Remote Sensing of Environment 297 (2023) 113779

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas

Yinxia Cao [a,b], Xin Huang [b,*], Qihao Weng [a]

[a] Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[b] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China

ARTICLE INFO

Keywords:
Built-up areas
Digital change detection
Weakly supervised learning
Class activation map
Noise correction
ZY-3

ABSTRACT

Accurate change detection of built-up areas (BAs) fosters a comprehensive understanding of urban development. The post-classification comparison (PCC) is a widely-used change detection method by classification and temporal comparison. For classification, image-level labeling is an efficient alternative to pixel-level one for pixel-wise weakly supervised segmentation, which frequently applies pixel-level pseudo labels generated from class activation map (CAM) to train semantic segmentation networks. CAM can be obtained from classification networks trained with image-level labels and can indicate the spatial location of objects. The existing studies are subject to the following issues: 1) They only rely on the single-scale and low-resolution CAM, but ignore the multi-scale property of BAs; 2) Pixel-level pseudo labels usually contain noises (e.g., omissions and false alarms); 3) The temporal correlation between multi-temporal images is less considered in PCC. To address these limitations, this paper proposed a multi-scale weakly supervised learning method, which utilized a large number of single-temporal high-resolution images and image-level labels to detect BA changes. This method consisted of three modules: 1) multi-scale CAM for BA pseudo label generation; 2) adaptive online noise correction for BA detection; and 3) generation of reliable pseudo labels for BA change detection. Based on ZY-3 images (2.5 m), we constructed the first multi-view datasets for both BA detection and change detection. Each ZY-3 image includes a multi-spectral image with red, green, blue, and near-infrared bands and a multi-view image with nadir-, forward-, and backward-views. The BA detection dataset contained 86,166 image-level samples (256 × 256 pixels for each sample), covering 48 major cities in China, while the BA change detection dataset consisted of ZY-3 bi-temporal images at rapidly urbanizing areas (i.e., Beijing and Shanghai). Experiments showed that the proposed method can detect BA changes and suppress pseudo changes effectively, yielding 88.2% F1-score in BA detection and 79.3% for Shanghai and 78.5% for Beijing in change detection. Further analysis demonstrated the proposed method to be advantageous in the following two fronts: 1) the image-level weak labels can achieve pixel-wise BA change detection at low cost; and 2) the multi-scale CAM and temporal correlation are effective in the scenarios with limited labels. Datasets and codes will be accessed at https://github.com/lauraset/MSWS.

1. Introduction

The World Population Prospects 2022 indicates that the global population is expected to surpass 8 billion by mid-November 2022 (United Nations, 2022). To meet the living needs of the growing population, a large number of cities are expanding rapidly, accompanied by the conversion from natural and agricultural lands into built-up areas (BAs) (Liu et al., 2020). According to existing literature (Pesaresi et al., 2013; Wu et al., 2021), BAs can be defined as the areas being dominated by buildings, excluding main roads, parks, and large open spaces, and are usually used for residential, industrial, and commercial activities. Built-up area change is highly related to arable land conservation, urban planning, and land resource management (Chen et al., 2016; Deng et al., 2015; Musakwa and Van Niekerk, 2015). Thus, accurate change information of BAs is essential for a comprehensive understanding of urban development.

Medium-to-low resolution remote sensing images, such as DMSP/OLS, MODIS, and Landsat, have been widely used for urban area extraction (Huang et al., 2021; Pesaresi et al., 2016; Taubenbock et al., 2012; Uhl and Leyk, 2020; Zhou et al., 2018). High-resolution (HR)

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China.
E-mail address: xhuang@whu.edu.cn (X. Huang).

https://doi.org/10.1016/j.rse.2023.113779
Received 18 April 2023; Received in revised form 9 July 2023; Accepted 18 August 2023
Available online 24 August 2023

sensors (the spatial resolution ≤5 m), such as SPOT, QuickBird, WorldView, ZY-3, TerraSAR-X, and TanDEM-X, have pioneered the fine-scale ground observation mode (Gamba et al., 2011; Huang and Zhang, 2012), which can detect BA changes more accurately. This study focuses on efficient BA change detection from HR data.

Change detection aims to identify changes between different temporal images over the same region. Most change detection methods require simultaneous input of multi-temporal images to detect changes, e.g., image differencing, change vector analysis, and multi-date direct classification methods (Hussain et al., 2013). However, multi-temporal HR images have a relatively low temporal resolution, high acquisition cost, and limited spatial coverage. Among the change detection methods, the post-classification comparison is an appropriate method for HR imagery. This method decomposes the change detection task into classification and temporal comparison, which allows single-temporal images to be used for classification, and significantly reduces the cost of data acquisition. Thus, this study aims to investigate the post-classification comparison method to make full use of single-temporal images to detect BA changes.

Classification is the core of the post-classification comparison method and can be achieved through traditional methods and deep learning. Traditional methods need domain prior knowledge to manually design spectral, spatial, or contextual features, and then set thresholds or adopt classifiers to detect objects (Dalla Mura et al., 2010; Huang et al., 2014; Lu and Weng, 2007). Many studies have successfully identified BAs with HR images (Pesaresi et al., 2008; Shao et al., 2014; Tao et al., 2013). Since traditional methods rely on prior knowledge to design features, they become less effective when coping with complex image scenes. By contrast, deep learning is a data-driven approach and can automatically extract task-specific multi-level features, and thereby has been successfully applied to BA detection (Hafner et al., 2022; Wang et al., 2021; Wu et al., 2021). BA detection belongs to semantic segmentation, where each pixel is assigned a category label, i.e., BA or non-BA. To carry out semantic segmentation, deep learning needs plenty of accurate pixel-level samples, while annotating these samples is expensive and laborious. In this regard, weak labels, e.g., image-level labels (each image is tagged as one or more categories), points, scribbles, and bounding boxes, can greatly reduce the cost of label acquisition (Shen et al., 2023). Among them, image-level labels have the lowest acquisition cost and thus are focused on in this study.

Image-level labels inform the presence of objects but do not provide their locations, which makes accurate semantic segmentation challenging. According to the literature (Shen et al., 2023; Zhou, 2018), weak supervision means that training images have not accurate pixel-level labels, e.g., incomplete, inexact (i.e., coarse-grained), and inaccurate (i.e., noisy) labels. By contrast, for semi-supervision, the annotations are incomplete, meaning that only a subset of training images have per-pixel labels while the rest of the images are unannotated. This study focuses on a type of inexact labels, i.e., image-level labels (each image is labeled as a category), to achieve pixel-level BA segmentation at low labeling cost. Using such labels, weakly supervised segmentation approaches can be implemented by the following procedures: 1) Optimizing a classification network, e.g., Convolutional Neural Networks (CNNs) (He et al., 2016) or Transformer (Dosovitskiy et al., 2020; Wang et al., 2022) to acquire a class activation map (CAM). Considering that the CAM can indicate the detailed location of an object (Zhou et al., 2016), it can be easily converted to pixel-wise pseudo labels; and 2) Optimizing a segmentation network by using those pseudo labels for object segmentation (Chan et al., 2021). For the classification network, each image will be mapped to a category, while for the segmentation network, each pixel will be labeled as a category. Based on the above procedures, many approaches have been developed (Fan et al., 2020; Fang et al., 2022a; Kolesnikov and Lampert, 2016; Pathak et al., 2015; Wang et al., 2020). For example, Ahn et al. (2019) calculated the semantic affinity between pixels with pseudo labels, and utilized the affinity to optimize a class boundary extraction network for obtaining more complete objects. Li et al. (2021) introduced a fully connected conditional random field (CRF) loss function (Tang et al., 2018) for optimizing a building extraction network with pseudo labels, and successfully identified more complete buildings. These methods only employed the last feature layer of a classification network to compute CAM, which contained rich semantic information but was limited by the single scale and low spatial resolution. In general, existing CAM algorithms ignored and underexplored the multi-scale property of objects (e.g., BAs).

Pseudo labels generated from CAM can be used to train semantic segmentation networks. However, CAM mainly indicates the most discriminative regions of objects, which can lead to noises (e.g., omissions and false alarms) in pseudo labels. These noisy labels can lower the learning ability of deep networks (Song et al., 2022). Thus, some researchers had proposed advanced techniques to combat false labels. These techniques included robust loss functions (Ghosh et al., 2017; Ma et al., 2020), multi-network training (Malach and Shalev-Shwartz, 2017), and noise correction (Yi and Wu, 2019). Considering multi-network training is subject to the influence of many parameters, Dong et al. (2022) only used the prediction result of a single network to correct false labels and applied the corrected labels to re-train the network for land cover mapping. They found that noise correction outperformed robust loss functions. However, the method developed by Dong et al. (2022) still required offline storage of corrected labels and multi-round network training, which is time-consuming and inefficient. Besides, due to the lack of clean labels, the timing of noise correction is difficult to determine. Existing methods usually set the correction time empirically.

After completing the classification, the post-classification comparison method can identify changes by comparing the multi-date classification results. However, the direct pixel-level comparison ignores the temporal correlation of multi-temporal images, and thus, may introduce lots of pseudo changes (Singh, 1989). Previous studies have attempted to incorporate temporal correlation information into the classification to reduce pseudo changes (Huang et al., 2017; Wu et al., 2017; Xian et al., 2009). Recently, deep learning opens up new ways to mine temporal information, but it still relies on high-quality samples (Xia et al., 2022). To mitigate this issue, some studies exploited pseudo labels to train change detection networks (Fang et al., 2022b; Gong et al., 2017, 2020; Li et al., 2019). For instance, Fang et al. (2022b) designed a sample selection method for extracting pseudo labels to optimize a binary change detection network, demonstrating the effectiveness of pseudo labels. However, these pseudo labels are usually generated from classification models and thus they suffer from classification errors, leading to some unreliable regions, which will lower the generalization ability of deep networks.

In summary, the post-classification comparison method adopts a strategy of classification followed by temporal comparison for change detection, but the following issues remain to be examined:

1) For classification, image-level labels can reduce the label acquisition cost, but existing weakly supervised segmentation methods suffer from the single scale and low spatial resolution of CAM, ignoring the multi-scale property of BAs.
2) Pseudo labels generated from CAM usually contain noises (e.g., omissions and false alarms), which can lower the learning ability of deep networks. In this regard, noise correction can mitigate the effect of noisy labels, while it is difficult to determine the timing of correction, i.e., when to correct the noisy labels.
3) Concerning temporal comparison, existing methods commonly use pseudo labels from classification models for optimizing change detection networks to mine the temporal correlation of multi-temporal images, but these pseudo labels usually contain unreliable regions, lowering the performance of networks.

To address these limitations, this paper proposed a multi-scale weakly supervised learning method by using a large number of single-

Table 1
The built-up area (BA) detection dataset.

City	#BA	#Non-BA	City	#BA	#Non-BA	City	#BA	#Non-BA
Baotou	539	381	Hohhot	314	310	Shijiazhuang	1970	777
Beijing	2314	586	Huizhou	290	86	Taiyuan	645	210
Changchun	1021	308	Jinan	1092	1863	Tangshan	572	108
Changsha	774	726	Kunming	809	891	Tianjin	2015	3803
Changzhou	1017	103	Lanzhou	250	42	Urumqi	530	1611
Chengdu	2005	458	Lhasa	159	491	Weinan	126	82
Chongqing	937	788	Nanchang	810	520	Wuhan	1159	1674
Dalian	636	3415	Nanjing	1029	1323	Wuxi	1019	499
Dongguan	1839	398	Nanning	601	555	Xi'an	1820	441
Foshan	1250	81	Ningbo	1023	1185	Xining	379	635
Fuzhou	811	1644	Ordos	56	86	Yantai	476	245
Guangzhou	1810	876	Qingdao	1292	1952	Yinchuan	338	430
Haikou	294	71	Quanzhou	1283	1626	Yulin	110	528
Hangzhou	1073	575	Shanghai	4214	2976	Zhaoqing	149	107
Harbin	351	303	Shenyang	1599	1143	Zhengzhou	970	460
Hefei	964	736	Shenzhen	1827	1091	Zhuzhou	222	184

Note: The total number of samples is 86,166, consisting of 46,783 BA samples and 39,383 non-BA samples. Each sample has a spatial extent of 256 × 256 pixels.

temporal images and image-level labels to detect BA changes. The method consisted of three modules: 1) a multi-scale classification network using image-level labels to obtain multi-scale CAM for generating pixel-level BA pseudo labels; 2) a BA detection network with adaptive online noise correction using pseudo labels; and 3) extraction of reliable change pseudo labels based on the BA detection network for optimizing a BA change detection network. By utilizing ZY-3 multi-view images, we constructed the multi-view BA detection and change detection datasets. The BA detection dataset contained 86,166 image-level samples (256 × 256 pixels for each sample), covering 48 major cities in China, while the BA change detection dataset consisted of ZY-3 bi-temporal images for two rapidly urbanizing regions, i.e., Beijing and Shanghai. Each ZY-3 image included a multi-spectral image (with red, green, blue, and near-infrared bands) and a multi-view image (with nadir, forward, and backward views), where the former can provide rich spectral-spatial details and the latter the vertical properties of BAs. To our best knowledge, these are the first multi-view satellite datasets for the purpose of BA detection and change detection.

The remainder of the paper has four sections. Sections 2 and 3 present data and method, respectively. Sections 4 and 5 describe the results and discussions, respectively. Conclusions are summarized in Section 6.

2. Data

2.1. The BA detection dataset

We constructed a BA detection dataset for training (Table 1). In detail, we collected 61 ZY-3 image sets between 2014 and 2017 (Huang et al., 2020b), which cover 48 Chinese cities and exhibit diverse BAs (Fig. 1). ZY-3 images were provided by the China Centre for Resources Satellite Data and Application (http://www.cresda.com/). Each image set is composed of two images covering the same area: 1) a multi-

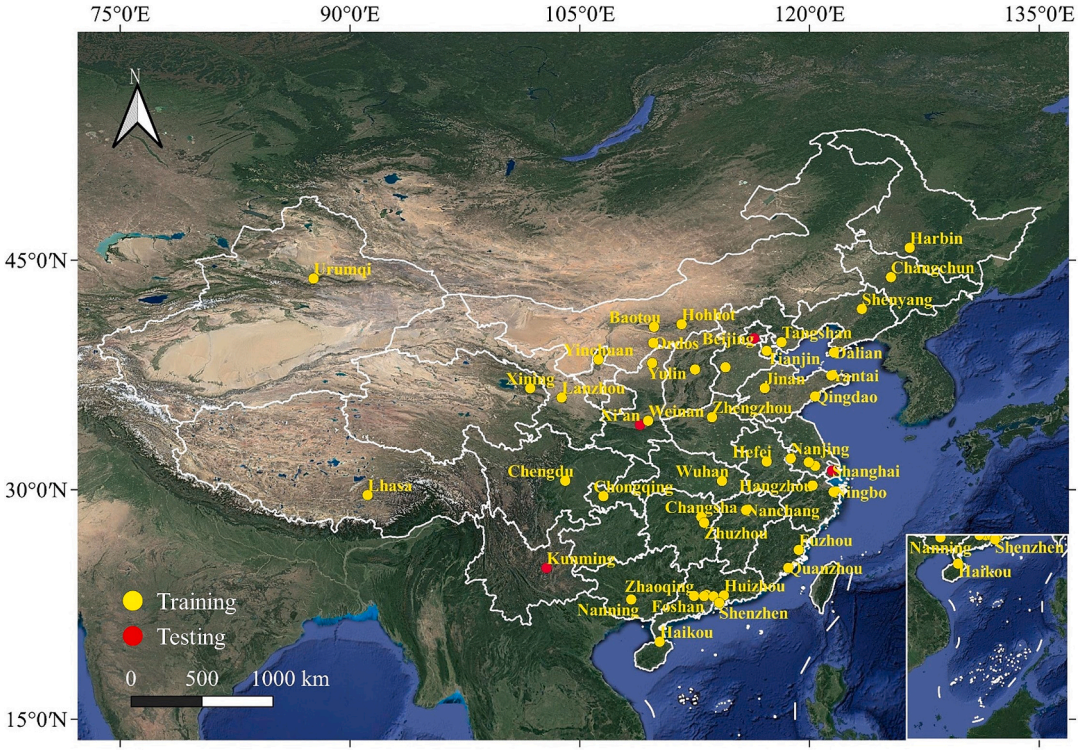

Fig. 1. The spatial distribution of the BA detection dataset.

