Remote Sensing of Environment 284 (2023) 113371

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A full-level fused cross-task transfer learning method for building change detection using noise-robust pretrained networks on crowdsourced labels

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

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, PR China

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Building change detection
Transfer learning
Crowdsourced label
High-resolution remote sensing
Pseudo label

ABSTRACT

Accurate building change detection is crucial for understanding urban development. Although fully supervised deep learning-based methods for building change detection have made progress, they tend to fuse temporal information only at a single level (e.g., input, feature, or decision levels) to mitigate the data distribution differences between time-series images, which is highly prone to introduce a large number of pseudo changes. Moreover, these methods rely on a large number of high-quality pixel-level change labels with high acquisition costs. In contrast, available crowdsourced building data are abundant but are less considered for change detection. For example, OpenStreetMap (OSM), Google Map, and Gaode Map provide lots of available building labels, yet they usually contain noise such as false alarms, omissions, and mismatches, limiting their wide application. In addition, when the building extraction task is transferred to the building change detection task, the temporal and regional differences between different images may lead to undesired pseudo changes. Given these issues, we propose a full-level fused cross-task transfer learning method for building change detection using only crowdsourced building labels and high-resolution satellite imagery. The method consists of three steps: 1) noise-robust building extraction network pretraining; 2) uncertainty-aware pseudo label generation; and 3) full-level fused building change detection. We created building extraction and building change detection datasets. The former (building extraction dataset) contains 30 scenes of ZY-3 images covering 27 major cities in China and crowdsourced building labels from Gaode Map for training, while the latter (building change dataset) contains bi-temporal ZY-3 images in Shanghai and Beijing for testing. The results show that the proposed method can identify changed buildings more effectively and better balance false alarms and omissions, compared to the existing state-of-the-art methods. Further analysis indicates that the inclusion of samples from multiple cities with various spatial heterogeneities is helpful to improve the performance. The experiments show that it is promising to apply the proposed method to situations where true labels are completely lacking or limited, thus alleviating the issue of high label acquisition cost. The source code will be available at https://github.com/lauraset/FFCTL.

1. Introduction

As the important places for human activities, urban areas are gradually expanding, accompanied by building change (Grimm et al., 2008; Seto et al., 2012). Particularly in Asia and Africa, urban expansion is triggering the conversion of large amounts of arable land into building areas to better accommodate population growth (D'Amour et al., 2017; Liu et al., 2020). At the same time, a large number of building demolition and redevelopment projects are carried out within cities in order to make efficient use of limited land resources and promote sustainable urban development (He et al., 2020; Huang et al., 2017; Lai et al., 2021). Therefore, building change detection is vital for understanding urban development and has been effectively applied to geodatabase update (Matikainen et al., 2010), disaster assessment (Anniballe et al., 2018; Zheng et al., 2021), urban sprawl studies (Qin et al., 2015), and illegal building monitoring (Moghadam et al., 2015), among others.

Remote sensing images, with various spectral, spatial, and temporal resolutions and wide coverage, are widely used for urban-related studies, such as medium-to-low resolution images (e.g. AVHRR, DMSP/OLS, VIIRS-DNB, MODIS, Landsat, and Sentinel-1/2) (Huang

* 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.2022.113371
Received 18 July 2022; Received in revised form 8 November 2022; Accepted 17 November 2022
Available online 24 November 2022

et al., 2021; Li et al., 2018; Liu et al., 2020; Mertes et al., 2015; Stokes and Seto, 2019; Taubenböck et al., 2012; Voogt and Oke, 2003; Zhou et al., 2018). However, these images have relatively coarse spatial resolution, making it difficult to capture fine-scale ground objects within urban areas, e.g., buildings (Pesaresi et al., 2016; Zhang et al., 2022). Recently, there is a significant increase in the amount of available high-resolution satellite imagery (with a spatial resolution $\leq$5 m), such as WorldView, PlanetScope, ZY-3, GF-1/2, and TerraSAR-X, and their rich spatial details can support the change detection at the building scale (Gamba et al., 2011; Huang et al., 2020; Marin et al., 2015). Therefore, this study aims to use high-resolution optical satellite imagery to monitor urban building changes.

Change detection methods, according to whether training samples are used, can be classified as unsupervised and supervised ones. Unsupervised methods do not rely on training samples and are easy to implement, such as image differencing (Bruzzone and Prieto, 2000), change vector analysis (CVA) (Bovolo and Bruzzone, 2006), principal component analysis and k-means clustering (PCA-kmeans) (Celik, 2009), and multivariate alteration detection (MAD) (Nielsen et al., 1998). Note that these methods focus on changed and unchanged categories, rather than specific ground objects (e.g., buildings). A large number of studies have examined building change detection methods (Leichtle et al., 2017; Tang et al., 2013; Tian et al., 2014; Wang and Li, 2020; Zhang et al., 2017). For instance, Huang et al. (2014) calculated the difference value of the bi-temporal morphological building index (MBI) (Huang and Zhang, 2012) to indicate the intensity of building change, and then combined it with spectral and shape conditions to obtain changed buildings by thresholding. For building change detection, the object-based analysis is popular (Chen et al., 2012). For instance, Leichtle et al. (2017) used the building footprint as the analysis unit to extract object-based spectral and textural features of bi-temporal images, and then employed principal component analysis and k-means clustering on these features to identify building changes. However, these unsupervised methods depend on handcrafted features, leading to the insufficient use of the prior knowledge of images, and they usually need to manually select the optimal threshold in the decision phase.

Supervised methods, compared to unsupervised ones, can perform change detection by leveraging samples to train classifiers, e.g., random forest (RF) and support vector machine (SVM) (Huang et al., 2017). However, traditional supervised methods require domain-specific knowledge to manually design features, e.g., spectral, shape, and textural features (Dalla Mura et al., 2010; Pacifici et al., 2009). In recent years, deep learning, especially convolutional neural networks (CNNs), has been widely used for building change detection (Chen et al., 2022a; Fang et al., 2022b; Peng et al., 2019; Zhang et al., 2020a), since it can automatically learn discriminative and representative features from images. Building change detection usually uses the encoder-decoder structure, e.g., U-Net (Weng and Zhu, 2021), SegNet (Badrinarayanan et al., 2017), and PSPNet (Zhao et al., 2017). This structure encodes the input image to extract deep features, and then uses a decoder to recover the feature size and output the predicted result of the same size as the input image. In this case, each pixel on the input image is labeled as a changed or unchanged building. Due to the high cost of labeling, semi-supervised learning has been adopted for building change detection (Peng et al., 2021; Sun et al., 2022a). Under the supervision of a few labeled temporal images, semi-supervised learning for change detection can introduce unsupervised regularization on a large number of unlabeled temporal images to improve the generalization performance of models (Zhang et al., 2018). In this way, semi-supervised learning can lower the dependence on pixel-level change labels.

Given that change detection usually focuses on bi-temporal images, according to the bi-temporal image fusion stage, it can be divided into the input-level, feature-level, and decision-level fusion (Caye Daudt et al., 2018; Shi et al., 2020a):

1) In the input-level fusion, bi-temporal images are stacked as a single image along the channel dimension before being passed into the network. For instance, Sun et al. (2020) first stacked bi-temporal images for feature extraction, and then adopted building extraction and change detection decoders to acquire buildings for each temporal image and building changes, respectively.
2) In the feature-level fusion, bi-temporal images are separately processed by two identical encoders with shared weights (i.e., the Siamese structure) to obtain bi-temporal features, and then these features are fused by differencing or stacking for detecting changes by the decoder. For instance, Liu et al. (2021) designed a dual-task constrained deep Siamese network to extract bi-temporal features and then applied the feature difference for predicting buildings for each temporal image and building changes. Chen et al. (2022a, 2022b) proposed a transformer module to model the spatio-temporal context of bi-temporal features, and incorporated this module into the feature difference-based Siamese network for change detection.
3) The decision-level fusion refers to the post-classification comparison strategy, where each image is classified and then the classification results of all images are directly compared to obtain changes (Aguirre-Gutiérrez et al., 2012; Ye et al., 2016). Although the decision-level fusion approach can obtain the change trajectories, it relies on the classification result of each image, which may suffer from error accumulation. Thus, in deep learning-based building change detection methods, the classification result of each image is usually taken as an auxiliary task to optimize the change detection task (Liu et al., 2021; Sun et al., 2022b).

Note that these methods require lots of high-quality pixel-level change labels for mitigating the data distribution differences between time-series images at the input, feature, or decision levels, in order to highlight true changes while suppressing pseudo changes. However, due to the difference in imaging conditions, e.g., atmosphere, illumination, and viewing angle variations, color inconsistency still exists in the time-series images. Moreover, fusing temporal information at a single level (e.g., input, feature, or decision levels) may cause lots of pseudo changes. Meanwhile, although available images are abundant, it is time-consuming and labor-intensive to collect high-quality pixel-level change labels, which limits the application of these methods.

As aforementioned, building change labels are typically expensive to collect. In contrast, available crowdsourced building data are abundant but are less considered for change detection. Commonly-used building data, such as the Massachusetts building dataset (Mnih, 2013), the Inria aerial image labeling dataset (Maggiori et al., 2017a), the WHU building dataset (Ji et al., 2019), the ISPRS Vaihingen and Potsdam datasets (Rottensteiner et al., 2014), and the SpaceNet challenge dataset (Van Etten et al., 2018), have significantly contributed to the development of deep learning-based building extraction methods (Hosseinpour et al., 2022; Li et al., 2020b; Shi et al., 2020b; Zhu et al., 2021). However, these building data rely on the region-, time-, and sensor-specific images, which makes it difficult to transfer them to other images. Fortunately, crowdsourced data, e.g., OpenStreetMap (OSM) and public maps (e.g., Google Map, Gaode Map, and Baidu Map), provide a large number of available building labels with a great potential for building extraction. For example, Kaiser et al. (2017) applied the OSM building labels to train a building detection network and obtained satisfactory results. However, these crowdsourced building labels are usually not exactly matched with the images in space and time, resulting in noise such as false alarms, omissions, and mismatches (Mnih and Hinton, 2012). These noisy labels can significantly reduce the generalization performance of the network (Zhang et al., 2021). Given the high cost of manually correcting noisy labels, researchers have proposed a range of automatic noisy label learning methods, such as transfer learning (Maggiori et al., 2017b) and noise-robust models (Ahmed et al., 2021; Mnih and Hinton, 2012; Zhang et al., 2020b). For instance, Maggiori et al. (2017b) first trained a building detection network with a large number of imperfect building labels from OSM, and then fine-tuned the network with a small number of accurate labels to mitigate the interference of noisy labels. Zhang et al. (2020b) designed a noisy label adaptive layer at the end of the conventional building extraction network, in order to model the probabilistic propagation relationship between noisy and true labels. Notably, these approaches still rely on a number of accurate labels or a specific network structure.

By courtesy of building data, building change detection can be performed by the post-classification comparison method, which allows for the simple transferring of the building extraction task to the building change detection task. This method first trains the building extraction network with building data to obtain buildings for each temporal image, and then generates changed buildings by direct comparison. However, this method relies on the building extraction result of each temporal image, which may suffer from error accumulation and consequently cause a large number of pseudo changes. This kind of error mainly comes from temporal and regional differences. On the one hand, in contrast to the building extraction task, the building change detection task focuses more on building changes on different time-series images. However, the time-series images usually have different imaging conditions, such as atmosphere and solar illumination variations. On the other hand, buildings in the study areas for both tasks may exhibit different spectral, contextual, and shape characteristics. A simple solution to mitigating these differences is to fine-tune the network using a small number of true labels from the study area (Maggiori et al., 2017b). However, the cost of collecting true labels is relatively high. In this context, some researchers adopt automatically generated change pseudo labels to replace true labels, and have successfully carried out change detection (Fang et al., 2022a; Gong et al., 2017, 2020; Tang et al., 2022). For example, Gong et al. (2017) employed a sparse autoencoder to learn temporal features, followed by fuzzy c-means (FCM) clustering to generate change pseudo labels, and finally used these labels to train a change detection network. Recently, Fang et al. (2022a) selected the consistent results of CVA and post-classification comparison methods as reliable pseudo labels, and then applied these labels to train a lightweight change detection network. However, these methods focus on binary change types (i.e., changed and unchanged types), rather than the change of specific ground objects (e.g., buildings). Compared with vegetation and bare soil, buildings exhibit diverse colors and shapes and are not affected by phenological conditions. Thus, it is essential to design pseudo label generation methods specifically for building change detection, in order to mitigate the temporal and regional differences in cross-task transfer learning.

In summary, although existing deep learning-based methods for building change detection have made progress, they still have the following limitations:

1) Existing approaches usually rely on a large number of high-quality pixel-level labels with high acquisition costs.
2) Existing approaches generally fuse temporal information only at a single level (e.g., input, feature, or decision levels) to mitigate the data distribution differences between time-series images, which may introduce lots of pseudo changes.
3) Available open-source or crowdsourced building data are abundant but are less considered for change detection. Meanwhile, although these data provide a large number of available building labels, they contain lots of noise, e.g., false alarms, omissions, and mismatches, which can significantly reduce the generalization performance of the network.
4) When the building extraction task is transferred to the building change detection task, the temporal and regional differences between different images may easily cause a large number of pseudo changes.

