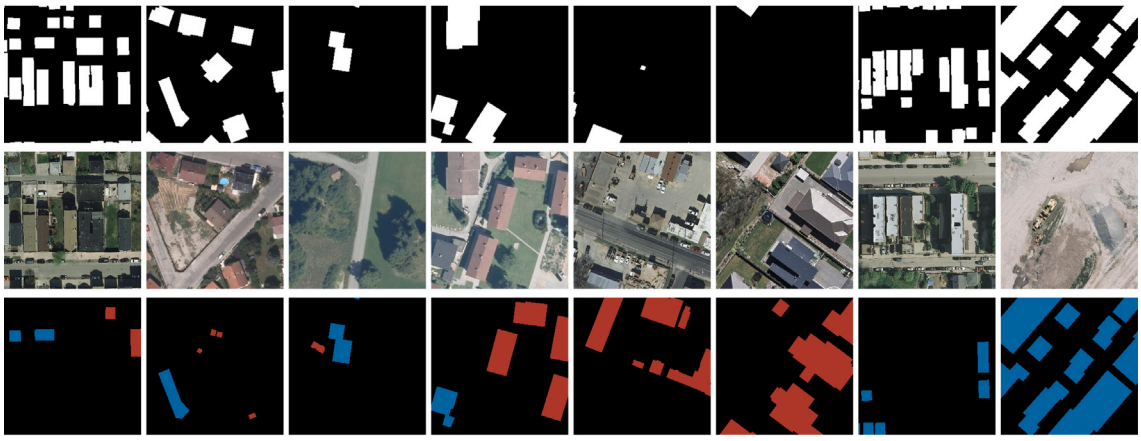


Fig. 12. Rows from top to bottom are: map, remote sensing image, ground reference mask of newly built (red) and demolished (blue) buildings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

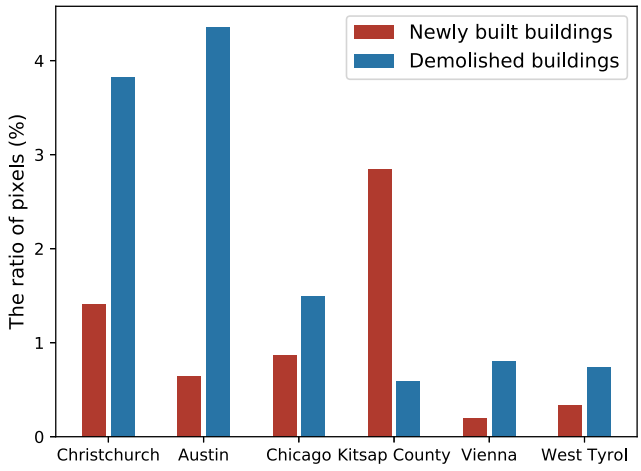


Fig. 13. Distribution of types of changed buildings in the IMBCD dataset.

6. Conclusion

Buildings are one of the most important terrestrial objects in cities and are continuously being constructed or demolished due to urban expansion and city renewal. Remote sensing imagery offers great potential for updating buildings on cadastral maps. However, the traditional visual inspection is tedious and slow when there are an immense number of changed buildings that have to be updated on cadastral maps. Therefore, automatically monitoring changed buildings is crucial. In this study, we address the task of extracting changed buildings between remote sensing images and maps. We create a public dataset called IMBCD to investigate this task. Meanwhile, we propose a new method BANet which is based on the idea of bilateral attention mechanisms to segment changed buildings. Experimental results on the IMBCD dataset clearly illustrate the efficacy of BANet. Specifically, BANet outperforms state-of-the-art methods by at least 5.64% and 3.67% in IoU on IMBCD-WHU and IMBCD-Inria datasets, respectively. Our work provides a practical strategy that can take full advantage of cadastral maps and remote sensing imagery for the update of cadastral maps. While BANet demonstrates strong performance, its computational cost may limit real-time applications. Additionally, performance may vary in regions with extreme landscape heterogeneity, suggesting future work to enhance generalizability. Future research will explore lightweight versions of BANet for real-time cadastral updates and extend the framework to handle multi-source data (*e.g.*., LiDAR or SAR) for improved robustness in complex environments.