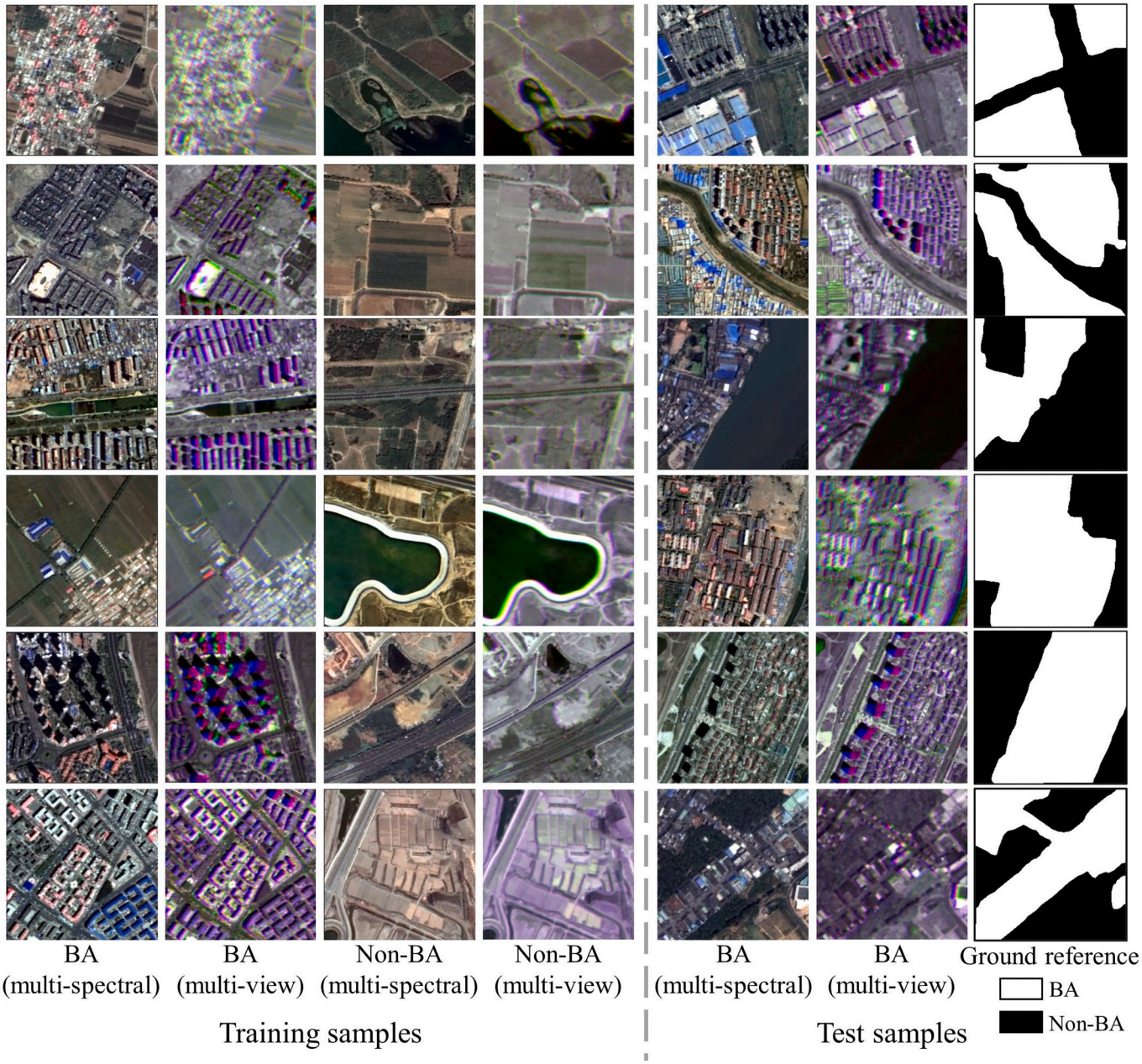

Fig. 2. Examples of the built-up area (BA) detection dataset. The multi-view images are visualized as the combination of the nadir (Red), forward (Green), and backward (Blue) panchromatic images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spectral image with red, green, blue, and near-infrared bands (spatial resolution = 5.8 m); 2) a multi-view image with three observation angles, i.e., nadir (2.1 m), +22° forward (2.5–3.5 m), and − 22° backward (2.5–3.5 m). Preprocessing steps, including radiometric correction, orthorectification, and image registration, were performed on all images (Huang et al., 2020b). The reference image for image registration is the nadir image. Then, we resampled the images to 2.5 m. Subsequently, we fused the nadir and multi-spectral images by pan-sharpening (Laben and Brower, 2000). Finally, multi-spectral and multi-view images were stacked along the channel dimension, resulting in 7-band feature images.

We cropped all the images (without overlapping) into blocks of 256 × 256 pixels. In the network training, we designated and collected blocks with building coverage over 15% as BA samples and blocks without buildings as non-BA samples. To ensure the performance of classification networks, we chose the percentage of 15% and this value has also been adopted by the existing study (Yan et al., 2022). After manual inspection, we obtained 86,166 samples, consisting of 46,783 BA samples and 39,383 non-BA samples (Table 1). To avoid the overfitting problem, we excluded Beijing, Shanghai, Kunming, and Xi'an from the BA datasets. The four cities were used for testing, while the remaining cities were for training (Fig. 1). For the training cities, the samples of each class were randomly split into 6:4 as the training set and the validation set, respectively. Note that these samples are at the image level, without the pixel-level annotation. Therefore, to evaluate the proposed method on pixel-level BA detection, we randomly selected 200 BA samples from the validation set as the test set, which was manually interpreted to obtain pixel-level BA labels (the ground reference in Fig. 2). As shown in Fig. 2, it can be observed that multi-spectral images provide rich spectral-spatial details and multi-view images the vertical information of ground objects. For example, BAs have different responses to different viewing angles, especially for high-rise buildings.

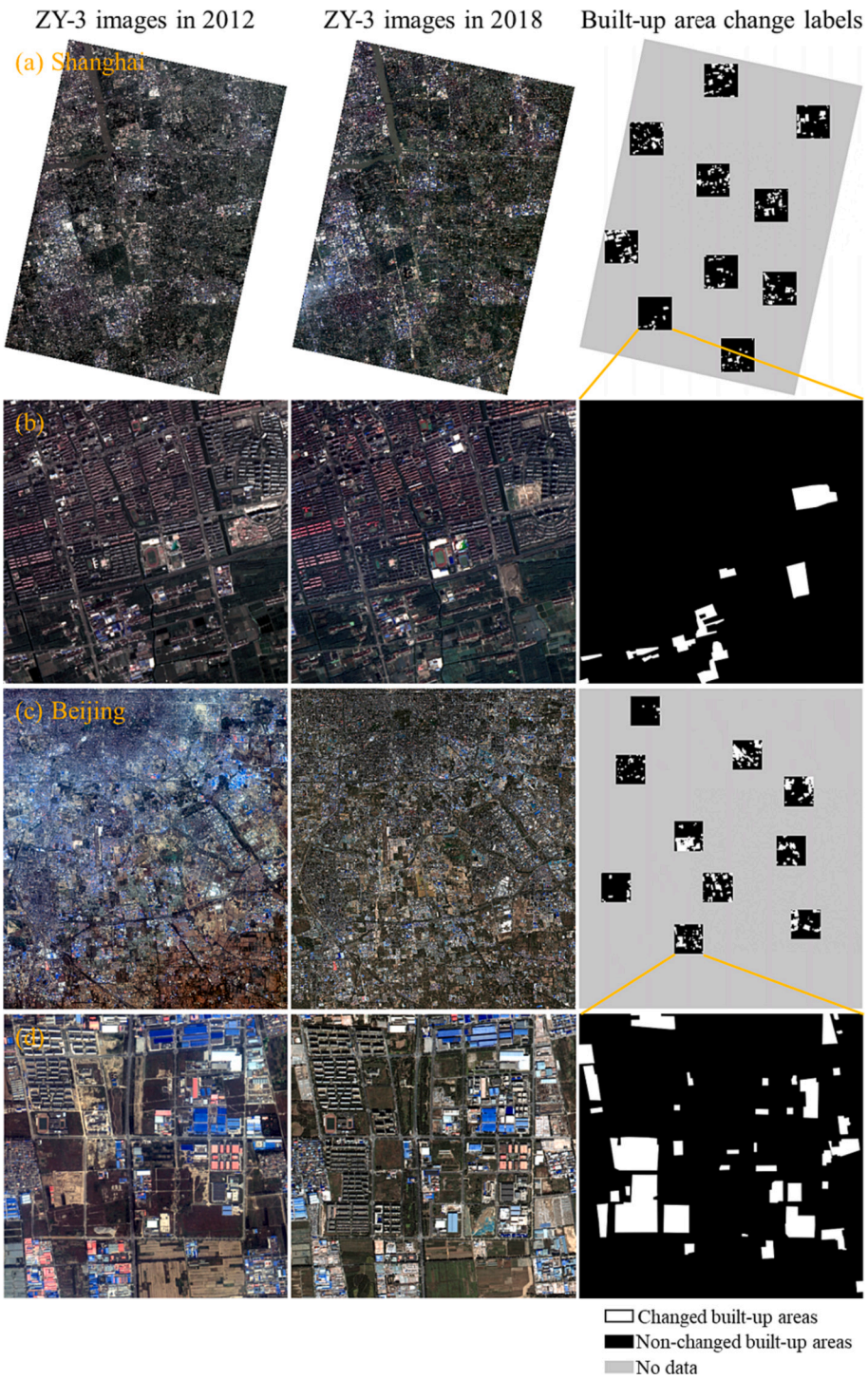

Fig. 3. Illustration of the BA change detection dataset. (a) and (c) are located in Shanghai and Beijing, respectively. (b) and (d) are test samples with a size of 1024 × 1024 pixels.

However, non-BAs, e.g., roads, water bodies, and arable land, have similar spectral responses to different viewing angles. This multi-view property can complement the spectral-spatial features provided by multi-spectral images, and their combination can provide a more complete picture of BAs.

2.2. The BA change detection dataset

We produced a change detection dataset for two rapidly urbanizing areas for testing (Fig. 3). One is located in Shanghai, where two ZY-3 images (8681 × 10,965 pixels) were obtained on Sept. 18, 2012, and Sept. 30, 2018, respectively. The other is distributed in Beijing, where two ZY-3 images (9916 × 11,122 pixels) were acquired on Oct. 11, 2012, and Oct. 6, 2018, respectively. The image preprocessing steps can be seen in Section 2.1. For testing accuracy, we randomly generated 10 blocks of 1024 × 1024 pixels for each city and manually interpreted the BA changes. These patches cover urban fringe areas that have undergone extensive building construction and demolition, which, therefore, are challenging for accurate change detection. Note that these pixel-level samples are only used for testing (not for training). The labels that are

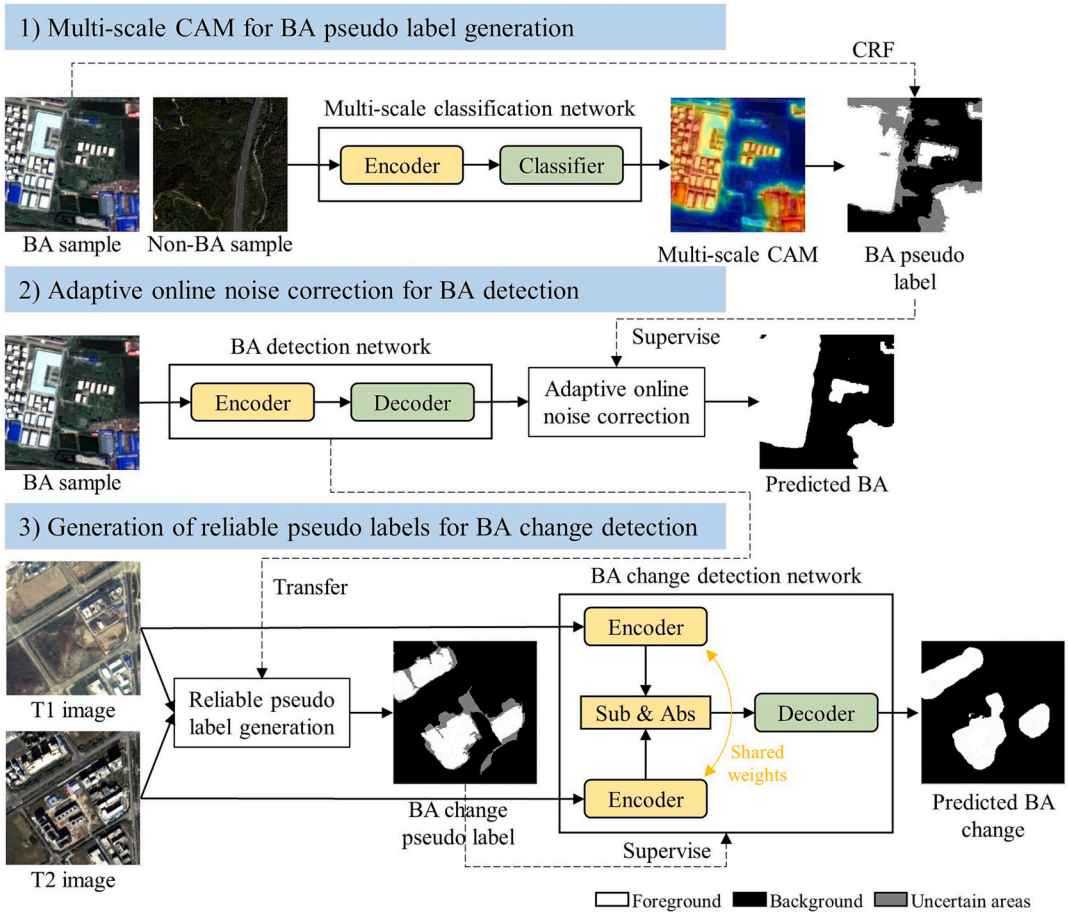

Fig. 4. Flowchart of the proposed approach.

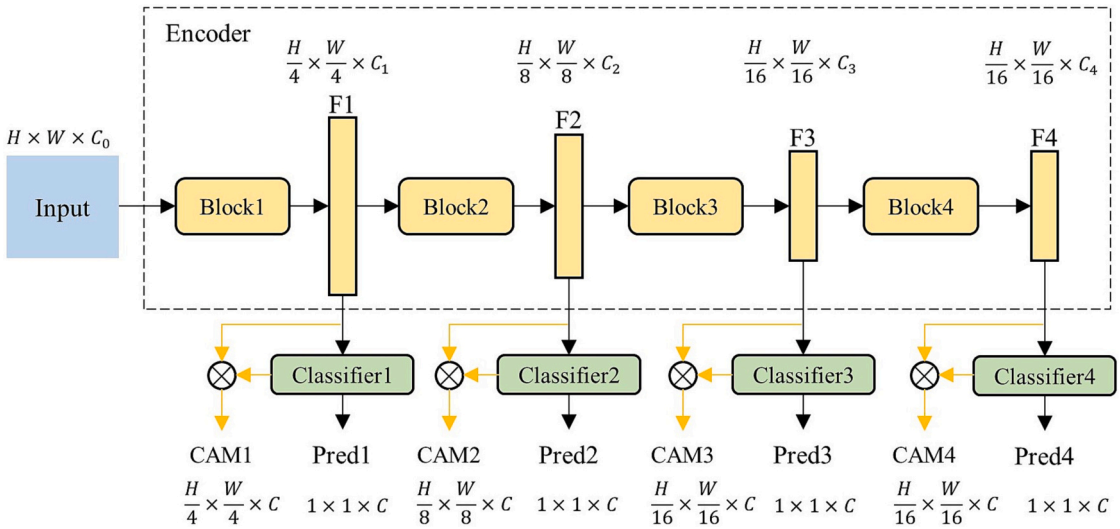

Fig. 5. Illustration of the classification network.

generated by the proposed pseudo label generation algorithm (see Section 3.3) are used for training the change detection network, without the need to annotate the images. To train the proposed method, we cropped all the images into samples of 256 × 256 pixels except for the test region. The overlapping percentage of samples was set to 50%. Finally, we generated 3575 training samples (256 × 256 pixels for each sample) in Shanghai and 6300 training samples in Beijing.