Given these issues, we propose a full-level fused cross-task transfer learning method for building change detection using only crowdsourced building labels and high-resolution satellite imagery. The method consists of three steps: 1) we first train a noise-robust building extraction network with crowdsourced building labels and high-resolution satellite imagery; 2) then, we apply the well-trained building extraction network to predict the building map for each temporal image, and design the uncertainty-aware analysis to obtain reliable change pseudo labels; 3) finally, we train a full-level fused building change detection network with the reliable change pseudo labels. The main contributions of this paper are summarized below:

- Crowdsourced building labels from 27 Chinese cities are used for building change detection to reduce the high acquisition cost of pixel-level labels.
- A noise-robust building extraction network is proposed to correct noisy labels, which can improve the generalization performance of the network across multiple cities.
- An uncertainty-aware pseudo label generation method is designed to mitigate the temporal and regional differences in cross-task transfer learning.
- A full-level fused building change detection network is developed to simultaneously reduce the data distribution differences between time-series images at the input, feature, and decision levels.

The rest of this paper is organized as follows. Section 2 introduces the experimental datasets. Section 3 describes the proposed method. Subsequently, the experimental results and discussions are given in Sections 4 and 5, respectively. Finally, we conclude this paper in Section 6.

2. Dataset description

2.1. Building detection dataset

To train the building extraction network, we created a building detection dataset. This dataset contains 30 scenes of ZY-3 images and the corresponding crowdsourced building labels (Table 1 and Fig. 1). It covers the 27 major cities in China and contains diverse buildings with different heights, shapes, sizes, and colors. ZY-3 images with less than 10% cloud coverage and imaging time between 2014 and 2017 were collected from the Land Satellite Remote Sensing Application Center (LASAC) of China (http://www.cresda.com/). Note that the ZY-3 satellite is capable of providing both multispectral images (with blue, green, red, and near-infrared bands) and multi-view images (with nadir, $+22^\circ$ forward, and -22° backward viewing angles) of the same area (Huang et al., 2017). We only collected multispectral images (with a spatial resolution of 5.8 m) and nadir-view images (2.1 m). All the images were preprocessed by radiometric correction and orthorectification (Liu et al., 2019), and were resampled to 2.5 m. Subsequently, image-to-image registration was applied to the nadir and multispectral images, and then they were fused by the Gram-Schmidt pan-sharpening algorithm (Laben and Brower, 2000) to enhance the spatial details of the multispectral images. Then, all the images were stretched to the range of [0, 255] (i.e., 8 bit) with the optimized linear stretch algorithm in ENVI software. Finally, we obtained 30 scenes of multi-spectral images at a spatial resolution of 2.5 m.

Crowdsourced building vector labels were collected by manual interpretation and were publicly released by Gaode Map (https://ditu.amap.com/). However, these labels and the corresponding ZY-3 images may be not consistent in space and time, leading to lots of noise, e.g., false alarms, omissions, and mismatches. To reduce the mismatch noise, we spatially registered each ZY-3 image and the corresponding crowdsourced building vector labels by the spatial adjustment tool in the ArcGIS software. Then, we converted the matched vector labels into the raster ones with the same resolution as ZY-3 images, i.e., 2.5 m (Fig. 1). Note that with these pre-processing steps, crowdsourced building labels still contain noise that needs a lot of time and labor to correct. To alleviate this issue, we propose an automatic noise correction method to

Table 1
Composition of the building detection dataset. The total number of samples is 34,616, each of which contains an image patch (256 × 256 pixels) and the corresponding building labels. All cities were classified into four geographic regions, i.e., east, west, south, and north.

City	#Sample	Imaging date	Region	City	#Sample	Imaging date	Region
Changzhou	1619	20,140,406	East	Yinchuan	666	20,170,706	West
Hefei	1717	20,160,828	East	Yulin	247	20,140,903	West
Jinan	453	20,170,824	East	Changsha	1287	20,171,029	South
Jinan	1529	20,170,530	East	Foshan	1433	20,150,414	South
Nanjing	1196	20,140,520	East	Guangzhou	1604	20,150,414	South
Wuxi	352	20,170,429	East	Haikou	380	20,160,506	South
Yantai	718	20,150,410	East	Huizhou	454	20,170,216	South
Chengdu	1783	20,170,513	West	Nanning	1024	20,150,413	South
Chengdu	2246	20,170,508	West	Zhengzhou	546	20,160,604	South
Kunming	1104	20,170,406	West	Beijing	3518	20,170,515	North
Lhasa	183	20,140,603	West	Harbin	404	20,170,529	North
Lanzhou	341	20,170,809	West	Shijiazhuang	1124	20,170,816	North
Urumqi	824	20,160,603	West	Shijiazhuang	1529	20,171,217	North
Xi'an	2749	20,150,512	West	Tianjin	2034	20,150,414	North
Xining	563	20,160,726	West	Taiyuan	989	20,150,423	North

train a noise-robust building extraction network (see Section 3.1). To train this network, we cropped all ZY-3 images and the crowdsourced building labels into patches of 256 × 256 pixels without overlapping, and obtained 34,616 samples, each of which contains an image patch and the corresponding building labels (Table 1). We randomly selected 90% of the samples for training and the remaining 10% for validation.

2.2. Building change detection dataset

To evaluate the proposed change detection method, we created the building change detection dataset in Shanghai and Beijing. Each region contains bi-temporal ZY-3 images (Fig. 2). The bi-temporal ZY-3 images of Shanghai were acquired on September 18, 2012 and September 30, 2018, while those of Beijing were acquired on October 11, 2012 and October 6, 2018. These images were preprocessed in the same way as mentioned in Section 2.1. Through image clipping, we obtained the ZY-3 images of Shanghai and Beijing with dimensions of 8681 × 10,965 pixels and 9916 × 11,122 pixels, respectively. Finally, we used the pseudo-invariant features method (Schott et al., 1988) to reduce the radiometric difference between bi-temporal images.

To test the accuracy of change detection, we manually interpreted the changed buildings by randomly selecting 10 sample patches (1024 × 1024 pixels for each patch) in each region (Fig. 2), aided by the Google Earth high-resolution historical images and ZY-3 time-series images. These sample patches are randomly distributed, and cover residential, commercial, and industrial areas. They mainly contain newly-built and demolished buildings, which is suitable for evaluating the generalization performance of the proposed method. It should be noted that all of these building change labels are used for testing the accuracy of the algorithms. To train the change detection network, we cropped the ZY-3 images of each region into patches of 256 × 256 pixels, and obtained 3575 samples (in patches) in Shanghai and 6300 samples (in patches) in Beijing after excluding the test sample patches. Note that the training set for the change detection is totally based on the pseudo labels that are generated from the crowdsourced building data (see the methodology section for details). This approach greatly alleviates the problem of the high label acquisition cost.

3. Methodology

The proposed building change detection method consists of three steps: 1) noise-robust building extraction network pretraining (Section 3.1); 2) uncertainty-aware pseudo label generation (Section 3.2); and 3) full-level fused building change detection (Section 3.3). The workflow is presented in Fig. 3. Details of each step are given below.

3.1. Noise-robust building extraction network pretraining

We proposed a noise-robust building extraction method (i.e., step 1 in Fig. 3) that was pretrained on crowdsourced building labels from the 27 major cities in China, to mitigate the interference of label noise. The method includes three steps: 1) network initialization; 2) noisy label correction; and 3) network retraining.

Firstly, we initialized the building extraction network using the original crowdsourced building labels (Section 2.1). The structure of the building extraction network is displayed in Fig. 4. Particularly, the encoder was set to the standard residual neural network, ResNet-50 (He et al., 2016), considering its powerful feature extraction capability and wide range of applications. Notice that deep networks tend to learn correctly labeled samples first in the early stage and start to learn mislabeled samples later (Arazo et al., 2019). Therefore, after initialization, the building extraction network already has building feature extraction capability. Secondly, to further optimize the network parameters while avoiding the interference of noisy labels, we used the initialized network as a starting point for subsequent building prediction and noisy label correction. In detail, for the current epoch (t), the supervision of the network comes from the original labels and the corrected ones that are the network predictions of the previous epoch (t-1). Finally, we obtained clean labels for retraining the building extraction network from scratch.

The loss function of the network training is formulated as follows:

$$L_{noise} = \lambda \cdot L_{initial}(q, p) + \beta \cdot L_{update}(\widehat{q}, p) \tag{1}$$

where q denotes the original building label, $\widehat{q}$ is the corrected building label, and p is the probability value of the network prediction. The loss functions $L_{initial}$ and L_{update} are both the sum of the binary cross-entropy loss and the dice coefficient loss to alleviate the class imbalance problem (Peng et al., 2019). The parameters λ and β are the loss weights of the original and corrected labels, respectively. In the network initialization phase, we used only the original crowdsourced building labels, i.e., $\lambda = 1$ and $\beta = 0$. Note that the parameters of the network were initiated by the pretrained weights from ImageNet (Jia Deng et al., 2009). To avoid the network fitting noise, the learning rate was fixed at 0.001 and the total number of epochs for training was set to 20. In the noisy label correction phase, we considered both the original and corrected labels. To reduce the interference of the noise introduced by the original labels, we selected the parameters $\lambda = 0.2$ and $\beta = 1$. To avoid overfitting, the number of epochs was set to 10 and the learning rate was fixed at 0.001. Finally, we retrained the network with only the corrected labels, i.e., $\lambda = 0$ and $\beta = 1$. To fully optimize the network, the learning rate was initially set to 0.001, with a drop of 0.1 every 20 epochs, and the total number of epochs for training was set to 50. For all phases, the optimizer was set to Adam (Kingma and Ba, 2014). The batch size, i.e., the number of image patches for training in each iteration, was set to 32, due to the GPU

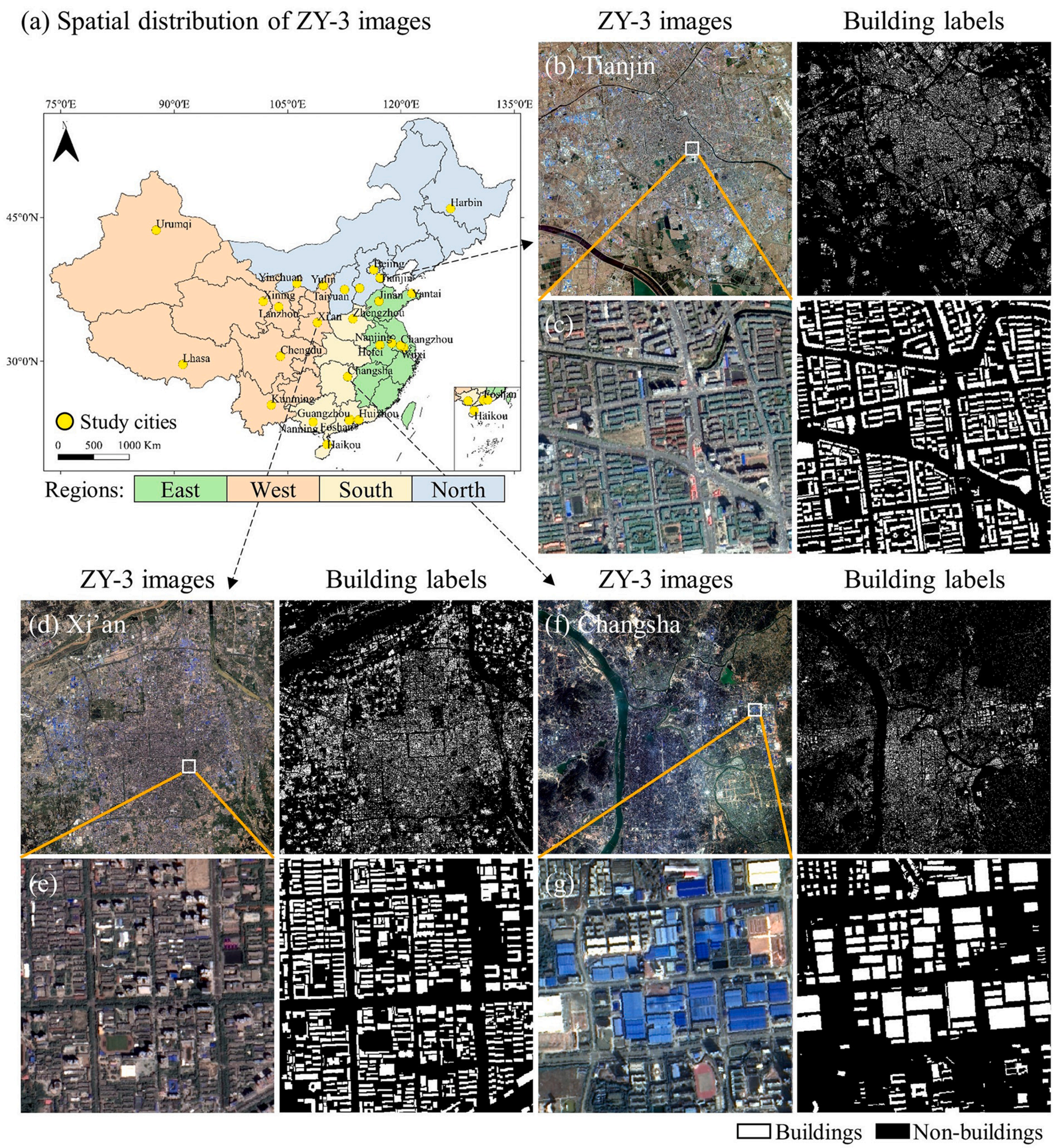

Fig. 1. The building detection dataset. (a) Spatial distribution of ZY-3 images. Each region from Tianjin (b), Xi'an (d), and Changsha (f), includes the ZY-3 images (30 km × 30 km) and the corresponding crowdsourced building labels. Each enlarged view (c, e, and g) has a spatial extent of 1.28 km × 1.28 km (i.e., 512 × 512 pixels).

memory limitation (11G for a single GeForce GTX 1080ti). In this study, the data augmentation includes random rotation, randomly horizontal and vertical flipping, and random grid shuffle, to improve the generalization performance of the network. For the network training, each band of images was linearly stretched to the range of [0,1] by the minimum and maximum values of each band. The validity of noisy label correction is discussed in Section 5.1.