CRediT authorship contribution statement

Qingyu Li: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lichao Mou:** Writing – review & editing, Methodology, Conceptualization. **Yilei Shi:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Xiao Xiang Zhu:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Xiaoxiang Zhu reports financial support was provided by German Federal Ministry of Education and Research. Xiaoxiang Zhu reports financial support was provided by German Research Foundation.

Acknowledgments

The work is jointly supported by the Excellence Strategy of the Federal Government and the Länder through TUM Innovation Network EarthCare, by the German Federal Ministry of Education and Research (BMBF) in the framework of the international future AI lab "AI4EO – Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond" (grant number: 01DD20001), by German Federal Ministry for Economic Affairs and Climate Action in the framework of the "national center of excellence ML4Earth" (grant number: 50EE2201C) and by the Munich Center for Machine Learning.

Data availability

No data was used for the research described in the article.

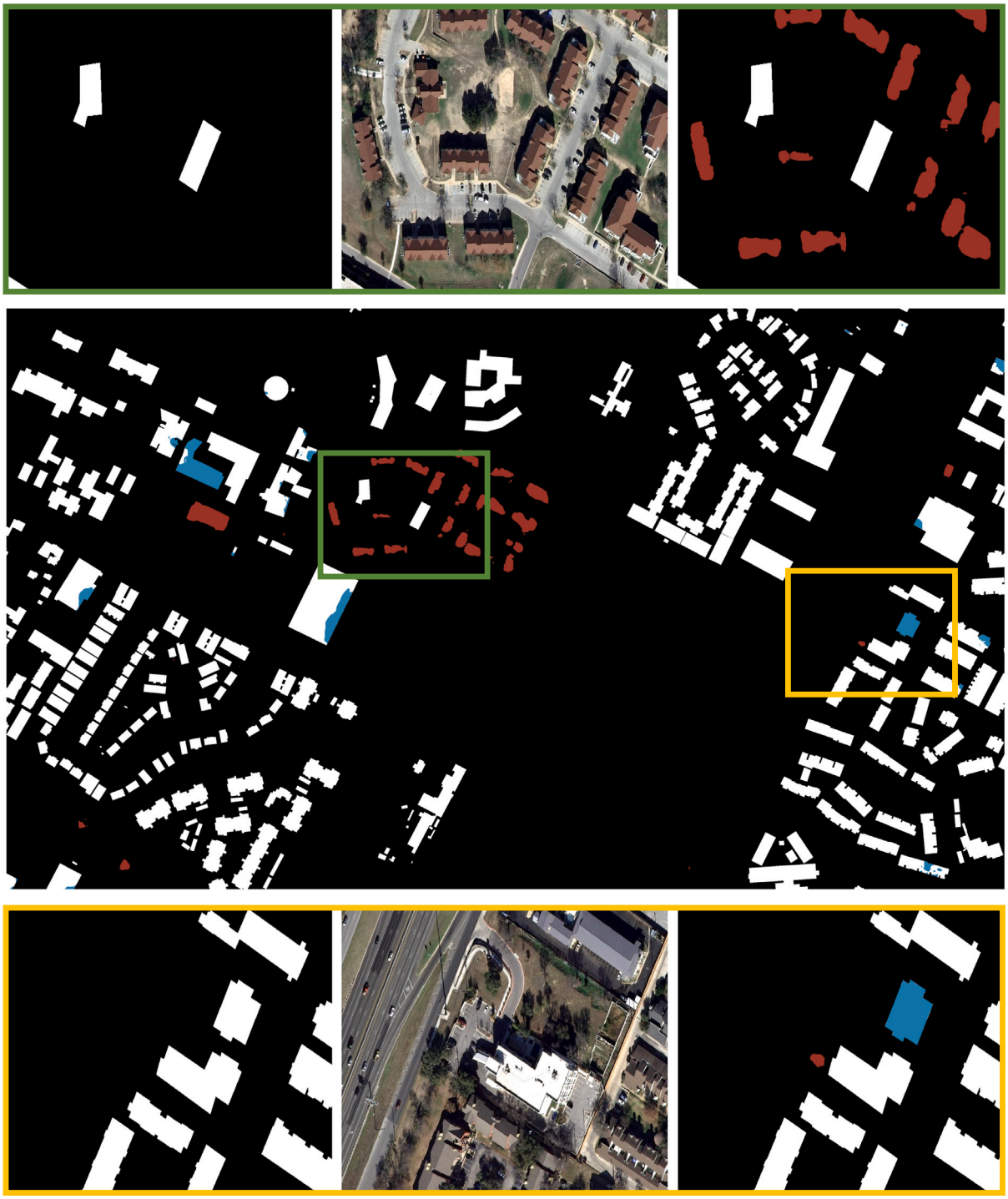


Fig. 14. Example segmentation results of newly built (red) and demolished (blue) buildings in an area of the city of Austin (1,350,000 m^2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Chen, J., He, P., Zhu, J., Guo, Y., Sun, G., Deng, M., Li, H., 2023. Memory-contrastive unsupervised domain adaptation for building extraction of high-resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote. Sens. 12 (10), 1662.

Chen, S., Yang, K., Stiefelhagen, R., 2021. DR-TANet: Dynamic receptive temporal attention network for street scene change detection. In: IV.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV.

Dai, X., Xia, M., Weng, L., Hu, K., Lin, H., Qian, M., 2023. Multi-scale location attention network for building and water segmentation of remote sensing image. IEEE Trans. Geosci. Remote Sens..