3. Methodology

The proposed method is composed of three modules (Fig. 4): 1) multi-scale CAM for BA pseudo label generation (Section 3.1); 2) adaptive online noise correction for BA detection (Section 3.2); and 3) generation of reliable pseudo labels for BA change detection (Section 3.3). Details are presented below.

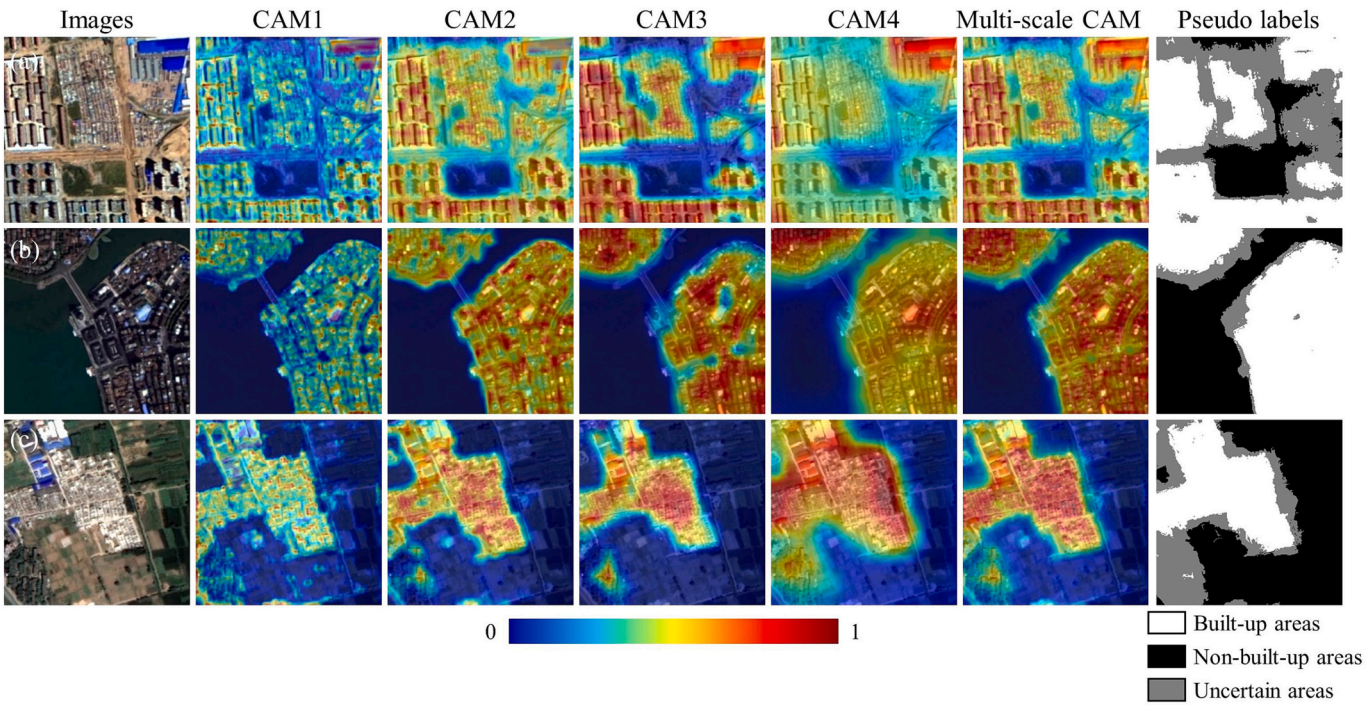

Fig. 6. Illustration of multi-scale CAM and pseudo labels.

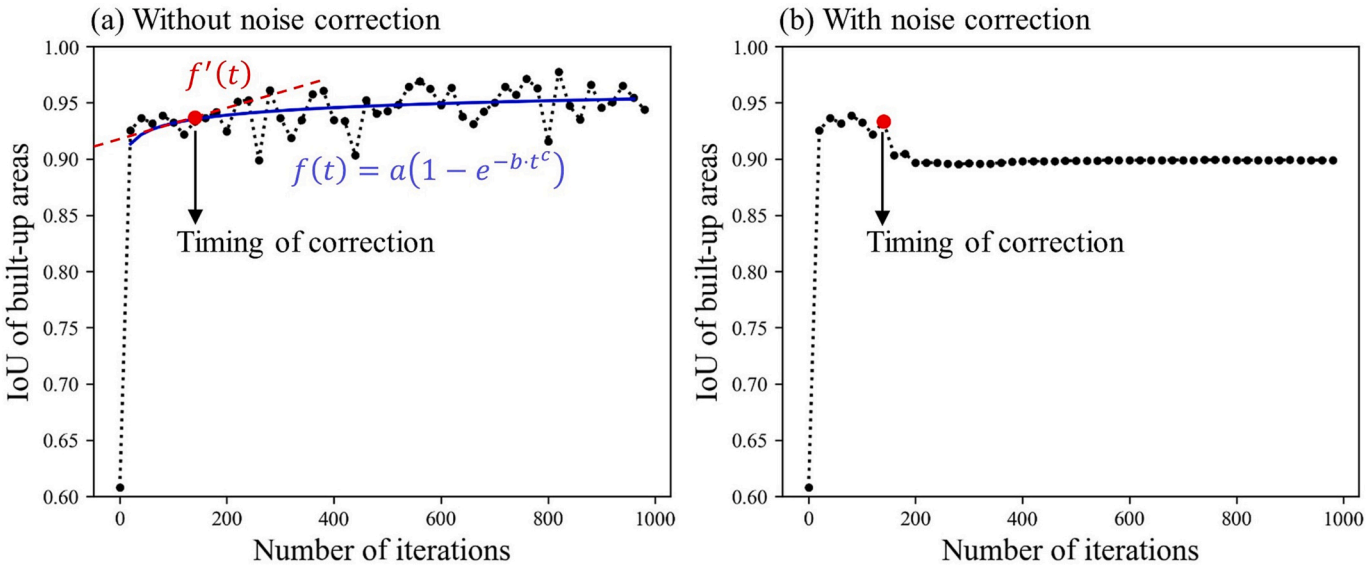

Fig. 7. The IoU curves of built-up areas without and with noise correction. Black line: the IoU curve of built-up areas. Blue line: the fitted exponential function (t) (Eq. (3)). Red line: the first-order derivatives ($f'(t)$) of $f(t)$. Red point: timing of correction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. Multi-scale CAM for BA pseudo label generation

We proposed a multi-scale CAM method to enhance the multi-scale representation capability and spatial resolution of CAM to obtain fine pseudo labels. The method is composed of three parts: 1) multi-scale classification network training; 2) multi-scale CAM generation; and 3) BA pseudo label generation.

3.1.1. Multi-scale classification network training

We trained a multi-scale classification network. Specifically, with image-level labels, image classification networks can extract multi-level features of the input image and then apply a classifier to determine the output class. In this study, we used MiT-B1 of the Mix-transformer (MiT) family (Xie et al., 2021) for classification (Fig. 5). Mit-B1 has the advantage of high computational efficiency and accuracy, and it can capture long-distance dependencies compared with CNNs. As shown in Fig. 5, MiT-B1 consists of four feature extractor modules (Block1–4) that can capture multi-level features (F1–4), and the number of channels C_1, C_2, C_3, and C_4 are 64, 128, 320, and 512, respectively. Notice that the spatial dimension of F4 in the original MiT-B1 is $\frac{H}{32} \times \frac{W}{32}$ (H: height, W: width). In this study, to improve the spatial resolution of features, the stride of Block4 was modified from 2 to 1, and hence, the dimension of F4 was modified as $\frac{H}{16} \times \frac{W}{16}$.

Furthermore, after each feature layer, we added a classifier to predict

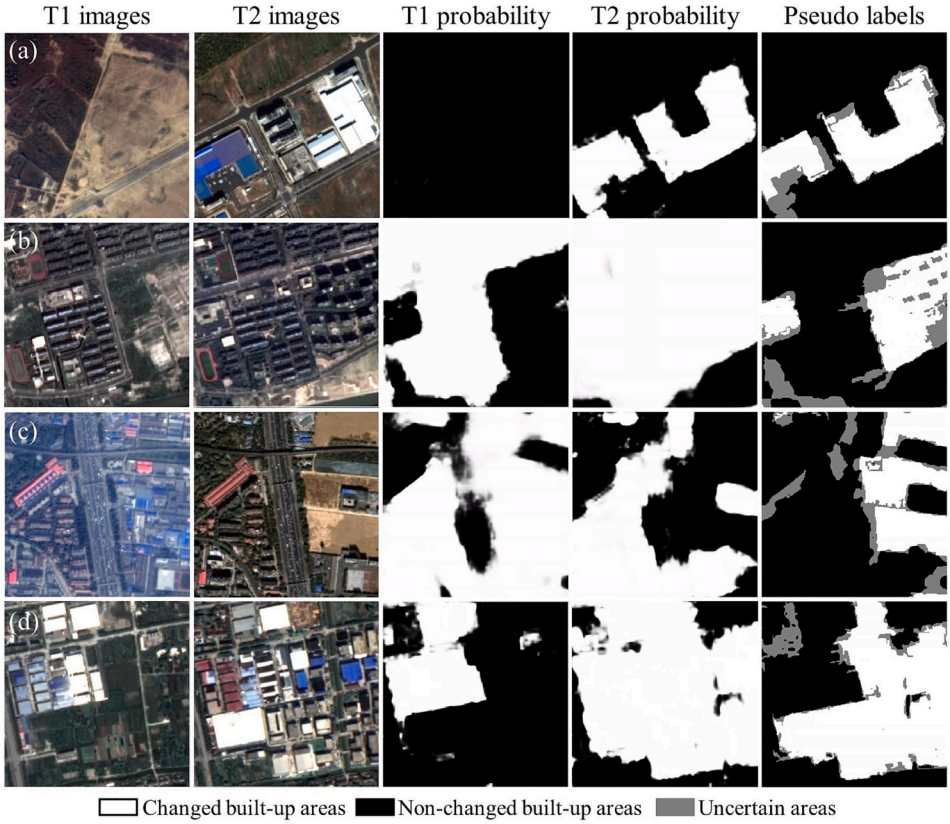

Fig. 8. Examples of reliable pseudo labels.

the output class. The classifier contains a convolutional layer (with weights of 1 × 1) and a global average pooling layer. Since this study involves binary classification (BA or non-BA), the prediction score (denoted by p, p ∈ {Pred1, Pred2,Pred3,Pred4}, see Fig. 5) has two channels (i.e., C = 2). We computed the sum of the cross-entropy loss L_{CE} between all prediction scores and image-level labels for network optimization:

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} q_{n,c}\cdot log\left(p_{n,c}\right) \quad (1)$$

where $q_{n,c}$ is the n-th referenced image-level label for class c. In this paper, $q_{n,c}$ is assigned as [0,1] (one-hot encoding) for BAs and [1,0] for non-BAs. log indicates the logarithmic function. N represents the number of samples. MiT-B1 was initialized by the ImageNet weight, and AdamW was selected as the optimizer. The learning rate was assigned as 6e-5, and it was updated by the polynomial descent with factor 1.0 by default (Xie et al., 2021). The input data was normalized by the min-max scaling. To avoid network overfitting, we adopted data enhancement techniques, including randomly horizontal or vertical flipping, random rotation (with the angle of (−90°, 90°)), and random grid shuffle. These training settings were used for subsequent experiments if not specified.

3.1.2. Multi-scale CAM generation

We employed the trained multi-scale classification network to generate feature maps (see Fig. 5). Then, we resampled feature maps to the same spatial extent as the input image, and calculated the multi-scale CAM using the following equation:

$$MSCAM_{i,j}^{c} = \sum_{s=1}^{S}\sum_{k=1}^{K} w_{k,s}^{c}\cdot F_{i,j,s}^{k} \quad (2)$$

Where $MSCAM_{i,j}^{c}$ denotes the multi-scale CAM of class c at the spatial location (i, j). The weight $w_{k,s}^{c}$ is derived from the convolution layer of the classifier at the scale s (see Classifier1–4 in Fig. 5). $F_{i,j,s}^{k}$ denotes the feature map with the dimension of H (height) × W (width) × K (the number of channels) at the scale s (see F1–4 in Fig. 5). Note that shallow features (e.g., F1) usually have a high spatial resolution but only provide low-level information, e.g., edges and corners, while deep features (e.g., F4) contain high-level information, e.g., semantics, but have low spatial resolution owing to the feature down-sampling. Therefore, we fused all the features to generate a multi-scale CAM, which can exhibit high-resolution multi-level information to explicitly describe the multi-scale properties of built-up areas. Examples of the multi-scale CAM are visualized in Fig. 6. It can be observed that the multi-scale CAM can retain better spatial details while identifying more complete BAs, compared to the single-scale.

3.1.3. BA pseudo label generation

To generate BA pseudo labels, we first linearly stretched the multi-scale CAM to the range of [0, 1]. Afterward, we applied the Otsu multi-threshold method with three thresholds (Otsu, 1979) on the CAM (written as T1, T2, T3). We labeled the pixels with CAM values > T3 as the foreground (i.e., BAs), those with values < T1 as the background (i.e., non-BAs), and those with values between T1 and T3 as uncertain areas. To improve the boundary of pseudo labels, we adopted the fully connected Conditional Random Field (CRF) algorithm (Krähenbühl and Koltun, 2011) to optimize the foreground and background regions. Examples of pseudo labels are displayed in Fig. 6. The training process of the multi-scale CAM for pseudo label generation is illustrated in Algorithm 1.

Algorithm 1. Multi-scale CAM for pseudo label generation.

Input: training images X, image-level labels Y, number of images N, batch size B, epoch E
Output: pixel-level pseudo labels $\widehat{\boldsymbol{Y}}$, multi-scale classification network parameters $\boldsymbol{\theta_{cls}}$, multi-scale CAM *MSCAM*
for $i = 1, \ldots, \frac{N}{B}E$ **do**
Fetch a batch of samples (x, y) from (X, Y)
Calculate $\boldsymbol{L_{CE}}$ by Eq. (1)
Update $\boldsymbol{\theta_{cls}}$
end for
Calculate *MSCAM* by Eq. (2)
Generate $\widehat{\boldsymbol{Y}}$ from *MSCAM* with the Otsu multi-threshold method and CRF

3.2. Adaptive online noise correction for BA detection

We proposed an adaptive online noise correction method and incorporated it into a segmentation network for BA detection to reduce the effects of false labels. The method adopts end-to-end training and can adaptively determine the timing of the correction. Specifically, we selected SegFormer (Xie et al., 2021) as the segmentation network, and MiT-B1 as the encoder (see Fig. 5). Supervised by noisy labels, the network usually learns correct labels quickly, and then gradually fits wrong labels (Liu et al., 2022). The learning process can be expressed by the exponential function:

$$f(t) = a\left(1 - e^{-b \cdot t^c}\right) \tag{3}$$

where a, b, and c are the fitting parameters and t is the current timing. We calculated the relative change (V) of the first-order derivatives ($f'(t)$) of $f(t)$ to indicate the speed of the network fitting:

$$\mathrm{V} = \frac{|f'(t) - f'(1)|}{|f'(1)|} \tag{4}$$

We applied $f(t)$ to fit the IoU curves of the training set and calculated V after each iteration. If V was greater than the threshold T_v, the noise correction would be conducted. Fig. 7 (a) visualizes the process of detecting the timing of correction. When starting noise correction, wrong labels will be corrected and then the network will start to fit corrected labels, mitigating the effect of wrong labels. Due to the presence of wrong labels in the training set, this process will lead to decrease of the IoU values of built-up areas, as reflected in Fig. 7 (b).