3.2. Uncertainty-aware pseudo label generation

We designed an uncertainty-aware pseudo label generation algorithm (i.e., step 2 in Fig. 3) to alleviate the temporal and regional differences in cross-task transfer learning. The algorithm includes three steps: 1) single-temporal building prediction; 2) object-to-pixel multi-temporal comparison; and 3) uncertainty-aware analysis for reliable pseudo label generation.

Step 1. Single-temporal building prediction. We applied the well-trained building extraction network that was pretrained from the 27 major cities of China in Section 3.1 to generate the building probability map (with a data range of [0,1]) for each temporal image. According to the existing literature (Ji et al., 2019), each pixel with a building probability greater than 0.5 is assigned as building, otherwise it is non-building.

Step 2. Object-to-pixel multi-temporal comparison. Firstly, we selected the building objects of each temporal image as the analysis unit for extracting the change objects. Specifically, for each building object (O1) at one time, we searched for the building object (O2) at another time that spatially overlaps with O1. If the overlap degree (i.e., the ratio

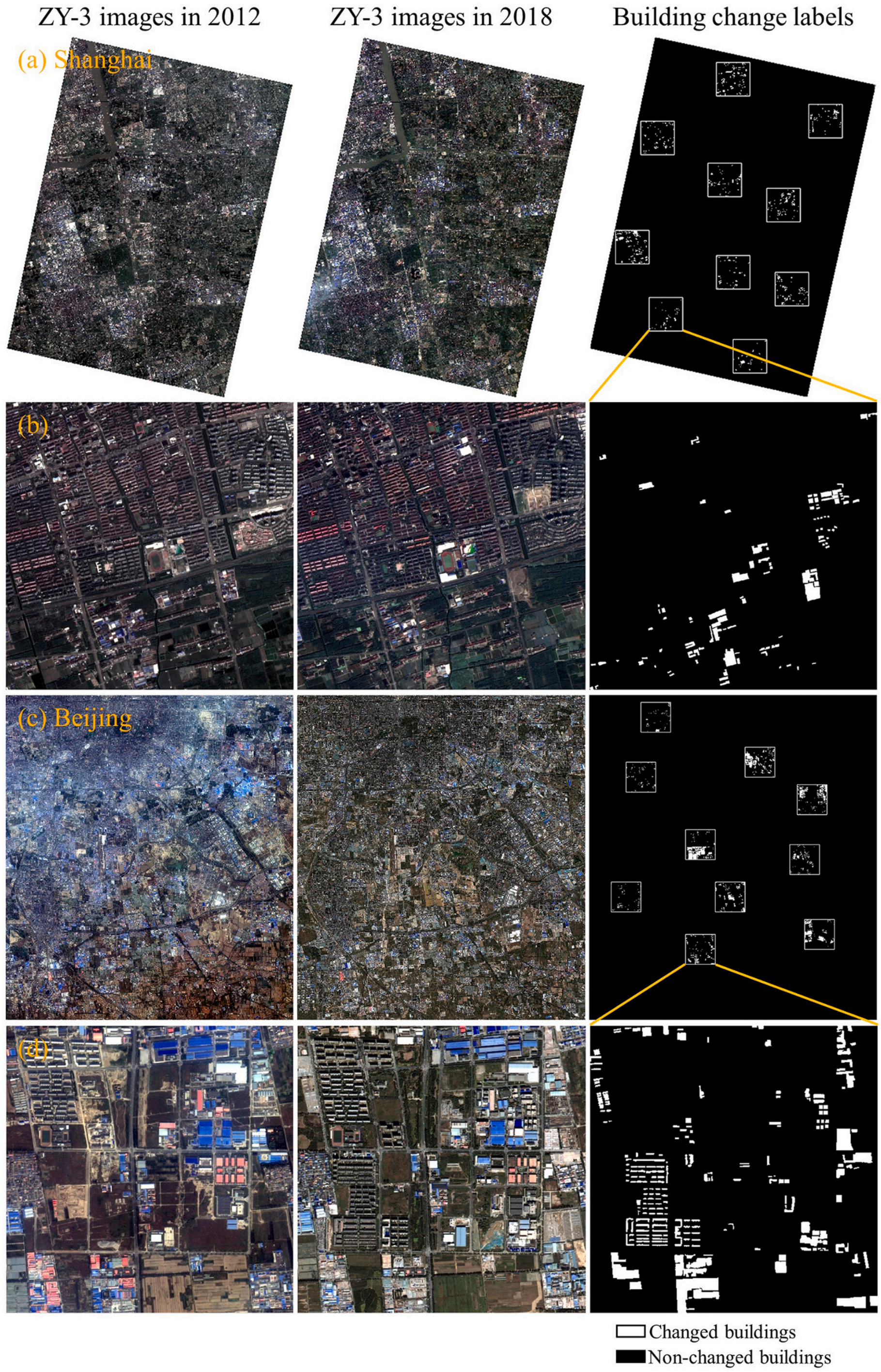

Fig. 2. The building change detection dataset, including bi-temporal ZY-3 images and the corresponding building change labels in Shanghai (a) and Beijing (c). The spatial extent of the ZY-3 image is 8681 × 10,965 pixels for Shanghai (a) and 9916 × 11,122 pixels for Beijing (c). Each enlarged view (b and d) has a spatial extent of 1024 × 1024 pixels.

of intersection and union) between O1 and O2 is greater than a threshold (0.5 in this study), the two objects are classified as "non-change", and as "change" otherwise. Subsequently, for changed objects (e.g., O1 and O2), we calculated the pixel-level change region, i.e., the non-overlapping region of the two objects. Taking Fig. 5 as an example, a and d are the same building and are considered unchanged due to their high overlap degree; c and f have not any spatial overlapping objects and thereby are assigned as changed buildings; b and e are spatial counterparts with low overlap degree, so the pixel-based comparison is further used to obtain the changed building region g. Note that object-based comparison can suppress the residual matching errors and the imaging condition difference between bi-temporal images, while the further pixel-based comparison of the changed objects can detect the specific change region. An in-depth analysis of the object-to-pixel comparison is presented in Section 5.2. Note that we did not use the objects from image segmentation as the analysis unit. The reason is that for the image segmentation techniques (Blaschke, 2010), the scale is a key parameter that influences the segmentation performance and usually needs to be set manually. By contrast, we directly obtained objects from the well-trained building extraction network, which has the capability of

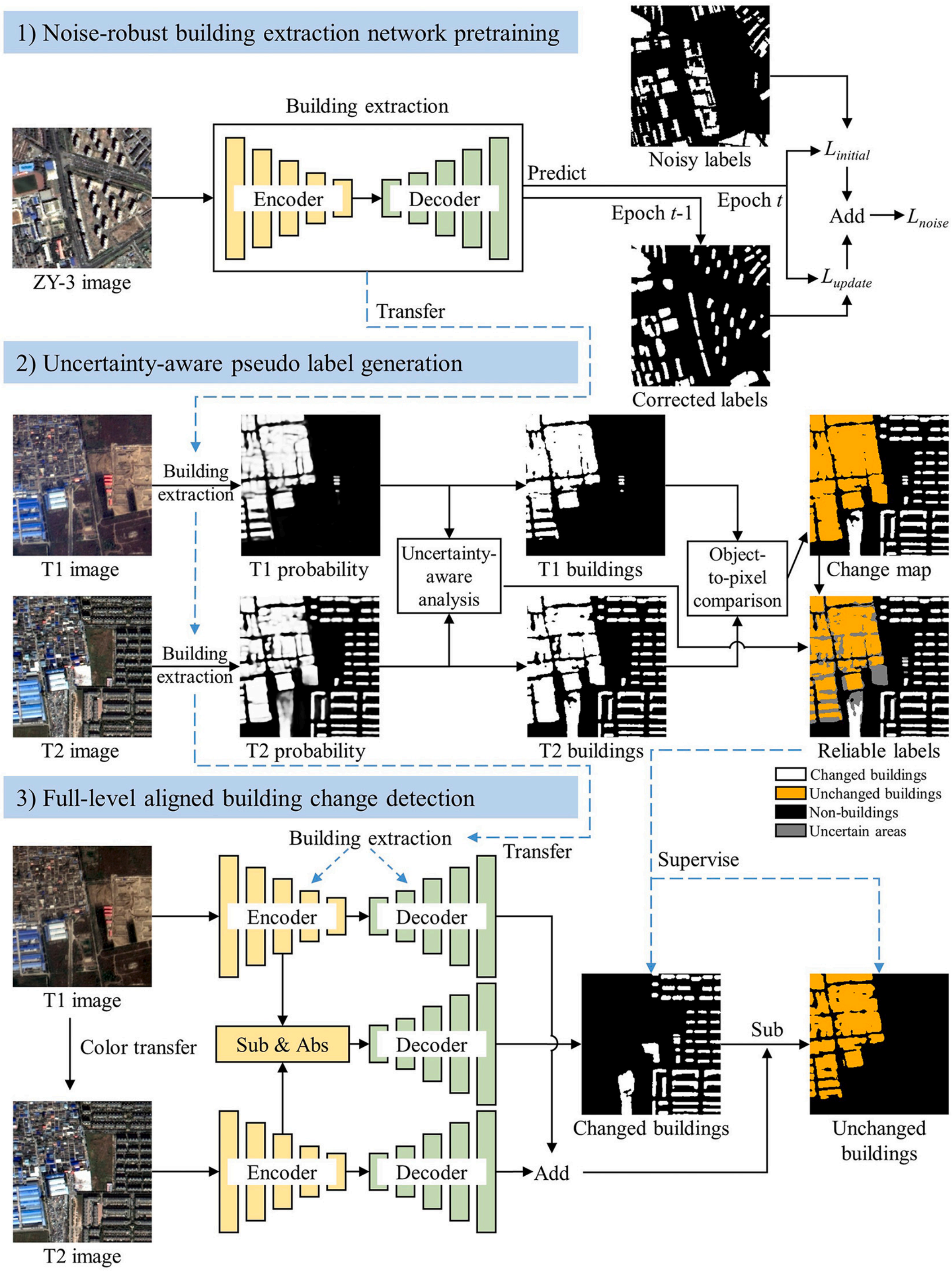

Fig. 3. The workflow of the proposed method. "Sub & Abs" in Step 3 means the absolute difference of bi-temporal features.

identifying buildings without the need of selecting the scale parameter.

Step 3. Uncertainty-aware analysis for reliable pseudo label generation. Based on the pseudo labels generated by the object-to-pixel comparison (Step 2), we further obtained reliable pseudo labels by the uncertainty-aware analysis. In detail, we calculated the absolute difference of bi-temporal building probability maps (see Step 1) as the building change probability (P) with a data range of [0,1]. A larger P indicates a higher change probability. We set the change probability threshold to 0.7 to filter out the uncertain region (i.e., 1-T < P < T) and retain the more reliable region (i.e., P ≥ T or $P \leq$ 1-T) for subsequent change detection. The sensitivity of the change probability threshold is analyzed in Section 5.2.

3.3. Full-level fused building change detection

We developed a full-level fused building change detection network (i.e., step3 in Fig. 3) to simultaneously mitigate the data distribution differences between time-series images at the input, feature, and decision levels. Without the need for manually labeled change samples, the network can rely merely on the automatically generated reliable pseudo labels (Section 3.2) built on the building extraction network (Section 3.1), which was fully pretrained on crowdsourced building labels covering the 27 major cities of China. The network consists of three parts: 1) color transfer at the input level; 2) layer-wise temporal difference at the feature level; and 3) simultaneous extraction of changed and

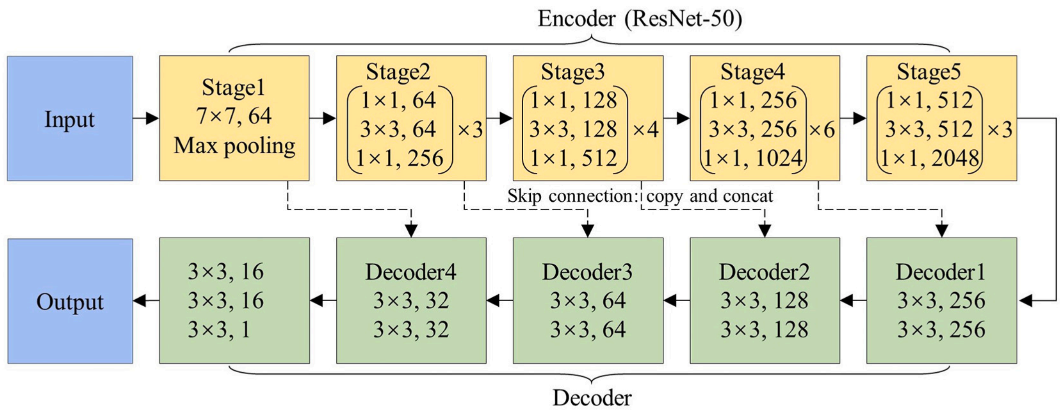

Fig. 4. The structure of the building extraction network. The sign "$k \times k$, c" represents the convolution layer with the kernel size of k and the number of channels of c.