Feng, Y., Jiang, J., Xu, H., Zheng, J., 2023. Change detection on remote sensing images using dual-branch multilevel intertemporal network. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Guo, H., Shi, Q., Marinoni, A., Du, B., Zhang, L., 2021. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sens. Environ. 264, 112589.

Henssen, J., 1995. Basic principles of the main cadastral systems in the world. In: Proceedings of the One Day Seminar Held During the Annual Meeting of Commission. vol. 7.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57 (1), 574–586.

Jiang, X., Li, D., Chen, H., Zheng, Y., Zhao, R., Wu, L., 2022. Uni6D: A unified cnn framework without projection breakdown for 6D pose estimation. In: CVPR.

Kraff, N.J., Wurm, M., Taubenböck, H., 2020. The dynamics of poor urban areas-analyzing morphologic transformations across the globe using earth observation data. Cities 107, 102905.

Lei, T., Geng, X., Ning, H., Lv, Z., Gong, M., Jin, Y., Nandi, A.K., 2023. Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 61, 1–14.

Li, M., Liu, X., Wang, X., Xiao, P., 2023. Detecting building changes using multi-modal siamese multi-task networks from very high resolution satellite images. IEEE Trans. Geosci. Remote Sens..

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2021. Building footprint generation through convolutional neural networks with attraction field representation. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Li, Q., Mou, L., Hua, Y., Shi, Y., Zhu, X.X., 2022a. CrossGeoNet: A framework for building footprint generation of label-Scarce Geographical Regions. Int. J. Appl. Earth Obs. Geoinf. 111, 102824.

Li, Q., Shi, Y., Auer, S., Roschlaub, R., Möst, K., Schmitt, M., Glock, C., Zhu, X., 2020. Detection of undocumented building constructions from official geodata using a convolutional neural network. Remote. Sens. 12 (21), 3537.

Li, Q., Taubenböck, H., Shi, Y., Auer, S., Roschlaub, R., Glock, C., Kruspe, A., Zhu, X.X., 2022b. Identification of undocumented buildings in cadastral data using remote sensing: Construction period, morphology, and landscape. Int. J. Appl. Earth Obs. Geoinf. 112, 102909.