Noise correction is done by the following steps. In detail, within each mini-batch, the segmentation network was applied to the input image to produce a class probability map. To obtain updated labels, we assigned pixels with probability values $> \gamma$ (i.e., reliable pixels) as predicted categories of the network, while the remaining pixels as uncertain ones. Considering that initial labels may contain lots of clean labels, we employed both initial and updated labels to optimize the network prediction (p). Note that the training process is online and once the noise correction starts, the updated labels will be iteratively corrected and be used to train the model. For reliable pixels (q_r) in the initial or updated labels, the cross-entropy loss (L_{CE}) was calculated (Eq. (1)), while for the uncertain pixels (q_u), the dense energy loss (L_{Energy}) was used (Zhang et al., 2020). The uncertain pixels are used as one class to compute the dense energy loss, which can fully utilize RGB colors and spatial positions and has been proved effective (Zhang et al., 2020). Therefore, the final loss function can be written as:

$$L_{total} = L_{CE}(q_r, p) + L_{Energe}(q_u, p) \tag{5}$$

The proposed online noise correction algorithm was efficient since it did not require offline storage of updated labels and can adaptively determine the correction timing. The training process of the adaptive online noise correction is presented in Algorithm 2.

Algorithm 2. Adaptive online noise correction.

Input: training images X, pixel-level pseudo labels $\widehat{\boldsymbol{Y}}$, number of images N, batch size B, epoch E, threshold T_V and γ
Output: BA segmentation network parameters $\boldsymbol{\theta_{seg}}$, the exponential function *f(t)*, the speed of the network fitting V, updated labels $Y_{updated}$
for $i = 1, \ldots, \frac{N}{B}E$ **do**
Fetch a batch of samples $(x, \widehat{y})$ from $(X, \widehat{\boldsymbol{Y}})$
Calculate *f(t)* by Eq. (3)
Calculate V by Eq. (4)
if $V > T_V$ **then** *// start noise correction*
Generate the network prediction p with $\boldsymbol{\theta_{seg}}$
Obtain $Y_{updated}$ by thresholding p with γ
Calculate L_{total} with $Y_{updated}$, $\widehat{\boldsymbol{Y}}$, and p by Eq. (5)
else
Calculate L_{total} with $\widehat{\boldsymbol{Y}}$ and p by Eq. (5)
end if
Update $\boldsymbol{\theta_{seg}}$
end for

3.3. Generation of reliable pseudo labels for BA change detection

We designed a reliable pseudo label extraction method to identify BA changes, which can incorporate temporal information to reduce pseudo changes. Firstly, the trained BA detection network (Section 3.2) was used to predict the extent and probability of BAs for each temporal image. To improve the completeness of BAs, we applied the multi-resolution segmentation algorithm (Blaschke, 2010) on the stacked multi-temporal image to obtain objects. In this study, the scale parameter was empirically set to 50. Subsequently, we calculated the mean value of each object to generate object-based probability maps. Next, the Otsu multi-threshold (Section 3.1) was performed on the absolute difference of object-based probability maps at two dates, and produced object-level pseudo labels with changed, unchanged, and uncertain regions. Likewise, we also computed pixel-level pseudo labels. We merged object-level and pixel-level pseudo labels to produce the final reliable pseudo labels, where their inconsistent regions were viewed as uncertain ones. Examples of reliable pseudo labels are shown in Fig. 8. It can

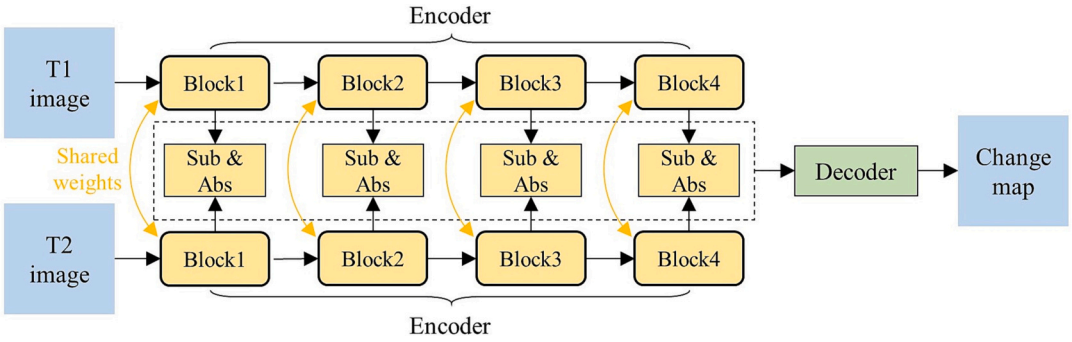

Fig. 9. The structure of the change detection network.

be observed that most of the newly built and demolished BAs were successfully identified, and their boundaries were well preserved because of the inclusion of object-based segmentation.

Finally, the generated reliable pseudo labels were employed to optimize a change detection network (Fig. 9). The meaning of "reliable pseudo labels" is explained below. Due to the lack of real labels, this study applied the trained BA detection network (Section 3.2) to generate pseudo labels for change detection. This approach can significantly lower the labeling cost. Considering that pseudo labels may contain unreliable pixels that can reduce the robustness of networks, we only selected reliable ones, according to the rules described in the first paragraph of Section 3.3. For simplicity, we referred to the reliable pixels in pseudo labels as reliable pseudo labels. The uncertain classes were ignored when training the change detection network. The network was composed of two encoders and one decoder. The two encoders were set to Mit-B1 (Fig. 5) and shared the same weight. To highlight changed areas, we calculated the absolute difference (see 'Sub & Abs' in Fig. 9) of features from the two encoders layer by layer, and applied it to the decoder. The decoder has the same structure as SegFormer (Xie et al., 2021), and can produce a pixel-level change map.

Table 2
Accuracy of different segmentation methods with image-level labels on the test samples of the BA detection dataset.

	CRF	OA	IoU	F1	Pre	Rec
SEC	√	0.832	0.757	0.862	0.847	0.877
IRNet	√	0.815	0.711	0.831	**0.913**	0.763
TSWS	√	0.842	0.767	0.868	0.860	0.877
Proposed method	×	0.846	0.769	0.870	0.875	0.865
Proposed method	√	**0.859**	**0.788**	**0.882**	0.879	**0.885**

Note: F1: F1-score; Pre: precision; Rec: recall. The highest scores are shown in bold.

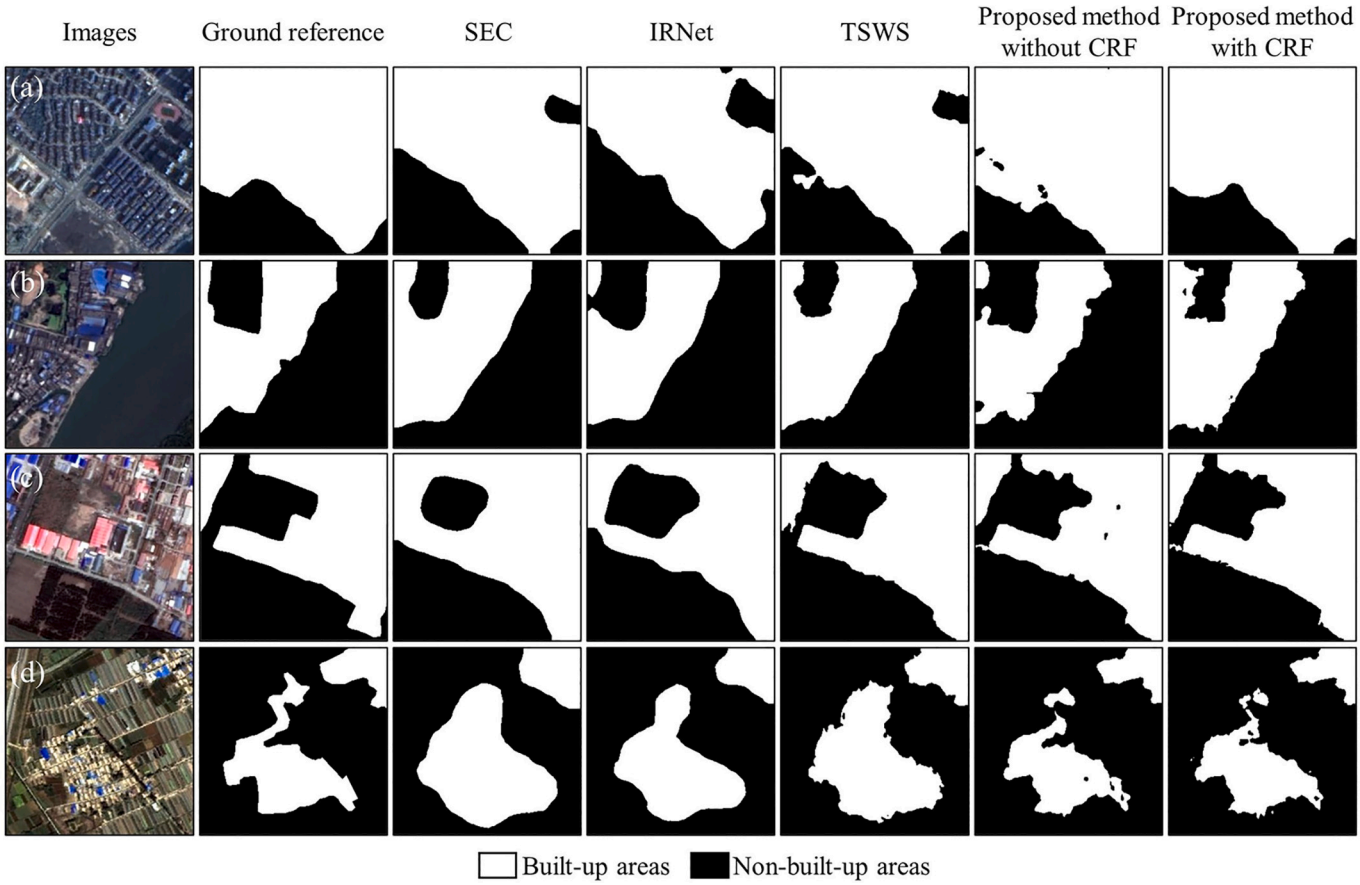

Fig. 10. Built-up area detection results with three advanced segmentation methods and the proposed one.

Table 3
The ZY-3 images in typical Chinese cities.

City	Imaging date (T1)	Imaging date (T2)	Image size (pixels)
Shenyang	2016-08-14	2021-03-25	20,898 × 22,814
Nanjing	2012-02-19	2014-11-13	10,515 × 13,751
Wuhan	2013-08-12	2018-10-28	23,340 × 23,974
Xi'an	2015-05-12	2017-06-27	18,715 × 22,011

3.4. Accuracy assessment

We selected five metrics for accuracy assessment as follows:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

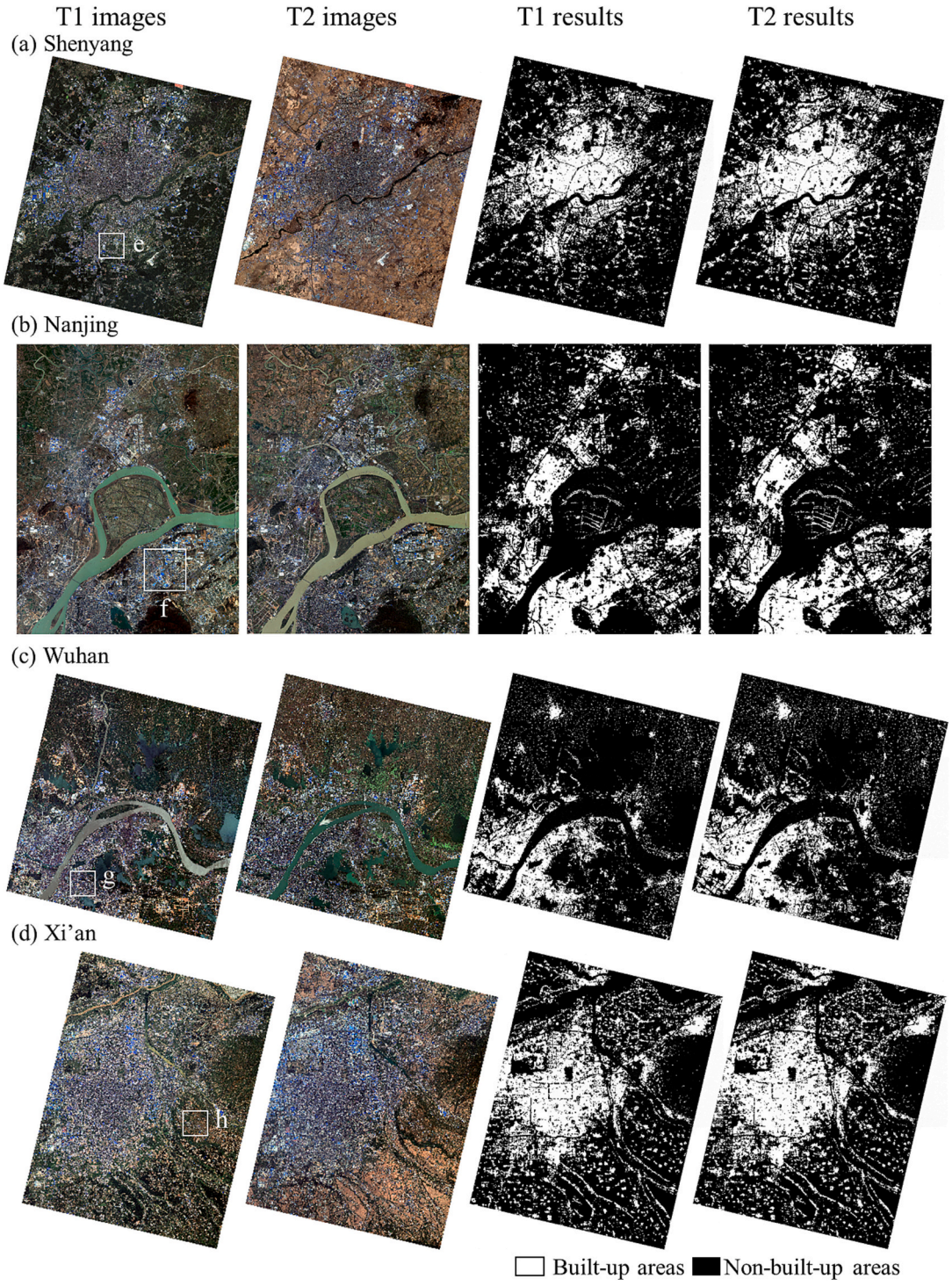

Fig. 11. Built-up area detection results in typical Chinese cities.