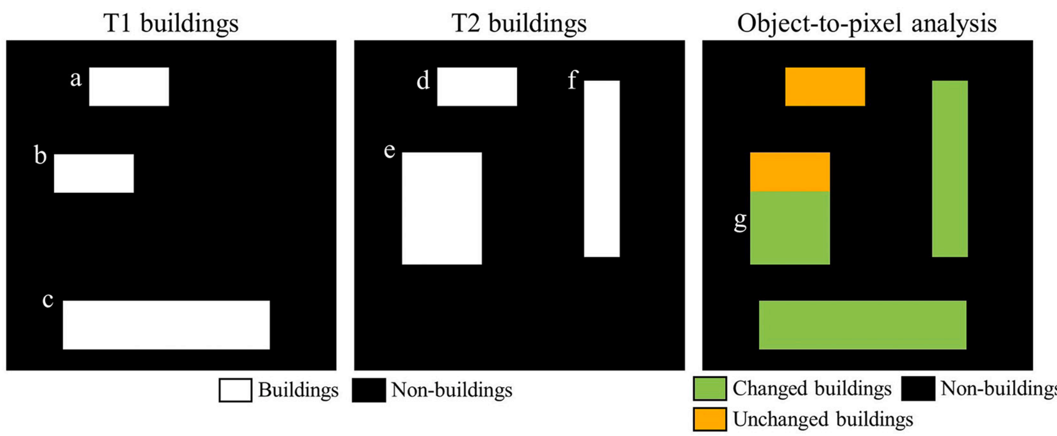

Fig. 5. Illustration of the object-to-pixel analysis.

unchanged buildings at the decision level.

Part 1. Color transfer at the input level. We applied the color transfer method (Xiao and Ma, 2006) to migrate the color space of one temporal image to another in order to reduce the difference in imaging conditions (e.g., atmosphere and illumination) between time-series images. This method is originally applied to natural images with red (R), green (G), and blue (B) bands, while in this study, we applied it to high-resolution satellite images with RGB and near-infrared (NIR) bands. Specifically, we first calculated the mean value of each band for both the source image (I_s) and the target image (I_t). We used the symbols $\overline{R_i}$, $\overline{G_i}$, $\overline{B_i}$, $\overline{N_i}$ to denote the mean values of the R, G, B, and NIR bands, respectively, with $i \in \{I_s, I_t\}$. Next, we calculated the covariance matrix (Cov_i) between all the four bands for both the source and target images, and then performed singular value decomposition (SVD) on Cov_i as follows:

$$Cov_i = U_i \cdot \Lambda_i \cdot V_i^T, i \in \{I_s, I_t\} \tag{2}$$

where U_i and V_i denote orthogonal matrices. Λ_i is a diagonal array with diagonal elements being the eigenvalues of Cov_i, i.e., λ_i^R, λ_i^G, λ_i^B and λ_i^N in sequence. Based on the mean values and the covariance matrices, we defined the rotation R_i, translation T_i, and scaling S_i parameters as follows:

$$R_i = \begin{cases} U_i, & i = I_s \\ U_i^{-1}, & i = I_t \end{cases} \tag{3}$$

$$T_i = \begin{pmatrix} 1 & 0 & 0 & 0 & t_i^R \\ 0 & 1 & 0 & 0 & t_i^G \\ 0 & 0 & 1 & 0 & t_i^B \\ 0 & 0 & 0 & 1 & t_i^N \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, t_i^j = \begin{cases} \overline{j_i}, & i = I_s \\ -\overline{j_i}, & i = I_t \end{cases}, j \in \{R, G, B, N\} \tag{4}$$

$$S_i = \begin{pmatrix} s_i^R & 0 & 0 & 0 & 0 \\ 0 & s_i^G & 0 & 0 & 0 \\ 0 & 0 & s_i^B & 0 & 0 \\ 0 & 0 & 0 & s_i^N & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, s_i^j = \begin{cases} \lambda_i^j, & i = I_s \\ 1 \Big/ \sqrt{\lambda_i^j}, & i = I_t \end{cases}, j \in \{R, G, B, N\} \tag{5}$$

Finally, we transformed the color space of the target image to that of the source image using the following equation:

$$\widehat{I_t} = T_{I_s} \cdot R_{I_s} \cdot S_{I_s} \cdot S_{I_t} \cdot R_{I_t} \cdot T_{I_t} \cdot I_t \tag{6}$$

For the color transfer between bi-temporal images, each temporal image can be used as the source or target image. An illustration of the color transfer is presented in Fig. 6. In this study, we chose the pre-temporal image (T1) as the target and the post-temporal image (T2) as the source, i.e., pre-temporal to post-temporal transfer (T1 to T2). In Section 5.3, we discussed the sensitivity of the color transfer direction,

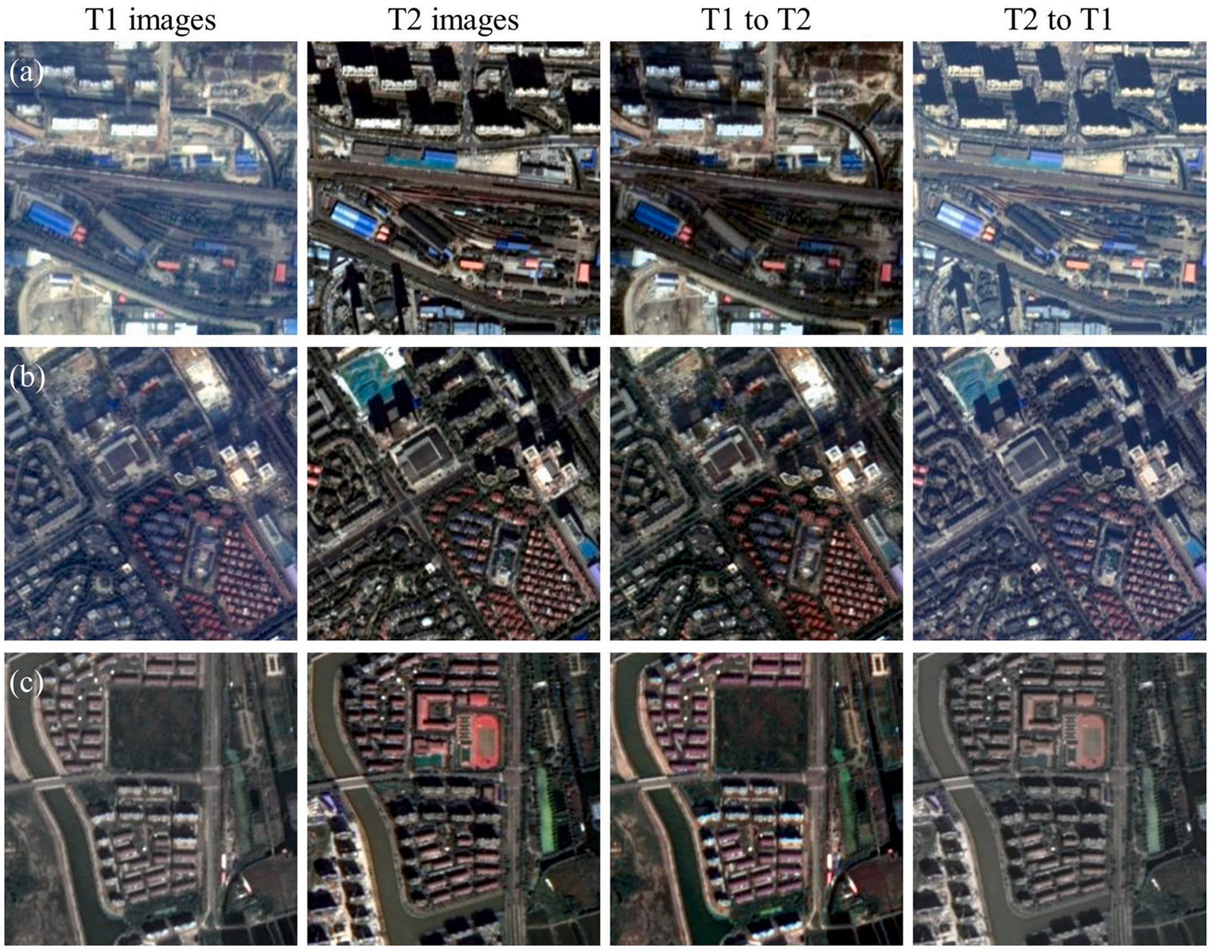

Fig. 6. Illustration of the color transfer in regions (a-c). "T1 (T2) to T2 (T1)" means that the color space of T1 (T2) images is transferred to that of T2 (T1) images. Each image has a spatial extent of 256 × 256 pixels.

including "T1 to T2" and its opposite direction "T2 to T1".

Part 2. Layer-wise temporal difference at the feature level. First, we employed the same feature encoder to extract multi-layer features for each temporal image. Then, we computed the absolute difference of bi-temporal features layer by layer (i.e., "Sub & Abs" in Fig. 7). Note that the feature difference is considered to suppress pseudo changes and highlight true changes. Moreover, the two encoders for feature extraction share parameters during training so that they can learn features from different temporal images simultaneously. This setting facilitates the network to learn general building features (e.g., shape and size) that

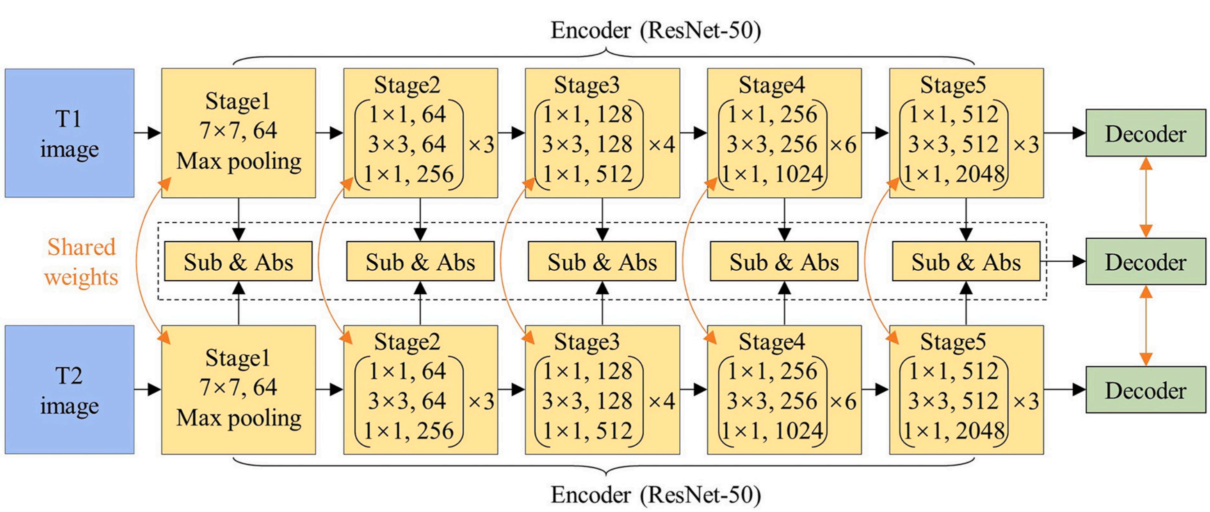

Fig. 7. Illustration of the layer-wise temporal difference at the feature level. "Sub & Abs" means the absolute difference of bi-temporal features. Orange arrows represent weight sharing.

are not limited to a specific imaging condition, thus improving the generalization performance of the network. The structure of the encoder was set to ResNet-50 with initialization parameters from Section 3.1. It is worth noting that in current studies, the network parameters are usually initialized by random generation or using the pretrained weights from natural image datasets, such as ImageNet (Jia Deng et al., 2009), and then are fine-tuned using high-quality true change labels. However, the true change labels are expensive and limited, which hinders the performance improvement of the network. In this context, this study leverages the pretrained building extraction network (Section 3.1) for parameter initialization, which can greatly reduce the network's reliance on a large number of manually labeled labels.

Part 3. Simultaneous extraction of changed and unchanged buildings at the decision level. At the decision level, we used three decoders with shared weights (consistent with the decoder of Section 3.1) to extract buildings for each temporal image and changed buildings. Subsequently, we computed the union of bi-temporal buildings (i.e., "Add" in the step 3 of Fig. 3), and then removed the changed buildings (i.e., "Sub" in the step 3 of Fig. 3) to obtain the unchanged buildings. The change detection network can simultaneously identify both changed and unchanged buildings, which is helpful for the decoder to perceive all the buildings in time-series images. The loss function is defined as:

$$L_{cd} = L_c\left(q_c^r, p_c\right) + L_u\left(q_u^r, p_u\right) \tag{7}$$

where q_c^r and q_u^r denote the reliable changed and unchanged building labels (see Section 3.2), respectively, while p_c and p_u represent the probabilities of changed and unchanged buildings predicted by the network, respectively. The loss functions L_c and L_u are both the sum of the binary cross-entropy loss and the dice coefficient loss.

When training the network, we set the initial learning rate to 0.001 and reduced the learning rate by a factor of 0.1 at the 10th and 15th epochs. The total number of epochs was set to 20 and the batch size was set to 16. We selected Adam as the optimizer (Kingma and Ba, 2014). The data augmentation is consistent with Section 3.1. Besides, in Section 5.3, we analyzed the performance of the three parts of the proposed network. Notice that compared with the building extraction network in Section 3.1, the proposed change detection network does not introduce any new parameters, and directly migrates the well-trained parameters of the former, lowering the reliance on a large number of high-quality true labels. Moreover, to alleviate the temporal and regional differences in cross-task transfer learning, we further leveraged the reliable change pseudo labels (see Section 3.2) to fine-tune the change detection network, without resorting to true labels. In this way, the high acquisition costs for the true labels of the buildings as well as their changes can be significantly lowered. The performance of the pseudo labels is evaluated by comparing them with the true labels in Section 5.4.