Li, Z., Yan, C., Sun, Y., Xin, Q., 2022c. A densely attentive refinement network for change detection based on very-high-resolution bitemporal remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Liao, C., Hu, H., Yuan, X., Li, H., Liu, C., Liu, C., Fu, G., Ding, Y., Zhu, Q., 2023. BCE-Net: Reliable building footprints change extraction based on historical map and up-to-date images using contrastive learning. ISPRS J. Photogramm. Remote Sens. 201, 138–152.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In: IGARSS.

OpenStreetMap contributors, 2017. Planet dump. retrieved from https://planet.osm.org, https://www.openstreetmap.org.

Revaud, J., Heo, M., Rezende, R.S., You, C., Jeong, S.-G., 2019. Did it change? Learning to detect point-of-interest changes for proactive map updates. In: CVPR.

Shao, R., Du, C., Chen, H., Li, J., 2021. SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network. Remote. Sens. 13 (18), 3750.

Shen, L., Lu, Y., Chen, H., Wei, H., Xie, D., Yue, J., Chen, R., Lv, S., Jiang, B., 2021. S2looking: A satellite side-looking dataset for building change detection. Remote. Sens. 13 (24), 5094.

Shu, Q., Pan, J., Zhang, Z., Wang, M., 2022. DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 112, 102940.

Varghese, A., Gubbi, J., Ramaswamy, A., Balamuralidhar, P., 2018. ChangeNet: A deep learning architecture for visual change detection. In: ECCVW.

Wang, X., Du, J., Tan, K., Ding, J., Liu, Z., Pan, C., Han, B., 2022. A high-resolution feature difference attention network for the application of building change detection. Int. J. Appl. Earth Obs. Geoinf. 112, 102950.

Wu, C., Du, B., Zhang, L., 2023. Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection. IEEE Trans. Pattern Anal. Mach. Intell..

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: NIPS.

Xu, L., Li, Y., Xu, J., Zhang, Y., Guo, L., 2023. BCTNet: Bi-branch cross-fusion transformer for building footprint extraction. IEEE Trans. Geosci. Remote Sens. 61, 1–14.

Xu, Z., Xu, C., Cui, Z., Zheng, X., Yang, J., 2022. CVNet: Contour vibration network for building extraction. In: CVPR.

Yuan, Y., Chen, X., Wang, J., 2020. Object-contextual representations for semantic segmentation. In: ECCV.

Zhang, J., Liu, R., Shi, H., Yang, K., Reiß, S., Peng, K., Fu, H., Wang, K., Stiefelhagen, R., 2023. Delivering arbitrary-modal semantic segmentation. In: CVPR.

Zhang, X., Tian, M., Xing, Y., Yue, Y., Li, Y., Yin, H., Xia, R., Jin, J., Zhang, Y., 2022. ADHR-CDNet: Attentive differential high-resolution change detection network for remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L., 2020. A single stream network for robust and real-time RGB-D salient object detection. In: ECCV. Springer.

Zheng, D., Li, S., Fang, F., Zhang, J., Feng, Y., Wan, B., Liu, Y., 2023. Utilizing bounding box annotations for weakly supervised building extraction from remote sensing images. IEEE Trans. Geosci. Remote Sens..

Zhou, F., Xu, C., Hang, R., Zhang, R., Liu, Q., 2023. Mining joint intra-and inter-image context for remote sensing change detection. IEEE Trans. Geosci. Remote Sens..

Zhou, W., Yang, E., Lei, J., Wan, J., Yu, L., 2022. PGDENet: Progressive guided fusion and depth enhancement network for RGB-D indoor scene parsing. IEEE Trans. Multimed..

Zorzi, S., Bazrafkan, S., Habenschuss, S., Fraundorfer, F., 2022. PolyWorld: Polygonal building extraction with graph neural networks in satellite images. In: CVPR.