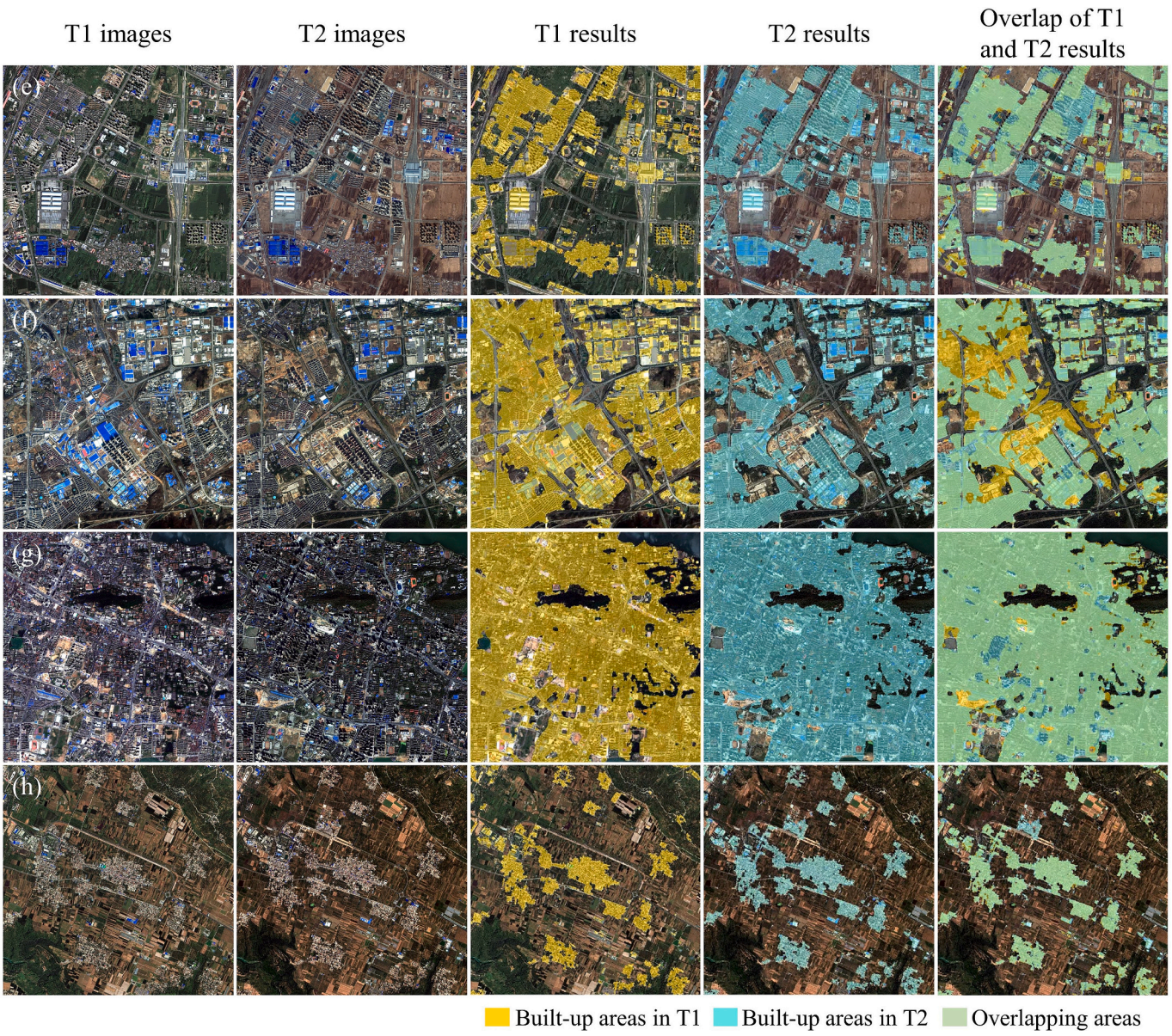

Fig. 12. Enlarged views of built-up area detection results for regions (e-h) in Fig. 11.

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where TP represents the number of pixels correctly assigned as foreground (e.g., BAs or changed BAs), FP represents the number of pixels incorrectly assigned as foreground, TN represents the number of pixels correctly assigned as background (e.g., non-BAs or non-changed BAs), and FN represents the number of pixels incorrectly assigned as background. OA denotes the overall accuracy of all categories. IoU reflects the degree of overlap between the predicted and referenced foreground. F1-score is a trade-off between precision and recall.

4. Results

4.1. Result of BA detection

To demonstrate the performance of the proposed method on BA detection, we compared three advanced segmentation methods: SEC (seed, expand and constrain) (Kolesnikov and Lampert, 2016), IRNet (inter-pixel relation network) (Ahn et al., 2019) and TSWS (two-step weakly supervised segmentation method) (Li et al., 2021). These methods can achieve pixel-level segmentation using image-level labels. Specifically, SEC designed three loss functions, i.e., seed, expand and constrain, to guide the semantic segmentation network to generate high-quality object boundaries. IRNet calculated the semantic affinity based on pseudo labels from CAM to train a class boundary extraction network and obtained more complete objects. TSWS introduced a fully connected conditional random field (CRF) loss function (Tang et al., 2018) to optimize a building segmentation network with pseudo labels from CAM and identified more complete buildings. The network structures for the three methods were made consistent with the proposed one. For a fair comparison, all methods used the pseudo labels with CRF (see Section

Table 4
Accuracy of BA change detection using the post-classification comparison (PCC) and the proposed method.

	Shanghai						Beijing				
	OA	IoU	F1	Pre	Rec		OA	IoU	F1	Pre	Rec
PCC	0.905	0.447	0.618	0.491	**0.833**	PCC	0.881	0.466	0.636	0.518	**0.825**
Proposed method	**0.962**	**0.657**	**0.793**	0.794	0.793	Proposed method	**0.947**	**0.646**	**0.785**	**0.813**	0.758
Network in Beijing	0.956	0.572	0.728	**0.838**	0.643	Network in Shanghai	0.939	0.603	0.753	0.777	0.730

3.1.3). To analyze the benefits of CRF, we compared the proposed method with and without CRF. Experiments were conducted with the test samples of the BA detection dataset (Section 2.1). As shown in Table 2, the proposed method consistently outperformed other ones in terms of overall metrics (i.e., OA, IoU, and F1). Moreover, it can be observed that CRF can significantly improve the overall metrics of BA segmentation. According to the visualization results in Fig. 10, our method detected more complete BAs and better retained the boundary information, due to the inclusion of multi-scale CAM and adaptive online noise correction (see Section 3).

We further visualized the detection results of built-up areas in typical Chinese cities, including Shenyang, Nanjing, Wuhan, and Xi'an, and the data details are present in Table 3. The image preprocessing steps can be seen in Section 2.1. The detection results of built-up areas are shown in Fig. 11 and Fig. 12. Overall, the proposed method can effectively identify built-up areas located in different geographical regions and exhibit

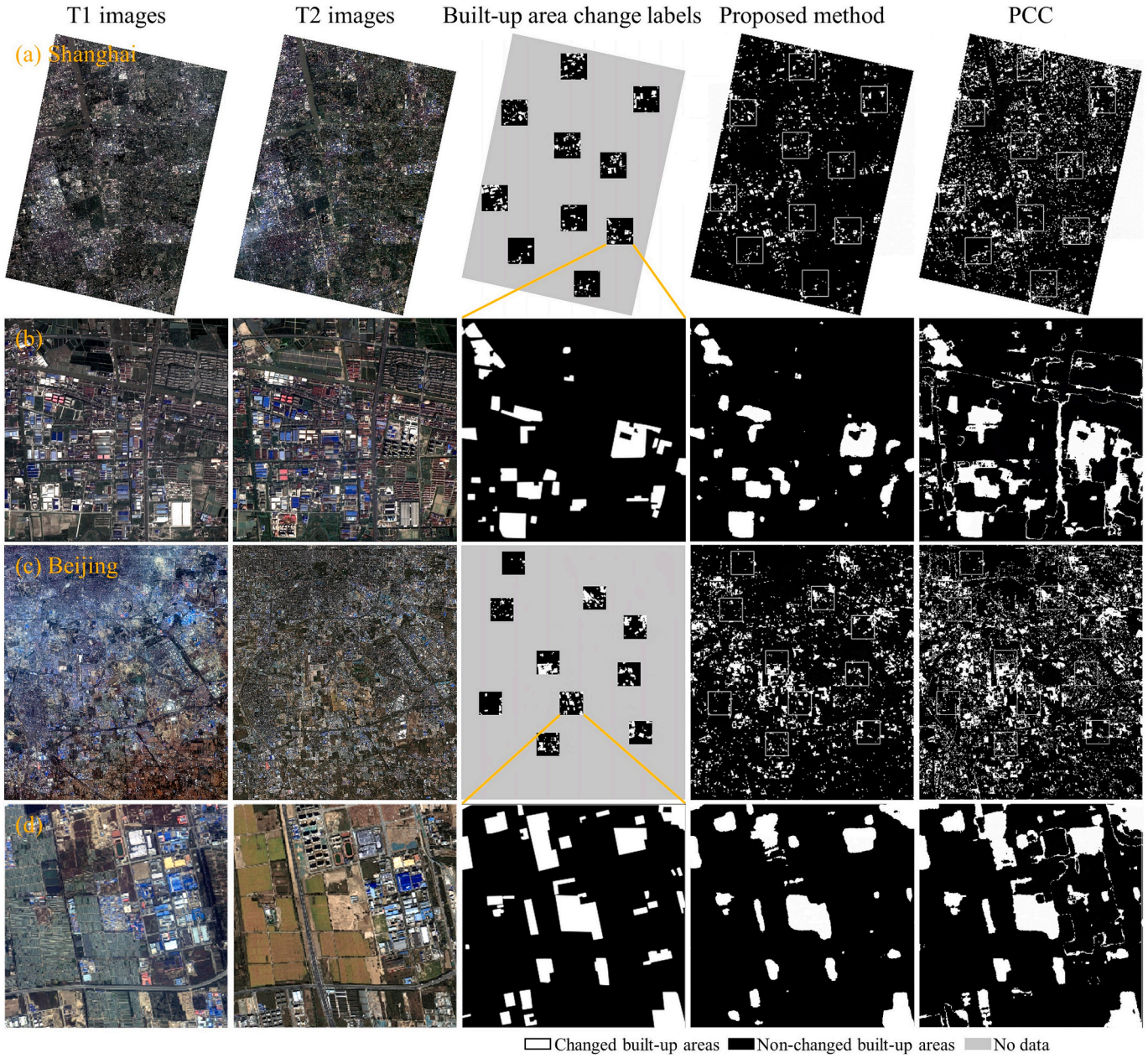

Fig. 13. Illustration of the post-classification comparison method (PCC) and the proposed one for built-up area change detection results.

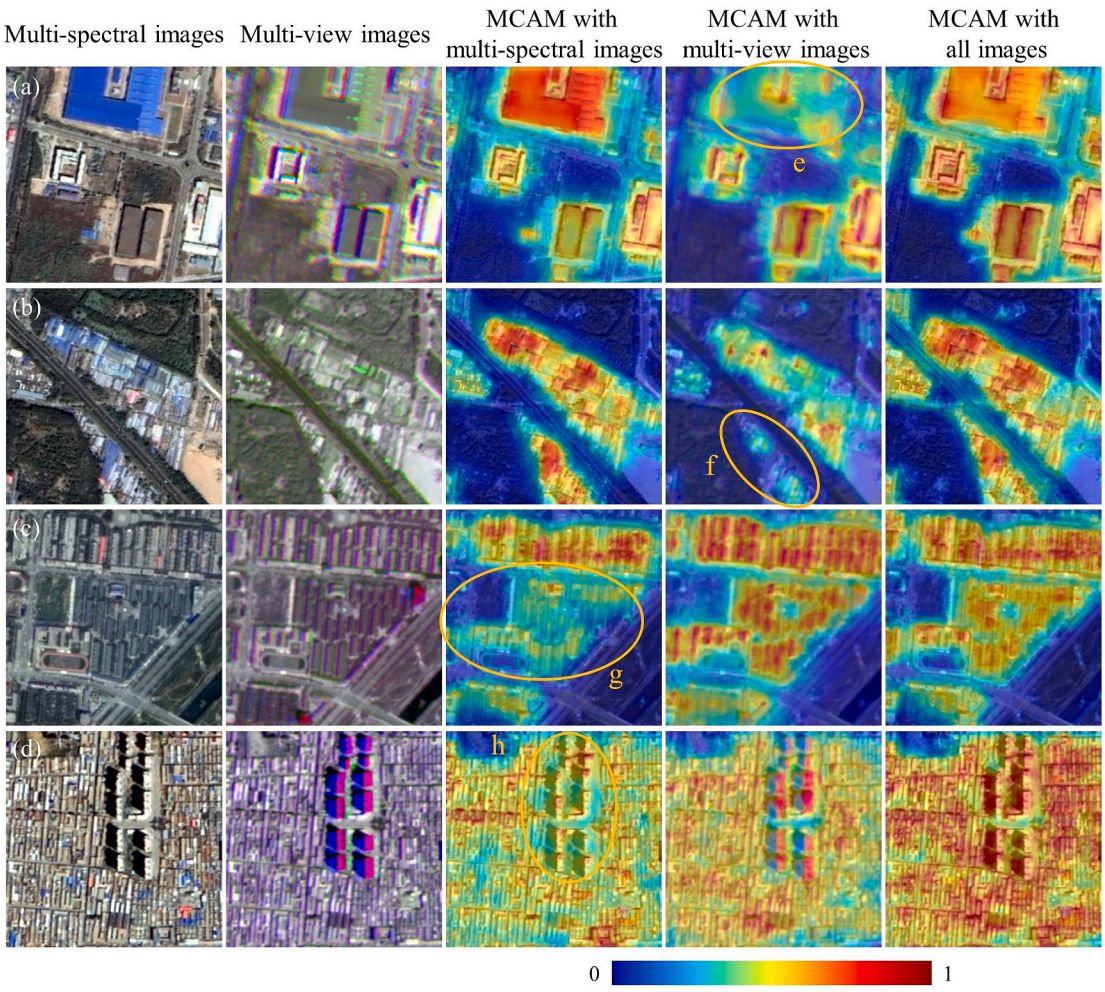

Fig. 14. Illustration of multi-scale class activation map (MCAM) generated with multi-spectral images, multi-view images, and their combination (i.e., all images). For multi-view images, nadir, forward, and backward images were displayed in red, green, and blue bands, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Accuracy of BA detection with multi-spectral images, multi-view images, and their combination (i.e., all images).

	OA	IoU	F1	Pre	Rec
Multi-spectral images	0.838	0.759	0.863	0.869	0.857
Multi-view images	0.800	0.693	0.819	0.890	0.758
All images	**0.859**	**0.788**	**0.882**	**0.879**	**0.885**

consistent performance on bi-temporal images, indicating that the proposed method is robust to differences in imaging conditions of bi-temporal images.

4.2. Result of BA change detection

To investigate the effectiveness of the proposed method for BA change detection, we compared it with the post-classification comparison method (PCC). For a fair comparison, we set the classification results of PCC as the extents of BAs predicted by the trained BA detection network (Section 3.2). For PCC, the change map was obtained by directly comparing the bi-temporal BAs, where the different areas of bi-temporal BAs were labeled as changes and the consistent areas as non-changes. Therefore, the difference between PCC and the proposed method is that the latter further generated reliable change pseudo labels for optimizing a change detection network (see Section 3.3) to mine the temporal correlation of multi-temporal images, which, is not considered by the former.

Experimental results on the test set of change detection are provided

Table 6
Ablation studies of the proposed method on BA detection.

	MSCAM	ONC	OA	IoU	F1	Pre	Rec
Baseline			0.830	0.751	0.858	0.857	0.859
Proposed method	✔		0.852	0.774	0.873	**0.895**	0.851
	✔	Fixed epoch = 1	0.846	0.767	0.868	0.887	0.850
	✔	Fixed epoch = 2	0.834	0.746	0.854	**0.895**	0.817
	✔	Adaptive	**0.859**	**0.788**	**0.882**	0.879	**0.885**

Note: MSCAM: multi-scale CAM. AONC: adaptive online noise correction.