3.4. Accuracy assessment

We used the building change samples (Section 2.2) to evaluate the proposed method, and calculated five accuracy metrics, including overall accuracy (abbreviated OA), intersection over union (IoU), F1-score (F1), precision (Prec), and recall (Rec). These metrics are commonly used for building change detection (Liu et al., 2021; Sun et al., 2022b) and are defined as follows:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

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

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

where TP (true positive) is the number of pixels correctly predicted as changed buildings, FP (false positive) is the number of pixels incorrectly predicted as changed buildings, TN (true negative) is the number of pixels correctly predicted as non-changed buildings, and FN (false negative) is the number of pixels incorrectly predicted as non-changed buildings. OA represents the precision of all classes (including changed and non-changed buildings), and IoU denotes the overlap degree of predicted and referenced changed buildings. Higher precision means fewer false alarms, and higher recall indicates fewer omissions. F1 balances the precision and recall of changed buildings, and outperforms OA in the case of class imbalance.

4. Results

4.1. Comparison with existing methods

To verify the effectiveness of the proposed method, we compared ten state-of-the-art change detection methods, including Deep CVA (Saha et al., 2019), Deep IRMAD (Nielsen, 2007), Deep PCA-Kmeans (Celik, 2009), FC-EF (Caye Daudt et al., 2018), FC-Siam-con (Caye Daudt et al., 2018), FC-Siam-diff (Caye Daudt et al., 2018), L-UNet (Papadomanolaki et al., 2021), SNUNet (Fang et al., 2022b), BIT (Chen et al., 2022b), and CSA-CDGAN (Wang et al., 2022b). For a fair comparison, the first three methods take the last layer of features of the trained building extraction network (Section 3.1) as input, while the last seven methods utilize the generated reliable pseudo labels (Section 3.2) to fine-tune the network parameters. In addition, for FC-EF, FC-Siam-con, and FC-Siam-diff, their network structures are consistent with the building extraction network (Section 3.1) used in this study. These methods are briefly introduced as follows:

1) Deep CVA. Deep change vector analysis (CVA) (Saha et al., 2019) takes a pretrained network as a feature extractor to obtain the features for each temporal image, and then applies the CVA algorithm (Bovolo and Bruzzone, 2007) on automatically selected features to detect changes.
2) Deep IRMAD. The iteratively reweighted multivariate alteration detection (IR-MAD) (Nielsen, 2007) adopts canonical correlation analysis (CCA) to calculate the difference of canonical variates and then iteratively assigns different weights to observations to identify changes. Note that we adopt the last layer of the trained building extraction network (Section 3.1) as the input of the traditional IR-MAD (Nielsen, 2007), and call this method "Deep IRMAD".
3) Deep PCA-Kmeans. PCA-Kmeans (Celik, 2009) applies principal component analysis (PCA) on the difference of bi-temporal images to extract the feature vector, and then uses the k-means clustering algorithm to partition the feature vector into changed and unchanged regions.
4) FC-EF. Fully convolutional early fusion (FC-EF) (Caye Daudt et al., 2018) stacks bi-temporal images along the channel dimension at the input level and then feeds them into a encoder-decoder network to identify changes.
5) FC-Siam-con. Fully convolutional Siamese-concatenation (FC-Siam-con) (Caye Daudt et al., 2018) adopts two encoders with the same structure and shared weights, i.e., the Siamese structure, to extract features for each temporal image, and then stacks bi-temporal features into a decoder for change detection.
6) FC-Siam-diff. Fully convolutional Siamese-difference (FC-Siam-diff) (Caye Daudt et al., 2018) is similar to FC-Siam-con. The

difference between them is that the former uses the absolute difference of the bi-temporal features as the input to the decoder.

7) L-UNet. L-UNet (Papadomanolaki et al., 2021) is built on a deep multitask learning framework that can perform both semantic segmentation and change detection. In particular, it incorporates fully convolutional long short-term memory (LSTM) blocks into every level of the encoder to capture the temporal relationship of spatial features from bi-temporal images.
8) SNUNet. SNUNet (Fang et al., 2022b) is a densely connected Siamese change detection network. It can maintain high-resolution features by densely connecting the encoder and the decoder, and employs an ensemble channel attention module to capture the most representative features for change detection.
9) BIT. Bi-temporal image transformer (BIT) (Chen et al., 2022b) uses the Siamese network to extract the features for each temporal image, then applies the transformer for modeling the spatio-temporal context and refining the bi-temporal features, and finally, the difference of the refined bi-temporal features is used to predict change regions.
10) CSA-CDGAN (Wang et al., 2022b) adopts an encoder-decoder network with a channel self-attention module as a generator to produce a change map, and then applies a discriminator to distinguish the change map and the ground truth. The generator and the discriminator compete with each other, such that the former can obtain a more accurate change map.

We tested the performance of these methods on the building change detection dataset (Section 2.2), and obtained quantitative (Table 2) and qualitative (Fig. 8) results. Table 2 shows that our method consistently outperforms others in terms of overall metrics, i.e., OA, IoU, and F1. This is mostly due to the fact that our method fully leverages change pseudo labels to optimize the network parameters, and simultaneously mitigates the data distribution differences between bi-temporal images at the input, feature, and decision levels. Moreover, we find that, Deep CVA, Deep IRMAD, and Deep PCA-Kmeans, do not leverage pseudo labels to optimize network parameters and perform significantly worse than other methods using pseudo labels.

Fig. 8 displays the visualization results of different change detection methods in dense and sparse building areas, respectively. Overall, by using only pseudo labels, our method can identify changed buildings more effectively and better balance false alarms and omissions, compared with other methods. In addition, we can observe that these methods that do not use pseudo labels (i.e., Deep CVA, Deep IRMAD, and Deep PCA-Kmeans) produce lots of false alarms in unchanged regions (e.g., Fig. 8(c) and (d)), while other methods using pseudo labels effectively suppress these false alarms. Moreover, it can be observed that all methods successfully detect most of changed buildings, thanks to the utilization of the building extraction network (Section 3.1) that was pretrained on crowdsourced building labels from the 27 major Chinese cities.

4.2. Results on the whole study areas

Fig. 9 shows the whole prediction results of our method on the building change detection dataset (Section 2.2). Fig. 9(b) and (d) display multiple changed buildings, such as newly-built, demolished, and reconstructed buildings. Overall, by courtesy of crowdsourced building labels from the 27 major Chinese cities, our method effectively identifies most building change types and suppresses pseudo changes in unchanged areas, such as the color change of the same building (Fig. 9(e)) and the illumination change in multi-temporal images (Fig. 9(f)).

5. Discussions

5.1. Performance of noisy label correction

To verify the effectiveness of the noisy label correction module, we compared the results with and without this module on two aspects, namely building change pseudo labels (obtained by the object-to-pixel multi-temporal comparison in Section 3.2) and building change detection (Section 3.3). As shown in Table 3, for both aspects, the noisy label correction module can significantly improve the overall metrics (i.e., OA, IoU, and F1), indicating its effectiveness. Furthermore, the overall accuracy is higher in the task of change detection than that in the task of generating pseudo labels, since the change detection results are further optimized and refined through the pseudo labels. This finding further confirms the necessity of fine-tuning the network using the pseudo labels, which is also reflected in Table 2.

Fig. 10 presents the original and corrected crowdsourced building labels. We can observe that, due to the inconsistency of the acquisition time with ZY-3 images, the original labels miss the newly-built buildings (e.g., Fig. 10(a-d)), and still retain the demolished buildings (e.g., Fig. 10 (e-f)). However, after noisy label correction, these omissions and false alarms are effectively removed. The corrected labels are relatively clean, which helps to improve the robustness of the network to label noise.

Fig. 11 displays the proportion of building pixels before and after noisy label correction (NLC), and the change proportion of building pixels after NLC for each training sample. We can see that for most samples, the proportion of building pixels increases after NLC, indicating that original samples suffer a lot from omissions. This phenomenon is also reflected in Fig. 10(a-d). Besides, we also calculated the average change proportion of building pixels ($P_{average}$) over all the training samples i.e., $P_{average} = \frac{1}{N}\sum_{i=1}^{N}\left(P^{i}_{after} - P^{i}_{before}\right)$, where P^{i}_{before} and P^{i}_{after} denote the proportion of building pixels of the i-th training sample before and after NLC, respectively, and N is the number of training samples. As is shown in Fig. 11(b), $P_{average}$ is 4.91% for those samples with ($P^{i}_{after} - P^{i}_{before}$)>0, and 1.02% for those samples with ($P^{i}_{after} - P^{i}_{before}$)<0.

Table 2
Accuracies of building change detection with different methods. The highest value for each metric is marked in bold.

Method	Shanghai					Beijing				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
Deep CVA	0.950	0.452	0.622	0.504	0.814	0.921	0.517	0.681	0.558	**0.876**
Deep IRMAD	0.950	0.453	0.623	0.505	0.813	0.921	0.517	0.682	0.558	**0.876**
Deep PCA-Kmeans	0.950	0.460	0.630	0.506	**0.834**	0.918	0.509	0.674	0.548	0.875
FC-EF	0.978	0.627	0.770	**0.820**	0.727	0.962	0.671	0.803	0.807	0.799
FC-Siam-Con	0.974	0.599	0.749	0.736	0.763	0.942	0.577	0.732	0.659	0.823
FC-Siam-Diff	0.977	0.630	0.773	0.786	0.761	0.952	0.626	0.770	0.716	0.833
BIT	0.976	0.605	0.754	0.780	0.729	0.949	0.613	0.760	0.703	0.827
LUNet	0.973	0.574	0.729	0.735	0.724	0.950	0.611	0.759	0.708	0.817
SNUNet	0.975	0.598	0.749	0.761	0.737	0.948	0.606	0.755	0.690	0.833
CSA-CDGAN	0.978	0.629	0.772	0.817	0.732	0.963	0.674	0.806	**0.811**	0.800
Ours	**0.979**	**0.644**	**0.783**	**0.820**	0.750	**0.964**	**0.691**	**0.817**	0.796	0.840

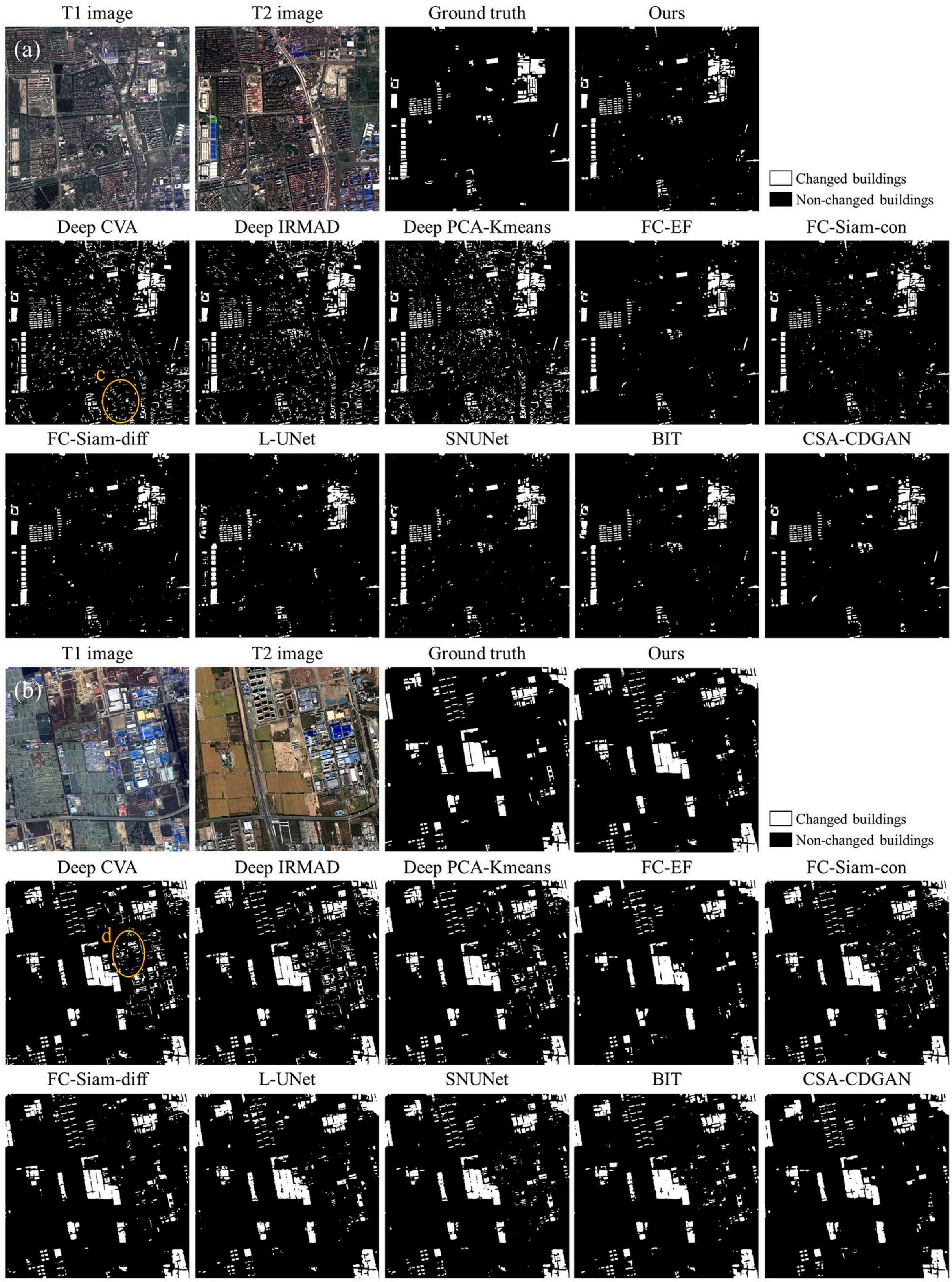

Fig. 8. Results of different change detection methods on two test samples (1024 × 1024 pixels for each) from Shanghai (a) and Beijing (b), respectively.