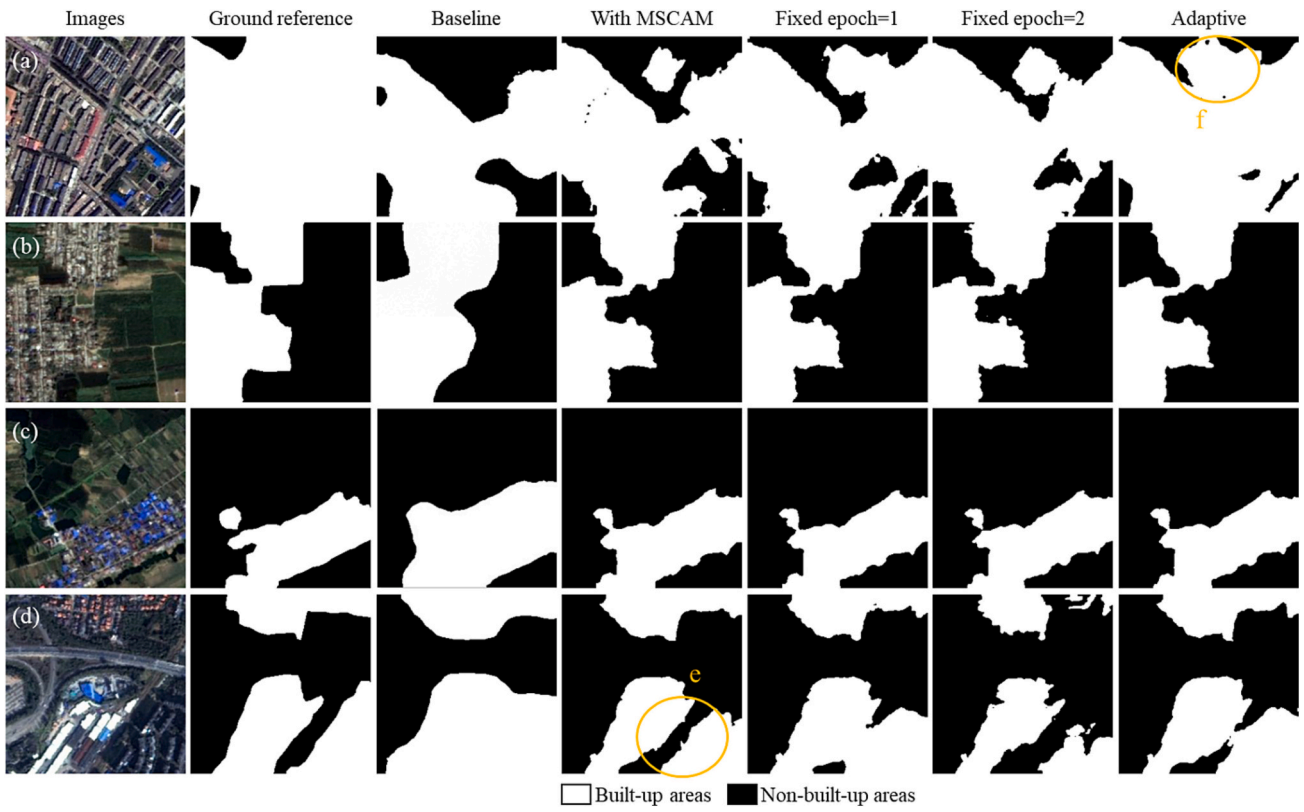

Fig. 15. Results of ablation studies of the proposed method on built-up area detection.

Table 7
Accuracy of different encoders on the test samples of the BA detection dataset.

Encoder	Params (M)	FLOPs (G)	MSCAM	OA	IoU	F1	Pre	Rec
Standard encoder of U-Net	18.9	13.1	×	0.619	0.379	0.550	0.926	0.391
	18.9	13.1	√	0.679	0.481	0.649	0.931	0.499
ResNet-18	11.2	1.8	×	0.681	0.500	0.667	0.880	0.537
	11.2	1.8	√	0.705	0.517	0.682	0.951	0.531
ResNet-50	23.5	4.1	×	0.658	0.451	0.622	0.909	0.473
	23.5	4.1	√	0.720	0.541	0.702	0.960	0.554
RegNetY-4G	19.6	4.0	×	0.682	0.492	0.659	0.909	0.517
	19.6	4.0	√	0.722	0.544	0.705	**0.962**	0.556
Mit-B1	13.2	2.1	×	0.839	0.776	0.874	0.819	**0.937**
	13.2	2.1	√	**0.859**	**0.788**	**0.882**	0.879	0.885

Note: M: megabyte, G: gigabyte, MSCAM: multi-scale class activation map.

in Table 4 and Fig. 13. As shown in Table 4, the proposed method significantly outperformed PCC in terms of OA, IOU, and F1. As displayed in Fig. 13, PCC introduced lots of false alarms, while the proposed method successfully detected most of the changes and at the same time, suppressed pseudo changes effectively. The main reason for the different performances was that PCC relied on the classification result of each date but ignored the temporal correlation of multi-temporal images, leading to error accumulation. By contrast, the proposed method utilized a change detection network (Fig. 9) to simultaneously extract multi-level features of multi-temporal images. In this way, the temporal correlation can be learned to be utilized for better distinguishing changed and unchanged areas.

4.3. Benefits of ZY-3 multi-view images

We constructed and publicly released the multi-view built-up area (BA) detection and change detection datasets, by using 61 ZY-3 multi-view image sets covering 48 Chinese cities. Note that there are few available semantic segmentation datasets based on multi-view satellite images, due to the high acquisition cost of multi-view images and high-quality labels. Recently, the 2019 Data Fusion Contest (DFC19) delivered a large-scale semantic 3-D dataset (Bosch et al., 2019; Le Saux et al., 2019), consisting of 69 WorldView-3 multi-view images, DSM (Digital Surface Models), and semantic labels involving water, ground, elevated roads and bridges, buildings, and high vegetation. Each WorldView-3 image was acquired at the angle optimal for stereo viewing, and thus most images have different imaging dates. However, the difference in imaging conditions for these WorldView-3 multi-view images can hinder vertical information extraction (Gong and Fritsch, 2018), and meanwhile, off-nadir multi-date images can easily cause pseudo changes on change detection. By contrast, ZY-3 satellites can simultaneously obtain multi-spectral and multi-view images, including nadir, forward, and backward viewing angles. In this context, the imaging condition of multi-view images is kept consistent, and nadir images can reduce pseudo changes caused by different viewing angles. Moreover, the joint use of multi-spectral and multi-view images can provide planar and vertical information for accurately describing BAs.

To validate the benefits of ZY-3 multi-view images, we utilized three

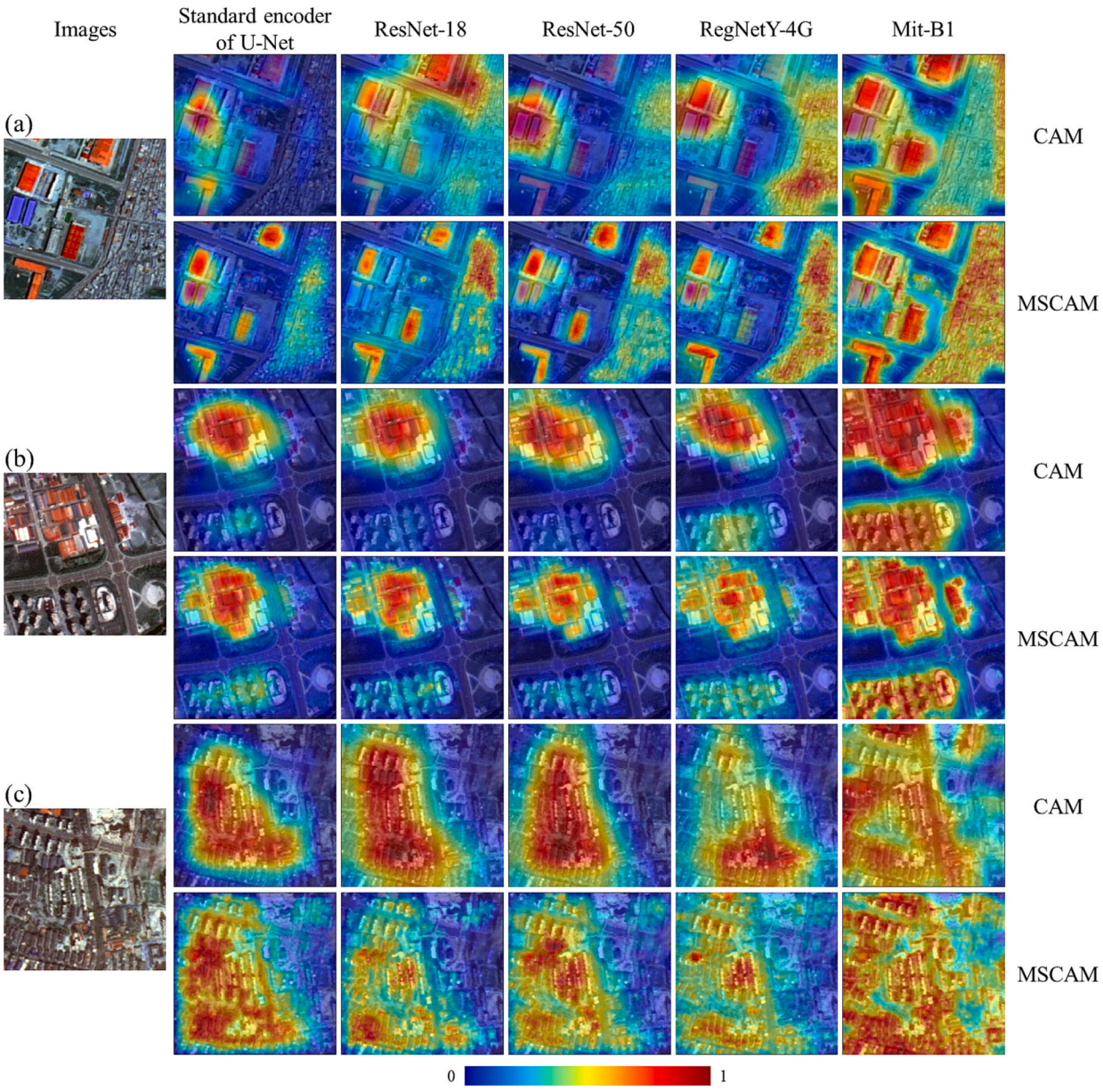

Fig. 16. Illustration of class activation map (CAM) and multi-scale CAM (MSCAM) generated by different encoders, including the standard encoder of U-Net, ResNet18, RestNet50, RegNetY-4G, and Mit-B1.

Table 8
Comparison of accuracy of built-up area change detection between the proposed method and the traditional one (Huang et al., 2020a).

	Shanghai					Beijing				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
Huang et al. (2020a)	0.858	0.284	0.442	0.347	0.612	0.855	0.249	0.399	0.42	0.379
Proposed method	**0.962**	**0.657**	**0.793**	**0.794**	**0.793**	**0.947**	**0.646**	**0.785**	**0.813**	**0.758**

types of input modes, including multi-spectral images, multi-view images, and their combination (i.e., all images), to generate the multi-scale class activation map (MCAM) and detect BAs. Experimental results are reported in Fig. 14 and Table 5. As illustrated in Fig. 14, MCAM with all images obtained the most complete coverage for BAs, compared to that with multi-spectral or multi-view images. It also can be observed that MCAM with multi-view images had a low response to low-rise BAs (e.g., Fig. 14 (e-f)), which, however, can be successfully identified by MCAM with multi-spectral images. Meanwhile, MCAM with multi-spectral images underestimated the middle to high-rise BAs (Fig. 14 (g-h)) owing to their similar colors to the background. However, this issue can be effectively dealt with by MCAM with multi-view images. The accuracy of BA detection in Table 5 shows that all images achieved the highest OA, IoU, F1, and recall values. These results indicated that multi-spectral and multi-view images can complement each other to improve the identification of BAs.

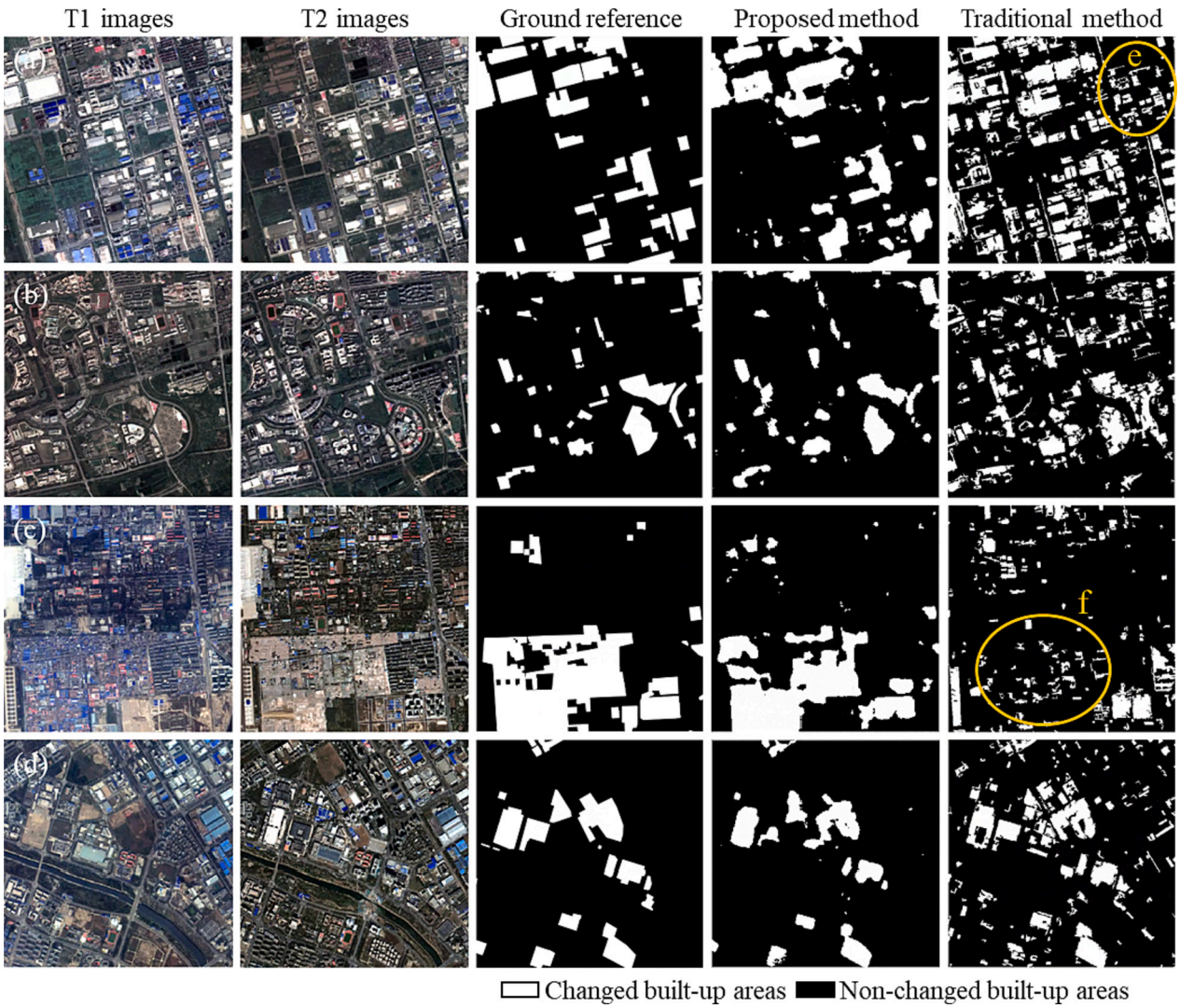

Fig. 17. Results of built-up area change detection with the proposed method and the traditional one by (Huang et al., 2020a).

Table 9
Accuracy of built-up area change detection with reliable pseudo labels at different units (i.e., pixel, object, and object+pixel).