5.2. Performance of uncertainty-aware pseudo label generation

To evaluate the performance of the uncertainty-aware pseudo label generation method, we discussed the impact of three aspects on the building change detection results: 1) the multi-temporal comparison based on different analysis units (i.e., pixel, object, and object-to-pixel); 2) the inclusion of the uncertainty-aware analysis; and 3) the threshold selection for the uncertainty-aware analysis.

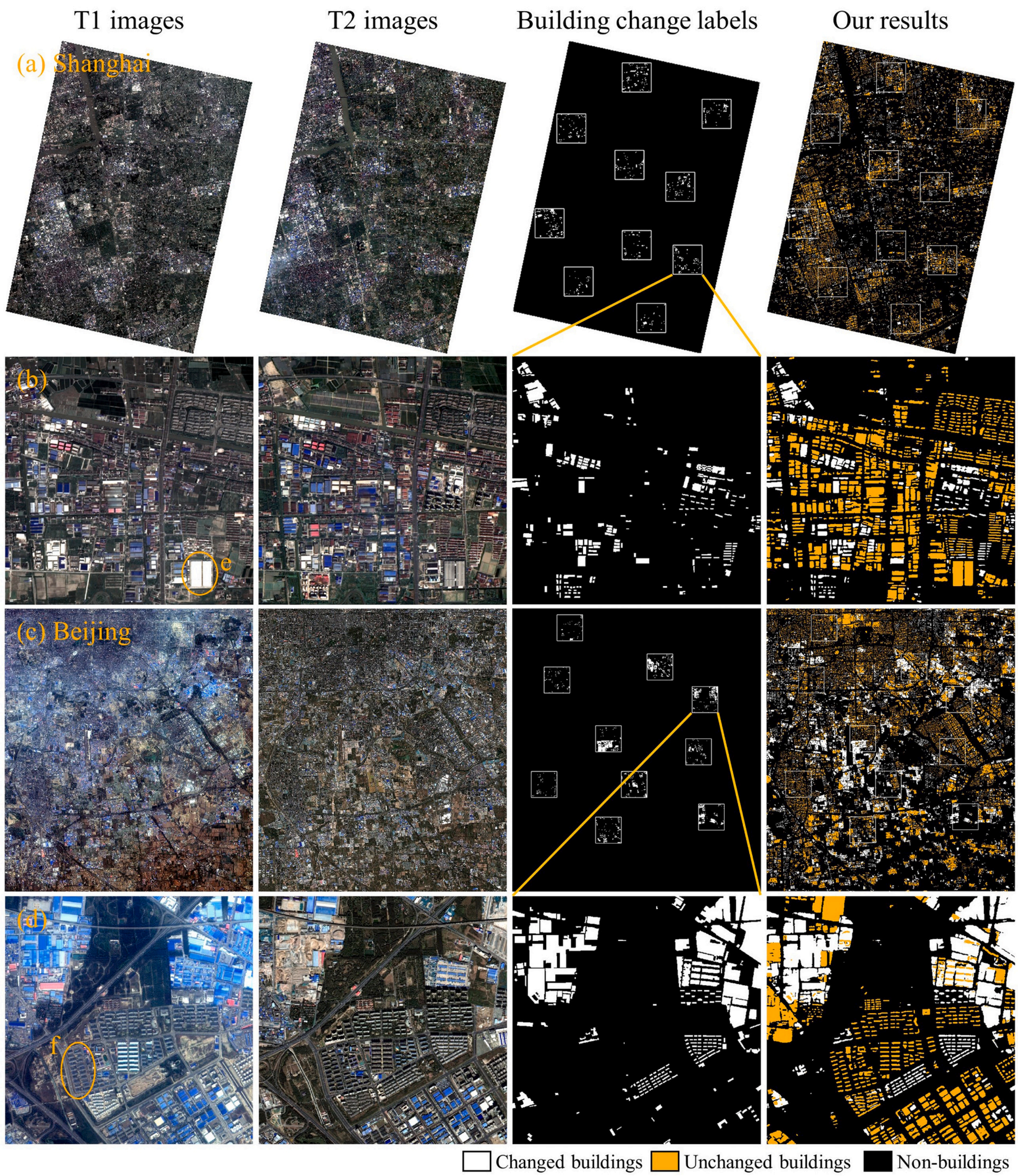

Fig. 9. Results of the proposed method on the building change detection dataset from Shanghai (a) and Beijing (c). Each enlarged view (b and d) has a spatial extent of 1024 × 1024 pixels.

Table 3
Accuracies of building change pseudo labels and building change detection with and without (w/o) noisy label correction (NLC). The highest value for each metric is marked in bold.

Task	NLC	Shanghai					Beijing				
		OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
Pseudo labels	w/o	0.959	0.431	0.602	0.592	0.613	0.901	0.384	0.555	0.491	0.637
	with	0.974	0.603	0.752	0.737	**0.768**	0.951	0.625	0.769	0.709	**0.841**
Change detection	w/o	0.970	0.441	0.612	**0.896**	0.465	0.945	0.454	0.625	**0.926**	0.471
	with	**0.979**	**0.644**	**0.783**	0.820	0.750	**0.964**	**0.691**	**0.817**	0.796	0.840

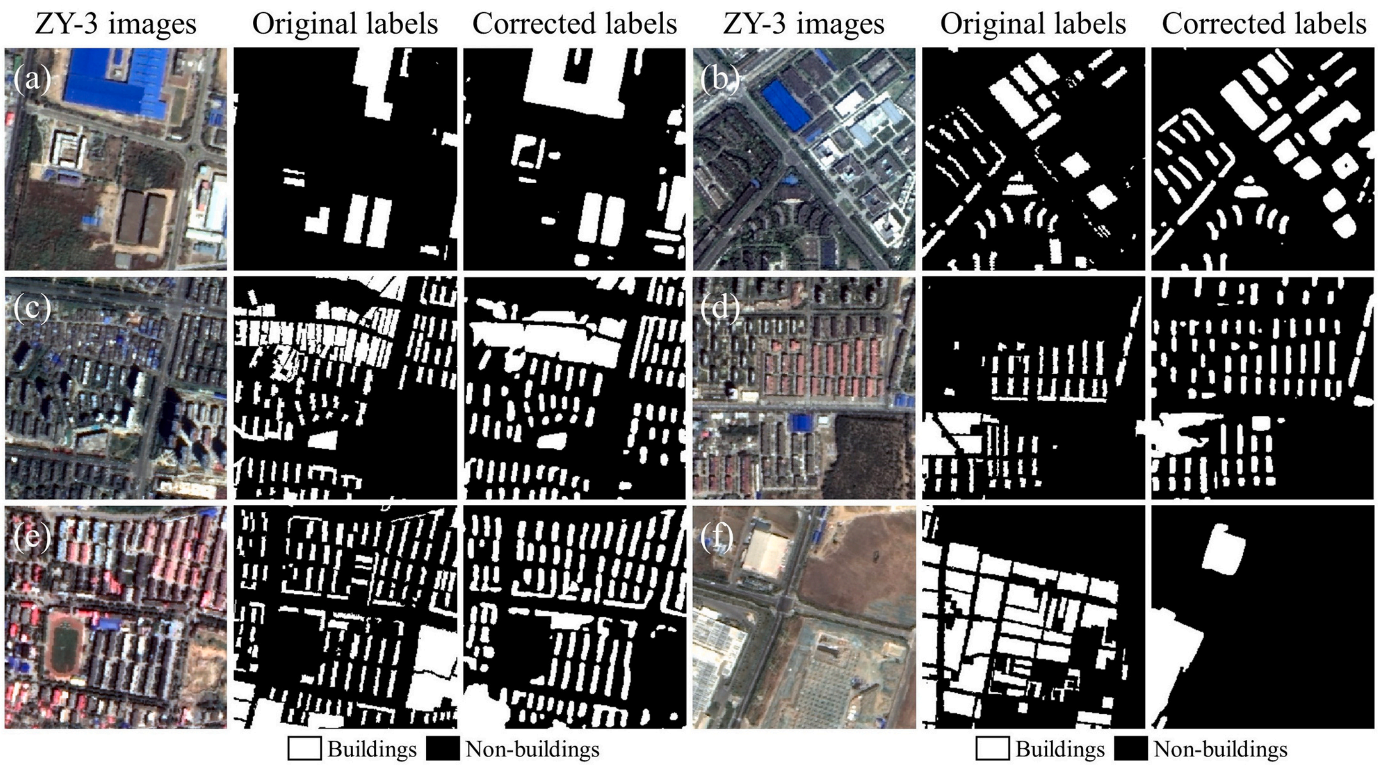

Fig. 10. Examples of the noisy label correction on regions (a-f). Each image has a spatial extent of 256 × 256 pixels.

Table 4 records the quantitative results of the first two aspects. In terms of the analysis unit, the object-to-pixel unit performs the best, followed by the object unit, and the worst is the pixel unit. Based on the object-to-pixel multi-temporal comparison, we further introduced the uncertainty-aware analysis to generate reliable pseudo labels for change detection. The results in Table 4 indicate that the inclusion of the uncertainty-aware analysis significantly improves overall metrics, i.e., the OA, IoU, and F1 values of building change detection, demonstrating its superiority.

Fig. 12 shows the building change pseudo labels generated by different analysis units and with and without the uncertainty-aware analysis. We can observe that the pixel-based comparison is susceptible to the residual matching errors and the imaging condition difference between bi-temporal images, thereby introducing undesired pseudo changes, e.g., Fig. 12(a). By contrast, the object-based comparison can reduce these pseudo changes. However, limited by the spatial resolution of ZY-3 images (2.5 m) and the quality of the crowdsourced building labels (containing noise, e.g., false alarms, omissions, and mismatches), the boundaries of buildings may not be accurately identified. This issue makes it difficult for the object-based comparison to well distinguish the changed and unchanged regions within dense building areas, e.g., Fig. 12(b). Fortunately, the proposed object-to-pixel comparison can effectively alleviate this issue. However, affected by the temporal and regional differences in cross-task transfer learning, pseudo labels may still contain some noise. The use of all pseudo labels for training tends to degrade the generalization performance of the network. Thus, we further designed the uncertainty-aware analysis to eliminate the uncertain labels (e.g., the gray areas in Fig. 12(c)) and trained the network with only the reliable labels.

The key parameter of the uncertainty-aware analysis is the change probability threshold (Section 3.2), and its influence on the results of building change detection is presented in Fig. 13. It can be seen that, as the threshold increases, the precision tends to increase, indicating a decrease in false alarms; but meanwhile, the recall tends to decrease, representing an increase in omissions. In order to better balance false alarms and omissions, the threshold was set to 0.7 in this study, since this value can obtain the satisfactory F1-score values in both study areas.

5.3. Performance of full-level fused building change detection

To validate the effectiveness of the full-level fused building change detection method, we analyzed the building change detection results under four conditions: 1) without color transfer at the input level; 2) without layer-wise temporal difference at the feature level; 3) without unchanged building extraction at the decision level; and 4) with different color transfer directions, including the pre-temporal (T1) to post-temporal (T2) transfer (i.e., T1 to T2) and its opposite direction (T2 to T1). As displayed in Table 5, the proposed method that considers the modules from all levels (i.e., the first three conditions) performs the best, where the feature-level module (i.e., layer-wise temporal difference) contributes the most. We can also observe that the color transfer direction has little influence on the results. Besides, it is possible to use the widely-used Wallis filter method (Li et al., 2006; Li et al., 2020c) for color transfer. The Wallis filter method can linearly transfer the mean value (μ) and standard deviation (σ) of source image (x^s) to those of target image (x^t), and obtain the transferred image (x^{ts}) by the equation: $x^{ts} = \sigma(x^s)\left(\frac{x^t-\mu(x^t)}{\sigma(x^t)}\right) + \mu(x^s)$. The performance of the Wallis filter is recorded in Table 5. We can observe that the Wallis filter obtains competitive results compared to the color transfer method used in this study. This result again verifies the necessity of considering the difference in imaging conditions (e.g., atmosphere and illumination) between time-series images at the input level. Note that when the number of available time-series images is more than two, we can consider some established color consistency correction methods for multiple images to mitigate the data distribution difference at the input level (Li et al., 2020a; Li et al., 2022).

Fig. 14 presents the visualization results under the aforementioned four conditions. We can find that the method without the module at any level (i.e., input, feature, or decision levels) introduces lots of pseudo changes (e.g., Fig. 14(c-h)), especially when the feature-level module is

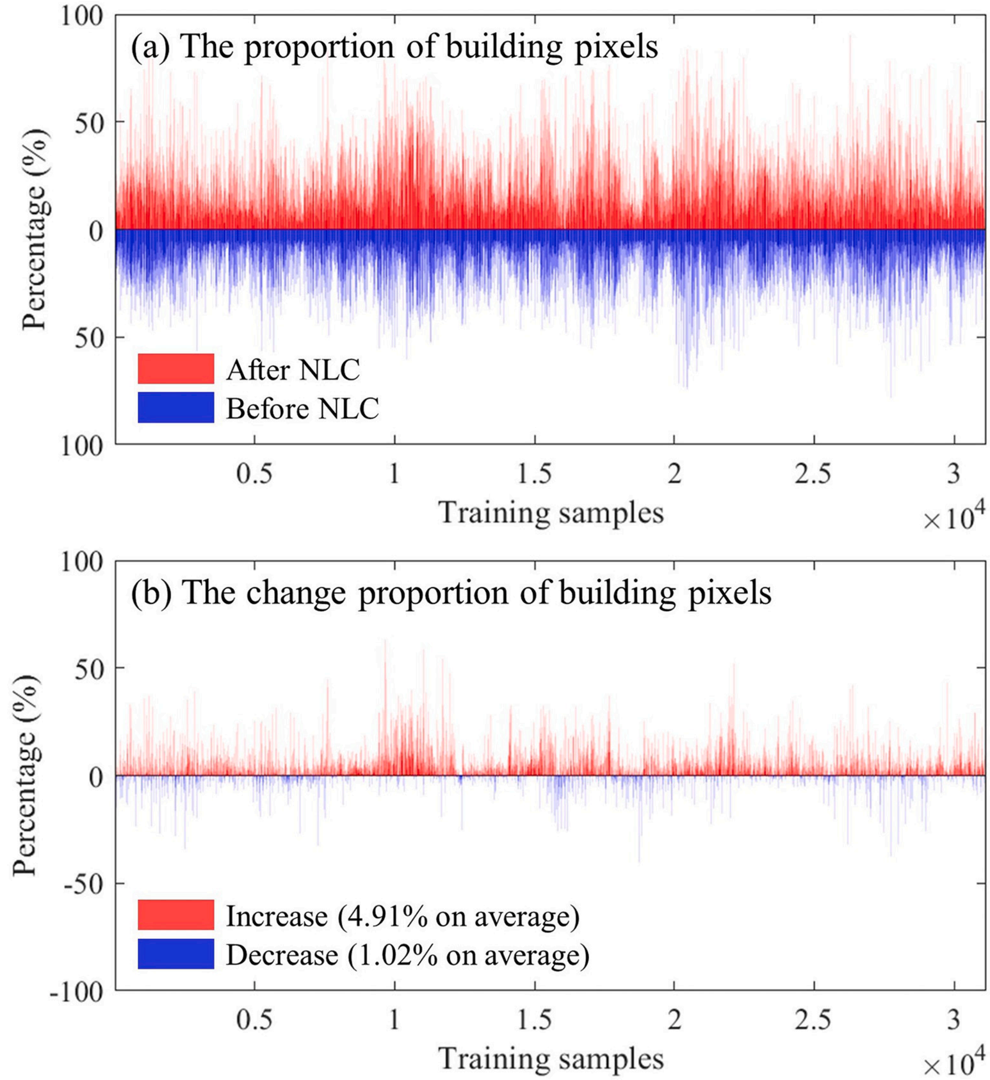

Fig. 11. (a) The proportion of building pixels before and after noisy label correction (NLC) for each training sample. (b) The change proportion of building pixels after NLC for each training sample.

Table 4
Accuracies of building change detection with different analysis units and with/without the uncertainty-aware analysis (UAA). The highest value for each metric is marked in bold.

Unit	UAA	Shanghai					Beijing				
		OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
Pixel	√	0.976	0.610	0.758	0.774	0.743	0.955	0.645	0.784	0.727	0.851
Object	√	0.977	0.628	0.772	0.787	0.756	0.960	0.672	0.804	0.769	0.841
Object-to-pixel	×	0.977	0.629	0.772	0.766	**0.779**	0.955	0.651	0.788	0.729	**0.858**
	√	**0.979**	**0.644**	**0.783**	**0.820**	0.750	**0.964**	**0.691**	**0.817**	**0.796**	0.840

removed. These pseudo changes are mainly caused by the data distribution differences between bi-temporal images. By contrast, the proposed method with the modules from all levels suppresses these pseudo changes better while ensuring high completeness, verifying its effectiveness. In addition, the prediction results of different color transfer directions are similar, indicating the low sensitivity of the color transfer to the direction.

5.4. Comparison between pseudo and true labels

Pseudo labels are the core of cross-task transfer learning, and are built on the building extraction network pretrained with crowdsourced building labels from the 27 major Chinese cities. This greatly mitigates the workload for manually labeling the buildings and their change samples. To investigate the performance of pseudo labels, we compared them with true labels. Specifically, for each study area, we randomly selected five patches (1024 × 1024 pixels for each patch) from the test samples as the training set (i.e., true labels) and the remaining patches as the testing set. For a fair comparison, we trained the change detection network under three settings: 1) using only true labels for training; 2) using only pseudo labels for training; and 3) first using pseudo labels for pretraining and then using true labels for fine-tuning. The experimental

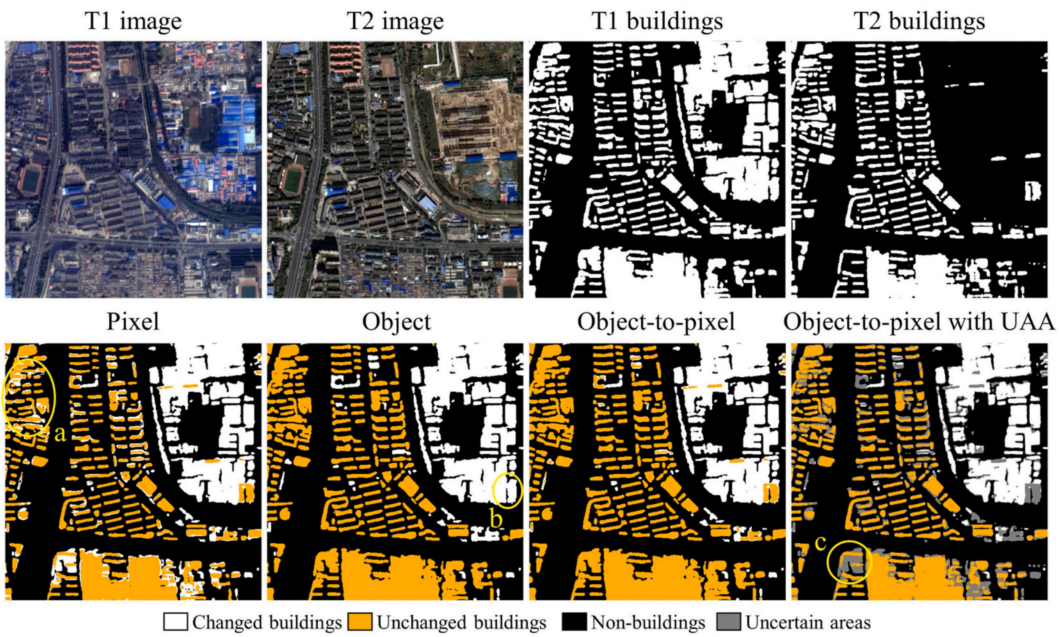

Fig. 12. Pseudo labels generated with different analysis units (i.e., pixel, object, and object-to-pixel) and with the uncertainty-aware analysis (UAA). Each image has a spatial extent of 512 × 512 pixels.

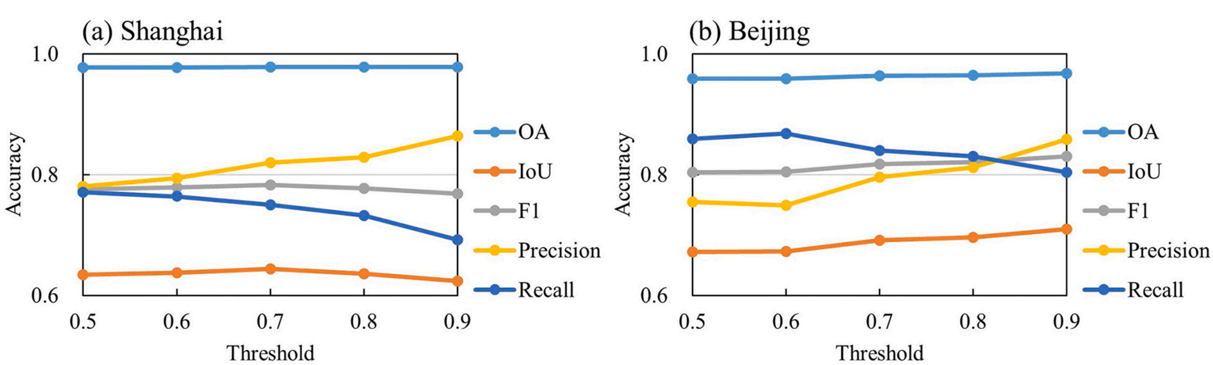

Fig. 13. Accuracies of building change detection with different thresholds for the uncertainty-aware analysis in Shanghai (a) and Beijing (b).

Table 5
Accuracies of building change detection without (w/o) input-, feature-, and decision-level modules and with two directions of color transfer (i.e., T1 to T2 and T2 to T1). "Full" indicates the modules from all levels. "Wallis" represents the Wallis filter for color transfer. The highest value for each metric is marked in bold.

Module	Shanghai					Beijing				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
w/o input-level	0.978	0.636	0.777	0.815	0.743	0.951	0.626	0.770	0.709	0.842
w/o feature-level	0.946	0.426	0.598	0.481	**0.787**	0.920	0.507	0.673	0.555	**0.853**
w/o decision-level	0.978	0.635	0.777	0.790	0.764	0.963	0.690	0.817	0.792	0.843
Full (T1 to T2)	**0.979**	**0.644**	**0.783**	0.820	0.750	0.964	0.691	0.817	0.796	0.840
Full (T2 to T1)	**0.979**	0.640	0.781	**0.822**	0.743	0.963	0.690	0.817	0.791	0.844
Wallis (T1 to T2)	0.978	0.638	0.779	0.805	0.754	**0.966**	**0.702**	**0.825**	**0.818**	0.831

results are recorded in Table 6. We can find that the method using only pseudo labels significantly outperforms that using only true labels, while the method using both labels performs slightly better than the former. The main reason for this phenomenon is that the number of available pseudo labels is much higher than that of true labels, which helps to sufficiently optimize the network parameters and improve the generalization performance of the network. These results confirm that pseudo labels hold great potential in situations where true labels are lacking or limited, thus mitigating the high acquisition cost of true labels.

Fig. 15 presents the predicted changed buildings using true labels, pseudo labels, and their combination. We can see that the method using only true labels is prone to miss the changed buildings (e.g., Fig. 15(c)), due to the limitation of label size, and meanwhile it is vulnerable to the imaging condition differences between time-series images, leading to the incorrect identification of unchanged areas (e.g., Fig. 15(d)). In contrast, the method using only pseudo labels can better balance the omissions and false alarms, and obtains more satisfactory results in building change detection. This phenomenon can be attributed to the full use of large-scale crowdsourced building labels.

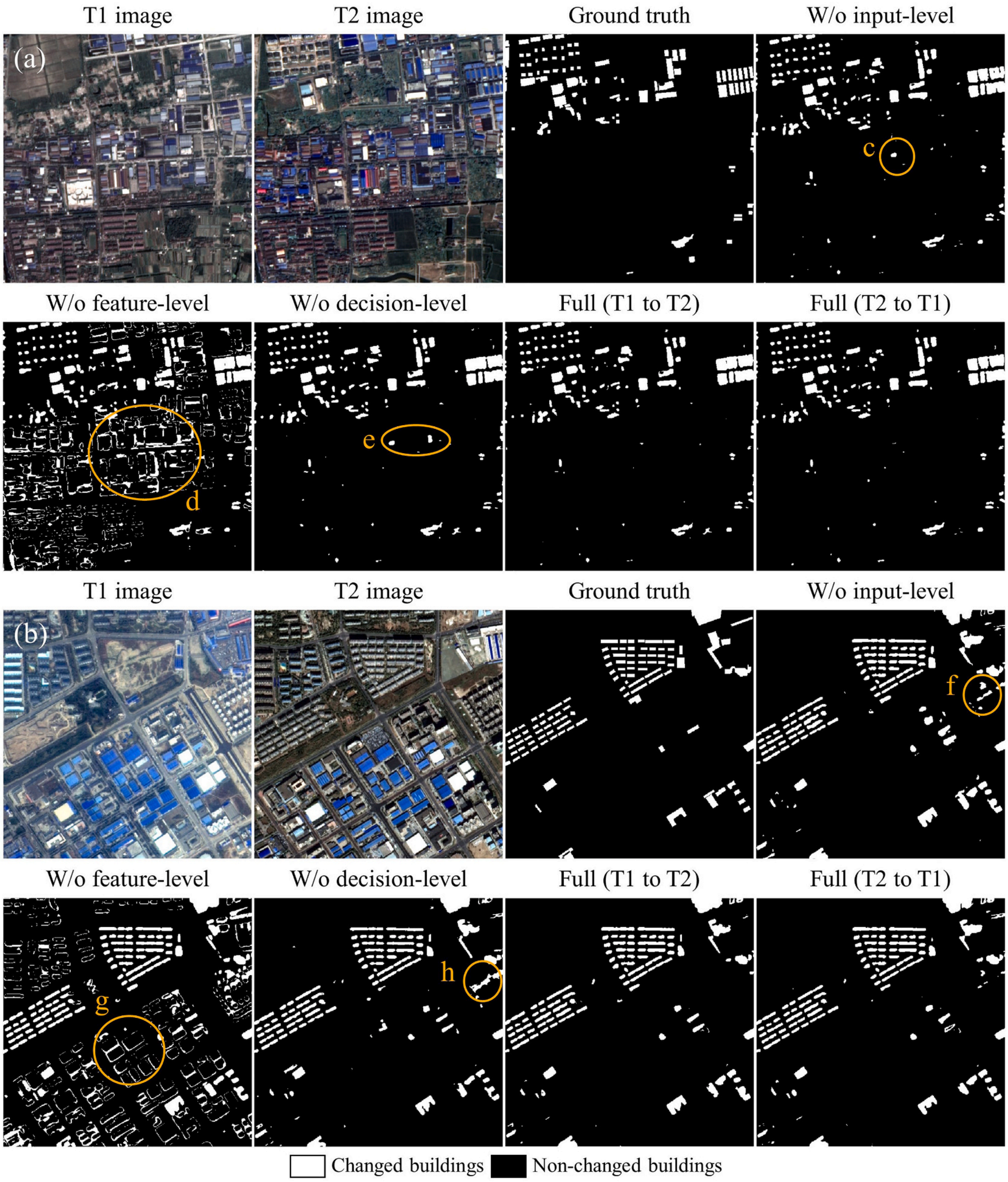

Fig. 14. Results of building change detection without (w/o) input-, feature-, and decision-level modules and with two directions of color transfer (i.e., T1 to T2 and T2 to T1) in areas from Shanghai (a) and Beijing (b). Each image has a spatial extent of 512 × 512 pixels.