	Shanghai					Beijing				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
Pixel	0.956	0.628	0.771	0.741	**0.804**	0.941	0.621	0.766	0.763	**0.769**
Object	**0.962**	0.646	0.785	**0.812**	0.760	**0.950**	**0.650**	**0.788**	**0.842**	0.740
Object+pixel	**0.962**	**0.657**	**0.793**	0.794	0.793	0.947	0.646	0.785	0.813	0.758

4.4. Effectiveness of the BA detection method

To evaluate the effectiveness of the proposed method on BA detection, we analyzed the effects of the two modules: 1) multi-scale CAM; and 2) adaptive online noise correction. Additionally, we discussed the effect of encoders. The 'baseline' method (Table 6) did not contain the two modules, and the other settings were kept consistent with the proposed method. Table 6 shows that both modules can effectively enhance the overall metrics (i.e., OA, IoU, and F1). Specifically, from Fig. 15, it can be seen that multi-scale CAM can enhance the spatial details of BAs (e.g., Fig. 15 (e)), while adaptive online noise correction can improve the completeness of BAs (e.g., Fig. 15 (f)). For adaptive online noise correction, the timing of correction is determined automatically. The detected timing of correction was the 140th iteration in this study. To evaluate the benefit of the adaptive timing of correction, we compared it with the fixed one.

Table 6 and Fig. 15 show that the adaptive timing of correction performed significantly better than the fixed one, and the performance of the latter decreased as the number of epochs increased. The main reason behind this phenomenon is that the fixed timing of correction can easily make the network overfit or underfit pseudo labels with noises, thus misleading the learning direction of the network and lowering its generalization ability.

To evaluate the effect of encoders on BA detection, we compared the performance of the different encoders with and without multi-scale CAM. Five encoders were considered, i.e., the standard encoder of U-Net (Ronneberger et al., 2015), ResNet-18 (He et al., 2016), ResNet-50 (He et al., 2016), RegNetY-4G (Radosavovic et al., 2020), and Mit-B1 (Xie et al., 2021). Note that the first four encoders are convolutional neural networks (CNNs), while only Mit-B1 belongs to transformer, which can capture long-distance dependencies compared with CNNs and thus was used in this study.

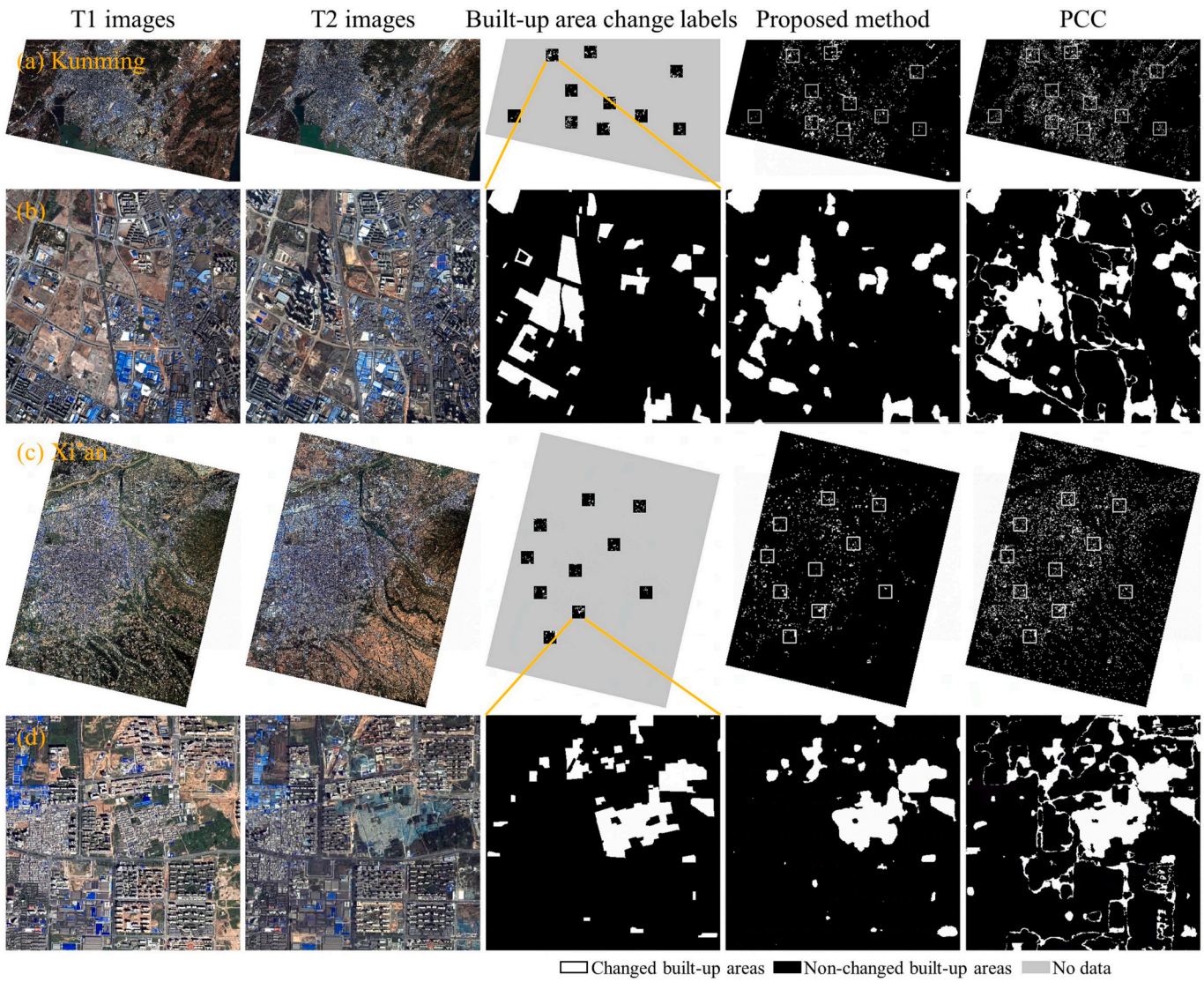

Fig. 18. Illustration of the post-classification comparison method (PCC) and the proposed one for built-up area change detection results.

Table 10
Accuracy of BA change detection using the post-classification comparison method (PCC) and the proposed one.

	Kunming					Xi'an				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
PCC	0.926	0.473	0.642	0.513	**0.857**	0.934	0.442	0.613	0.484	**0.837**
Proposed method	**0.968**	**0.645**	**0.784**	**0.812**	0.758	**0.969**	**0.582**	**0.736**	**0.799**	0.682

Table 7 shows the BA detection accuracy with different encoders. For each encoder, we computed the number of parameters (denoted as Params) and the number of floating point operations (FLOPs) to measure the computational complexity. Larger Params (or FLOPs) signify higher computational complexity. It can be seen that for all the encoders, multi-scale CAM can significantly improve the overall metrics (i.e., OA, IoU, and F1) of BA detection. Moreover, Mit-B1 obtained the highest overall metrics at low cost of Params and FLOPs, demonstrating its high computational efficiency and accuracy. By contrast, the standard encoder of U-Net performed the worst and it has large Params and FLOPs, leading to low computational efficiency.

Fig. 16 displays the class activation map (CAM) and multi-scale CAM (MSCAM) generated by different encoders. We can observe that for all encoders, MSCAM can provide more detailed boundary information while maintaining a high level of coverage of the built-up areas (BAs), compared to CAM. Furthermore, regardless of using MSCAM, Mit-B1 identified the most complete BAs, while other encoders, e.g., the standard encoder of U-Net, can miss a number of BAs, which is also reflected in its low recall value (Table 7).

4.5. Performance of the BA change detection method

In this study, we developed a deep learning-based method to identify BA changes. To illustrate the performance of the proposed method, we compared it with a traditional method and analyzed the effect of the pseudo label generation method introduced in Section 3.3. Firstly, we implemented the traditional method based on manually designed features (Huang et al., 2020a). The latter was designed to automatically

detect new built-up areas. It extracted planar and vertical features from the multi-temporal ZY-3 images to construct the time series curves of features, and then detected the extent of changes by a multi-temporal change detection model.

Table 8 reveals that the proposed method outperformed the traditional one (Huang et al., 2020a). Furthermore, Fig. 17 illustrates that the traditional method heavily suffered from false alarms (e.g., Fig. 17 (e)) and omissions (e.g., Fig. 17 (f)), while the proposed one can effectively alleviate these issues. The main reason for the different performance is that the traditional method relied on domain knowledge to manually design features, which can hardly cope with complex scenarios; by contrast, the deep learning method in this study was data-driven and was able to automatically extract representative features from complex image scenarios. It should be noted that currently, most of deep learning-based segmentation methods still require a large number of accurate pixel-level labels to optimize parameters. In this context, the proposed method was built on image-level labels, without the need for manually annotating pixel-level labels, and thereby it can lower the cost of label acquisition.

To investigate the performance of the pseudo label generation method, we analyzed the effects of three units on BA change detection, including pixel, object, and object+pixel (see Section 3.3). Table 9 demonstrates experimental results. Concerning overall metrics (i.e., OA, IoU, and F1), the object and object+pixel achieved similar results and both were better than the pixel unit. In terms of precision and recall, the object+pixel struck a balance between the object and the pixel, and therefore was used in this study.

4.6. Transferability of the BA change detection method

The change detection network used pseudo labels generated by the trained BA detection network (Section 3.2), rather than true labels. For each city (e.g., Shanghai or Beijing), we separately applied the pseudo labels for optimizing the change detection network (Section 3.3), so that the network can well adapt to each city.

We have attempted to transfer the change detection network trained in Beijing to Shanghai and the network trained in Shanghai to Beijing. Table 4 records the results. We can observe that for each city, the proposed method performed the best, followed by the network trained in other cities, while the post-classification comparison method performed the worst. This phenomenon indicates that the proposed method can adapt well to each city and is still effective to some extent even when transferred to other cities.

To further investigate the transferability of the proposed method, we collected the bi-temporal ZY-3 images in two Chinese cities, i.e., Kunming and Xi'an, which are located in the south and west of China, respectively (Fig. 1). For Kunming, two ZY-3 images (18,881 × 11,413 pixels) were obtained on Feb. 9, 2013, and Jan. 17, 2016, respectively. For Xi'an, two ZY-3 images (18,715 × 22,011 pixels) were acquired on May 12, 2015, and June 27, 2017, respectively. For each city, we randomly generated 10 blocks of 1024 × 1024 pixels and manually interpreted the BA changes for testing (see Fig. 18). These patches exhibit various types of building construction and demolition, which, therefore, are challenging for accurate change detection. According to the quantitative results in Table 10, the proposed method obtained higher overall metrics (i.e., OA, IoU, and F1) than the post-classification comparison one (PCC). Visualization results in Fig. 18 show that compared to PCC, the proposed method effectively suppressed pseudo changes and successfully detected most of the changes.

5. Discussions

Deep learning-based change detection methods have been widely used, but most of them rely on a large number of high-quality pixel-level labels (Caye Daudt et al., 2018; Chen et al., 2022; Shi et al., 2022). In this paper, we introduced image-level labels (i.e., one image is tagged as one or more categories) with low acquisition cost and proposed a multi-scale weakly supervised learning method to achieve pixel-level change detection of built-up areas (BAs). The proposed method adopted the strategy of 'first detection and then change detection', and the findings can be interpreted as follows:

Given that existing studies are subject to the single scale and low resolution of CAMs (class activation maps), we proposed a multi-scale CAM by fusing shallow and deep features at different scales to describe the multi-scale property of BAs. Results in Section 4.4 verified the effectiveness of the multi-scale CAM in enhancing the overall metrics and the spatial details of BAs.

Considering that pseudo labels generated from CAM may contain noises and existing methods set the timing of noise correction empirically, we proposed an adaptive online noise correction module. This module can correct noises online using the network prediction, and automatically determine the timing of correction by applying an exponential function to fit the IoU curves of training sets. Results in Section 4.4 showed that the noise correction module can improve the integrity of BAs effectively.

For change detection, the post-classification comparison method (PCC) detects changes by directly comparing the multi-date detection results, ignoring the temporal correlation of multi-temporal images. In this context, we proposed a reliable pseudo label generation method by utilizing the probability of detected results to measure the reliability of labels. The pseudo labels were then used to train the BA change detection network. In this way, the network can automatically learn the temporal correlation from pseudo labels and effectively suppress lots of false alarms caused by direct comparison (see Section 4.2).

Considering the less availability of multi-view datasets for semantic segmentation, we will release the multi-view built-up area (BA) detection and change detection datasets. The datasets were constructed based on ZY-3 multi-spectral and multi-view satellite images, which can provide planar and vertical information to accurately describe BAs. Results in Section 4.3 indicated that the joint use of multi-spectral and multi-view images can improve the completeness of BAs.

This study focused on built-up areas (BAs) at 2-m resolution. There are some studies on impervious areas (ISAs) at 10-m resolution (Huang et al., 2022a; Sun et al., 2022). The main differences between the two studies lie in the definition and the data source. BAs refer to the areas dominated by buildings, excluding main roads, parks, and large open spaces, while ISAs are usually defined as artificial lands which prevent water penetration, including buildings, roads, and open spaces (e.g., parking lots and squares). Therefore, in terms of the definition, ISAs contain BAs. For the data source, this study focused on 2-m resolution ZY-3 images, while ISA studies mainly used 10-m resolution Sentinel-1/2 images. ZY-3 images can provide richer spatial details, e.g., single buildings, and therefore are well suitable for city-scale studies. By contrast, Sentinel-1/2 images have limited spatial details but higher temporal resolution and larger spatial coverage, and thus are widely used for studies at region to global scales.

The proposed method has limitations too. For example, image-level labels only indicate whether an object (e.g., BAs) is present in an image and do not provide any locational information. This property might lead to the omission of some objects (e.g., small changed BAs in Fig. 13 (b) and (d))and the erroneous identification of boundaries. In addition, although the ZY-3 images used in this paper can be acquired at low cost, their spatial resolution (2.5 m) is not optimal for all built-up features, which may limit the accurate boundary delineation for the changed objects and make the detection of small changed BAs challenging. Note that in the study, "large area" means "large scale". We constructed a large-scale built-up area (BA) dataset. This dataset consists of 61 ZY-3 image sets between 2014 and 2017, which cover 48 Chinese cities and exhibit diverse BAs. The proposed method can obtain higher accuracies than other ones on this dataset.

Further research is warranted in two areas:

One possible future research issue is how to obtain fine pixel-level

pseudo labels from coarse image-level labels. This paper designed a multi-scale CAM to generate pixel-level pseudo labels. However, pixel-level pseudo labels still face the problem of missing small objects and rough boundaries. One strategy to alleviate omissions is to introduce other weak labels, e.g., point, scribble, and bounding box labels. These weak labels can provide spatial location information of objects, and have lower acquisition cost than pixel-level ones (Hua et al., 2021; Xu and Ghamisi, 2022). Solutions to improve boundary accuracy include introducing boundary optimization functions (Liang et al., 2022), incorporating boundary prior knowledge (Wei and Ji, 2022), or designing boundary-aware modules (Huang et al., 2022b).

Another possible research direction is how to use multi-date detection results to improve the accuracy of change detection in the absence of real change labels. This paper proposed a reliable pseudo label generation method, by using multi-date detection results to generate reliable change pseudo labels for optimizing the change detection network. However, limited by the data availability, this paper trained a change detection network separately for each study area (e.g., Shanghai or Beijing), making it difficult to directly transfer the network to other regions. Further research should collect time-series images in more cities to establish a change detection model with better generalization performance.

6. Conclusions

This study focuses on built-up area detection and change detection from high-resolution satellite imagery for understanding urban development. The traditional post-classification comparison method detects changes by classification and temporal comparison, in which image-level labeling is an efficient alternative to pixel-level labeling for pixel-wise classification. However, existing weakly supervised segmentation methods with image-level labels faced the problems of the single scale and low resolution of class activation map. In addition, pixel-level pseudo labels generated by CAM usually contained noise, but existing correction methods were inefficient and the timing of correction is difficult to determine. For temporal comparison, existing methods usually used pseudo labels to train change detection networks to mine the temporal correlation of multi-temporal images, but these labels contained unreliable regions, lowering the performance of networks. To address the above issues, this paper developed a multi-scale weakly supervised learning method by using the image-level labels to detect BA changes.

To test the proposed method, we constructed two datasets: 1) the BA detection dataset with 86,166 image-level samples (256 × 256 pixels for each sample), covering 48 major cities in China, for training; 2) the BA change detection dataset with two time-series ZY-3 images in two rapidly urbanizing areas (Beijing and Shanghai) for evaluation. Experiments showed that, compared to existing methods, the proposed method on BA detection identified more complete BAs while effectively retaining their boundaries. Meanwhile, the proposed method on BA change detection effectively detected most of changes and suppressed false alarms. Further analysis led to the following conclusions:

1) The proposed BA change detection method can effectively utilize the powerful learning ability of deep learning to alleviate false alarms and omissions that the traditional method can't.
2) In terms of BA detection, the designed multi-scale CAM can enhance the spatial details of BAs, while adaptive online noise correction can improve the integrity of BAs.
3) For reliable pseudo label generation, the object+pixel analysis unit can effectively balance false alarms and omissions.

In summary, the contributions of this study include the following aspects:

- Image-level labels are introduced as supervision for pixel-wise BA change detection, which can substantially lower the labelling cost.
- A multi-scale weakly supervised learning method is proposed for pixel-wise BA change detection, so that the multi-scale property and spatial details of BAs can be better utilized to generate more accurate pixel-level pseudo labels from the image-level ones.
- An adaptive online noise correction module is developed, so that incorrect pseudo labels can be automatically removed to obtain reliable BAs. Moreover, a reliable pseudo label generation module is proposed for BA change detection, so that temporal correlation between multi-temporal images can be incorporated to reduce pseudo changes.
- The multi-view BA detection and change detection datasets are constructed. To the best of the authors' knowledge, this is the first semantic segmentation dataset based on multi-view satellites.

CRediT authorship contribution statement

Yinxia Cao: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Xin Huang:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition. **Qihao Weng:** Writing – review & editing, Supervision, Funding acquisition.

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

I have shared the link to my data and code at the github website

Acknowledgments

The authors are grateful to the editors and anonymous reviewers for their constructive comments. The research was supported by the National Natural Science Foundation of China (under Grants 41971295 and 42271328).

Credit author statement.

References

Ahn, J., Cho, S., Kwak, S., 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2204–2213. https://doi.org/10.1109/CVPR.2019.00231.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65, 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004.

Bosch, M., Foster, K., Christie, G., Wang, S., Hager, G.D., Brown, M., 2019. Semantic stereo for incidental satellite images. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 1524–1532.

Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: Proc. - Int. Conf. Image Process. ICIP, pp. 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652.

Chan, L., Hosseini, M.S., Plataniotis, K.N., 2021. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. Int. J. Comput. Vis. 129, 361–384. https://doi.org/10.1007/s11263-020-01373-4.

Chen, H., Qi, Z., Shi, Z., 2022. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3095166.

Chen, Y., Chen, Z., Xu, G., Tian, Z., 2016. Built-up land efficiency in urban China: insights from the general land use plan (2006–2020). Habitat Int. 51, 31–38.

Dalla Mura, M., Benediktsson, J.A., Waske, B., Bruzzone, L., 2010. Morphological attribute profiles for the analysis of very high resolution images. IEEE Trans. Geosci. Remote Sens. 48, 3747–3762. https://doi.org/10.1109/TGRS.2010.2048116.

Deng, X., Huang, J., Rozelle, S., Zhang, J., Li, Z., 2015. Impact of urbanization on cultivated land changes in China. Land Use Policy 45, 1–7.

Dong, R., Fang, W., Fu, H., Gan, L., Wang, J., Gong, P., 2022. High-resolution land cover mapping through learning with noise correction. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3068280.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale arXiv Prepr. arXiv2010.11929.
Fan, J., Zhang, Z., Song, C., Tan, T., 2020. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4283–4292.
Fang, F., Zheng, D., Li, S., Liu, Y., Zeng, L., Zhang, J., Wan, B., 2022. Improved Pseudomasks Generation for Weakly Supervised Building Extraction From High-Resolution Remote Sensing Imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 1629–1642.
Fang, H., Du, P., Wang, X., 2022. A novel unsupervised binary change detection method for VHR optical remote sensing imagery over urban areas. Int. J. Appl. Earth Obs. Geoinf. 108, 102749 https://doi.org/10.1016/j.jag.2022.102749.
Gamba, P., Aldrighi, M., Stasolla, M., 2011. Robust Extraction of Urban Area Extents in HR and VHR SAR Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 4, 27–34. https://doi.org/10.1109/JSTARS.2010.2052023.
Ghosh, A., Kumar, H., Sastry, P.S., 2017. Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence.
Gong, K., Fritsch, D., 2018. Point cloud and digital surface model generation from high resolution multiple view stereo satellite imagery. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 42, 363–370. https://doi.org/10.5194/isprs-archives-XLII-2-363-2018.
Gong, M., Duan, Y., Li, H., 2020. Group self-paced learning with a time-varying regularizer for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. 58, 2481–2493. https://doi.org/10.1109/TGRS.2019.2951441.
Gong, M., Yang, H., Zhang, P., 2017. Feature learning and change feature classification based on deep learning for ternary change detection in SAR images. ISPRS J. Photogramm. Remote Sens. 129, 212–225. https://doi.org/10.1016/j.isprsjprs.2017.05.001.
Hafner, S., Ban, Y., Nascetti, A., 2022. Unsupervised domain adaptation for global urban extraction using sentinel-1 SAR and sentinel-2 MSI data. Remote Sens. Environ. 280, 113192.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
Hua, Y., Marcos, D., Mou, L., Zhu, X.X., Tuia, D., 2021. Semantic segmentation of remote sensing images with sparse annotations. IEEE Geosci. Remote Sens. Lett. 19, 1–5.
Huang, X., Cao, Y., Li, J., 2020a. An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images. Remote Sens. Environ. 244 https://doi.org/10.1016/j.rse.2020.111802.
Huang, X., Li, J., Yang, J., Zhang, Z., Li, D., Liu, X., 2021. 30 m global impervious surface area dynamics and urban expansion pattern observed by landsat satellites: from 1972 to 2019. Sci. China Earth Sci. 64, 1922–1933. https://doi.org/10.1007/s11430-020-9797-9.
Huang, X., Lu, Q., Zhang, L., 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. ISPRS J. Photogramm. Remote Sens. 90, 36–48. https://doi.org/10.1016/j.isprsjprs.2014.01.008.
Huang, X., Wang, Y., Li, J., Chang, X., Cao, Y., Xie, J., Gong, J., 2020b. High-resolution urban land-cover mapping and landscape analysis of the 42 major cities in China using ZY-3 satellite images. Sci. Bull. 65, 1039–1048. https://doi.org/10.1016/j.scib.2020.03.003.
Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. Remote Sens. Environ. 196, 56–75. https://doi.org/10.1016/j.rse.2017.05.001.
Huang, X., Yang, J., Wang, W., Liu, Z., 2022. Mapping 10 m global impervious surface area (GISA-10m) using multi-source geospatial data. Earth Syst. Sci. Data 14, 3649–3672.
Huang, X., Zhang, L., 2012. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 161–172. https://doi.org/10.1109/JSTARS.2011.2168195.
Huang, Z., Xiang, T.-Z., Chen, H.-X., Dai, H., 2022. Scribble-based boundary-aware network for weakly supervised salient object detection in remote sensing images. ISPRS J. Photogramm. Remote Sens. 191, 290–301.
Hussain, M., Chen, D., Cheng, A., Wei, H., Stanley, D., 2013. Change detection from remotely sensed images: from pixel-based to object-based approaches. ISPRS J. Photogramm. Remote Sens. 80, 91–106. https://doi.org/10.1016/j.isprsjprs.2013.03.006.
Kolesnikov, A., Lampert, C.H., 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp. 695–711. https://doi.org/10.1007/978-3-319-46493-0_42.
Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with Gaussian edge potentials. In: Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011, NIPS 2011, 24, pp. 109–117.
Laben, C.A., Brower, B.V., 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening.
Le Saux, B., Yokoya, N., Hansch, R., Brown, M., Hager, G., 2019. 2019 data fusion contest. IEEE Geosci. Remote Sens. Mag. 7, 103–105. https://doi.org/10.1109/MGRS.2019.2893783.
Li, X., Yuan, Z., Wang, Q., 2019. Unsupervised deep noise modeling for hyperspectral image change detection. Remote Sens. 11, 258. https://doi.org/10.3390/rs11030258.
Li, Z., Zhang, X., Xiao, P., Zheng, Z., 2021. On the effectiveness of weakly supervised semantic segmentation for building extraction from high-resolution remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3266–3281. https://doi.org/10.1109/JSTARS.2021.3063788.
Liang, Z., Wang, T., Zhang, X., Sun, J., Shen, J., 2022. Tree energy loss: Towards sparsely annotated semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16907–16916.
Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C., 2022. Adaptive early-learning correction for segmentation from noisy annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2606–2616.
Liu, X., Huang, Y., Xu, X., Li, Xuecao, Li, Xia, Ciais, P., Lin, P., Gong, K., Ziegler, A.D., Chen, A., Gong, P., Chen, J., Hu, G., Chen, Y., Wang, S., Wu, Q., Huang, K., Estes, L., Zeng, Z., 2020. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. Nat. Sustain. 3, 564–570. https://doi.org/10.1038/s41893-020-0521-x.
Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823–870. https://doi.org/10.1080/01431160600746456.
Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J., 2020. Normalized loss functions for deep learning with noisy labels. In: International Conference on Machine Learning. PMLR, pp. 6543–6553.
Malach, E., Shalev-Shwartz, S., 2017. Decoupling" when to update" from" how to update". Adv. Neural Inf. Process. Syst. 30.
Musakwa, W., Van Niekerk, A., 2015. Monitoring sustainable urban development using built-up area indicators: a case study of Stellenbosch, South Africa. Environ. Dev. Sustain. 17, 547–566.
Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man. Cybern. 9, 62–66.
Pathak, D., Krahenbuhl, P., Darrell, T., 2015. Constrained convolutional neural networks for weakly supervised segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1796–1804.
Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Julea, A., Kemper, T., Soille, P., Syrris, V., 2016. Operating procedure for the production of the global human settlement layer from landsat data of the epochs 1975, 1990, 2000, and 2014. Publ. Off. Eur. Union 1–62.
Pesaresi, M., Gerhardinger, A., Kayitakire, F., 2008. A robust built-up area presence index by anisotropic rotation-invariant textural measure. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 1, 180–192. https://doi.org/10.1109/JSTARS.2008.2002869.
Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., 2013. A global human settlement layer from optical HR/VHR RS data: concept and first results. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 2102–2131.
Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P., 2020. Designing network design spaces. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10425–10433. https://doi.org/10.1109/CVPR42600.2020.01044.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.
Shao, Z., Tian, Y., Shen, X., 2014. BASI: a new index to extract built-up areas from high-resolution remote sensing images by visual attention model. Remote Sens. Lett. 5, 305–314.
Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X., Tian, Q., 2023. A survey on label-efficient deep image segmentation: bridging the gap between weak supervision and dense prediction. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2023.3246102.
Shi, Q., Liu, M., Li, S., Liu, X., Wang, F., Zhang, L., 2022. A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3085870.
Singh, A., 1989. Review article digital change detection techniques using remotely-sensed data. Int. J. Remote Sens. 10, 989–1003.
Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G., 2022. Learning from noisy labels with deep neural networks: a survey. IEEE Trans. Neural Networks Learn. Syst. https://doi.org/10.1109/TNNLS.2022.3152527.
Sun, Z., Du, W., Jiang, H., Weng, Q., Guo, H., Han, Y., Xing, Q., Ma, Y., 2022. Global 10-m impervious surface area mapping: a big earth data based extraction and updating approach. Int. J. Appl. Earth Obs. Geoinf. 109, 102800.
Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y., 2018. On regularized losses for weakly-supervised CNN segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 507–522.
Tao, C., Tan, Y., Zou, Z., Tian, J., 2013. Unsupervised detection of built-up areas from multiple high-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. 10, 1300–1304.
Taubenbock, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., Dech, S., 2012. Monitoring urbanization in mega cities from space. Remote Sens. Environ. 117, 162–176. https://doi.org/10.1016/j.rse.2011.09.015.
Uhl, J.H., Leyk, S., 2020. Towards a novel backdating strategy for creating built-up land time series data using contemporary spatial constraints. Remote Sens. Environ. 238, 111197.
United Nations, 2022. World urbanization prospects: The 2022 revision. United Nations Department of Economic and Social Affairs, New York, NY, USA.

Wang, H., Gong, X., Wang, B., Deng, C., Cao, Q., 2021. Urban development analysis using built-up area maps based on multiple high-resolution satellite data. Int. J. Appl. Earth Obs. Geoinf. 103, 102500 https://doi.org/10.1016/j.jag.2021.102500.

Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M., 2022. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS J. Photogramm. Remote Sens. 190, 196–214. https://doi.org/10.1016/j.isprsjprs.2022.06.008.

Wang, S., Chen, W., Xie, S.M., Azzari, G., Lobell, D.B., 2020. Weakly supervised deep learning for segmentation of remote sensing imagery. Remote Sens. 12, 207. https://doi.org/10.3390/rs12020207.

Wei, Y., Ji, S., 2022. Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–12. https://doi.org/10.1109/TGRS.2021.3061213.

Wu, C., Du, B., Cui, X., Zhang, L., 2017. A post-classification change detection method based on iterative slow feature analysis and bayesian soft fusion. Remote Sens. Environ. 199, 241–255. https://doi.org/10.1016/j.rse.2017.07.009.

Wu, F., Wang, C., Zhang, H., Li, J., Li, L., Chen, W., Zhang, B., 2021. Built-up area mapping in China from GF-3 SAR imagery based on the framework of deep learning. Remote Sens. Environ. 262, 112515.

Xia, H., Tian, Y., Zhang, L., Li, S., 2022. A deep siamese postclassification fusion network for semantic change detection. IEEE Trans. Geosci. Remote Sens. 60, 1–16. https://doi.org/10.1109/TGRS.2022.3171067.

Xian, G., Homer, C., Fry, J., 2009. Updating the 2001 National Land Cover Database land cover classification to 2006 by using landsat imagery change detection methods. Remote Sens. Environ. 113, 1133–1147. https://doi.org/10.1016/j.rse.2009.02.004.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xu, Y., Ghamisi, P., 2022. Consistency-regularized region-growing network for semantic segmentation of urban scenes with point-level annotations. IEEE Trans. Image Process. 31, 5038–5051.

Yan, X., Shen, L., Wang, J., Wang, Y., Li, Z., Xu, Z., 2022. PANet: pixelwise affinity network for weakly supervised building extraction from high-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. 19 https://doi.org/10.1109/LGRS.2022.3205309.

Yi, K., Wu, J., 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 7010–7018. https://doi.org/10.1109/CVPR.2019.00718.

Zhang, B., Xiao, J., Wei, Y., Sun, M., Huang, K., 2020. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. In: AAAI 2020 - 34th AAAI Conf. Artif. Intell, pp. 12765–12772. https://doi.org/10.1609/aaai.v34i07.6971.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2921–2929. https://doi.org/10.1109/CVPR.2016.319.

Zhou, Y., Li, X., Asrar, G.R., Smith, S.J., Imhoff, M., 2018. A global record of annual urban dynamics (1992–2013) from nighttime lights. Remote Sens. Environ. 219, 206–220. https://doi.org/10.1016/j.rse.2018.10.015.

Zhou, Z.H., 2018. A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5, 44–53. https://doi.org/10.1093/nsr/nwx106.