Table 6
Accuracies of building change detection trained with pseudo labels, true labels, and their combination. The highest value for each metric is marked in bold.

Training labels	Shanghai					Beijing				
	OA	IoU	F1	Pre	Rec	OA	IoU	F1	Pre	Rec
True	0.965	0.503	0.669	0.816	0.567	0.939	0.498	0.665	0.817	0.561
Pseudo	0.974	0.642	0.782	0.835	**0.735**	0.957	0.675	0.806	0.789	0.823
Pseudo + True	**0.976**	**0.651**	**0.789**	**0.856**	0.731	**0.962**	**0.699**	**0.823**	**0.821**	**0.824**

5.5. *Effect of the spatial heterogeneity of samples*

In this study, considering that buildings in different cities may have different sizes, shapes, and colors, we used crowdsourced building labels from the 27 major cities in China to enhance the generalization performance of the network. In this section, we further explored the effect of the spatial heterogeneity of samples across multiple cities on change detection. Specifically, we divided these cities into four geographic regions across China: east, west, south, and north (Fig. 1). For noisy building samples, in each region (i.e., east, west, south, and north), we randomly selected 6000 patches (256 × 256 pixels for each patch) for training and 600 patches for validation, to train the building extraction network for a fair comparison. To test the accuracy of change detection, we collected bi-temporal ZY-3 images in four cities from different geographic regions, i.e., Shanghai (east), Kunming (west), Shenzhen (south), and Beijing (north). For each city, we manually interpreted

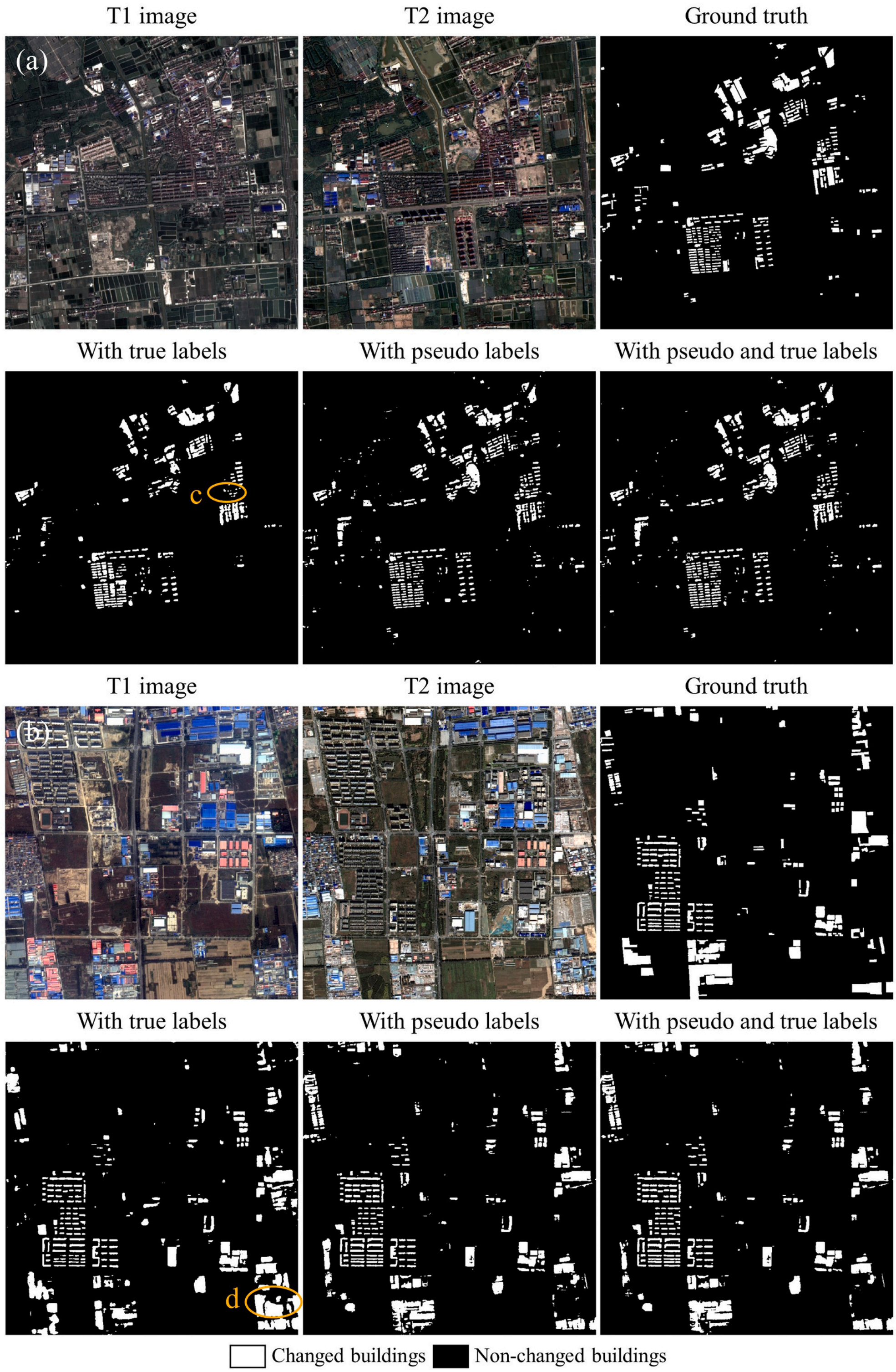

Fig. 15. Results of building change detection trained with true labels, pseudo labels, and their combination in areas from Shanghai (a) and Beijing (b). Each image has a spatial extent of 1024 × 1024 pixels.

changed buildings in 10 random patches (1024 × 1024 pixels for each patch) as test samples (Fig. 2 and Fig. 16). To train the building change detection network, we used all samples except test samples in four cities (i.e., Shanghai, Beijing, Kunming, and Shenzhen). The experimental results are presented in Fig. 17 and Fig. 18. Overall, it can be found that samples from different geographic regions has a significant influence on change detection. With respect to the overall metrics (i.e., OA, IoU, and F1), the method using samples from all regions has the best performance in the four test cities, indicating the necessity of adopting samples across multiple cities. Among the four geographic regions (i.e., east, west, south, and north), the methods using samples from eastern and western regions performs better than those using samples from southern and northern regions. This phenomenon may be attributed to the diversity of samples in eastern and western regions, which increases the spatial heterogeneity of samples and thus improves the generalization capability of the network.

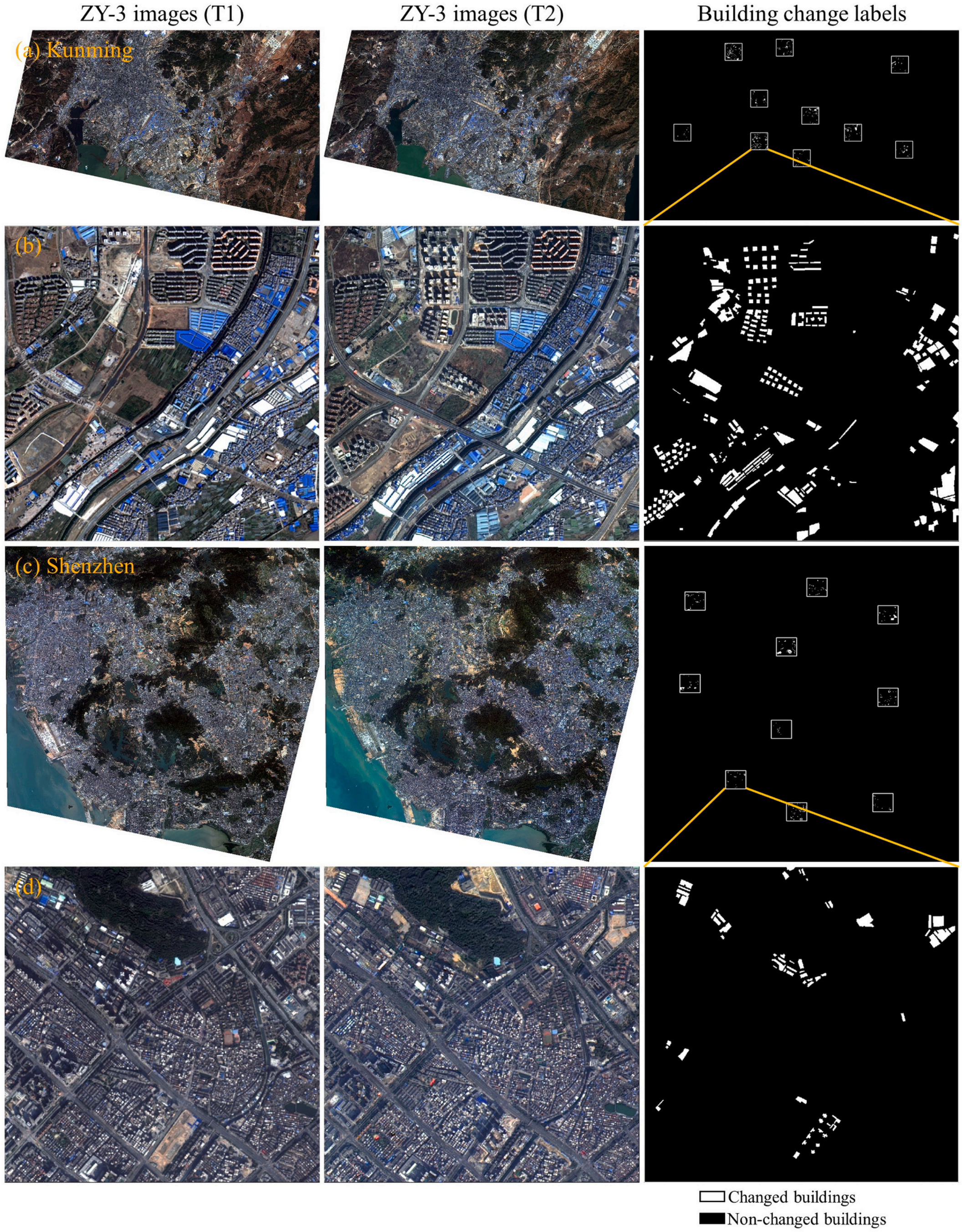

Fig. 16. Bi-temporal ZY-3 images and the corresponding building change labels in Kunming (a) and Shenzhen (c). The spatial extent of the ZY-3 image is 18,881× 11,413 pixels for Kunming (a) and 15,857 × 17,621 pixels for Shenzhen (c). Each enlarged view (b and d) has a spatial extent of 1024 × 1024 pixels. T1 and T2 in Kunming are Feb. 9, 2013 and Jan. 17, 2016, respectively, while T1 and T2 in Shenzhen represent Dec. 23, 2013 and Feb. 11, 2017, respectively.

6. Conclusions

This study is concerned with building change detection using bi-temporal high-resolution satellite imagery, which is important for understanding urban development. Although fully supervised deep learning-based methods for building change detection have made progress, they tend to fuse temporal information only at a single level (input, feature, or decision levels) to mitigate the data distribution differences between time-series images, which is highly prone to introduce a large number of pseudo changes. Moreover, these methods rely on a large number of high-quality pixel-level change labels with high acquisition costs. In contrast, available crowdsourced building data are abundant but are less considered for change detection. For example, OpenStreetMap (OSM), Google Map, and Gaode Map provide lots of available building labels, yet they usually contain noise such as false alarms, omissions, and mismatches, limiting their wide application. When the building extraction task is transferred to the building change detection task, the temporal and regional differences between different images are highly likely to introduce a large number of pseudo changes. To mitigate these limitations, we proposed a full-level fused cross-task transfer learning method, which can perform building change detection using only a large number of crowdsourced building labels and high-resolution satellite images. The method consists of three parts: 1) noise-robust building extraction network pretraining; 2) uncertainty-

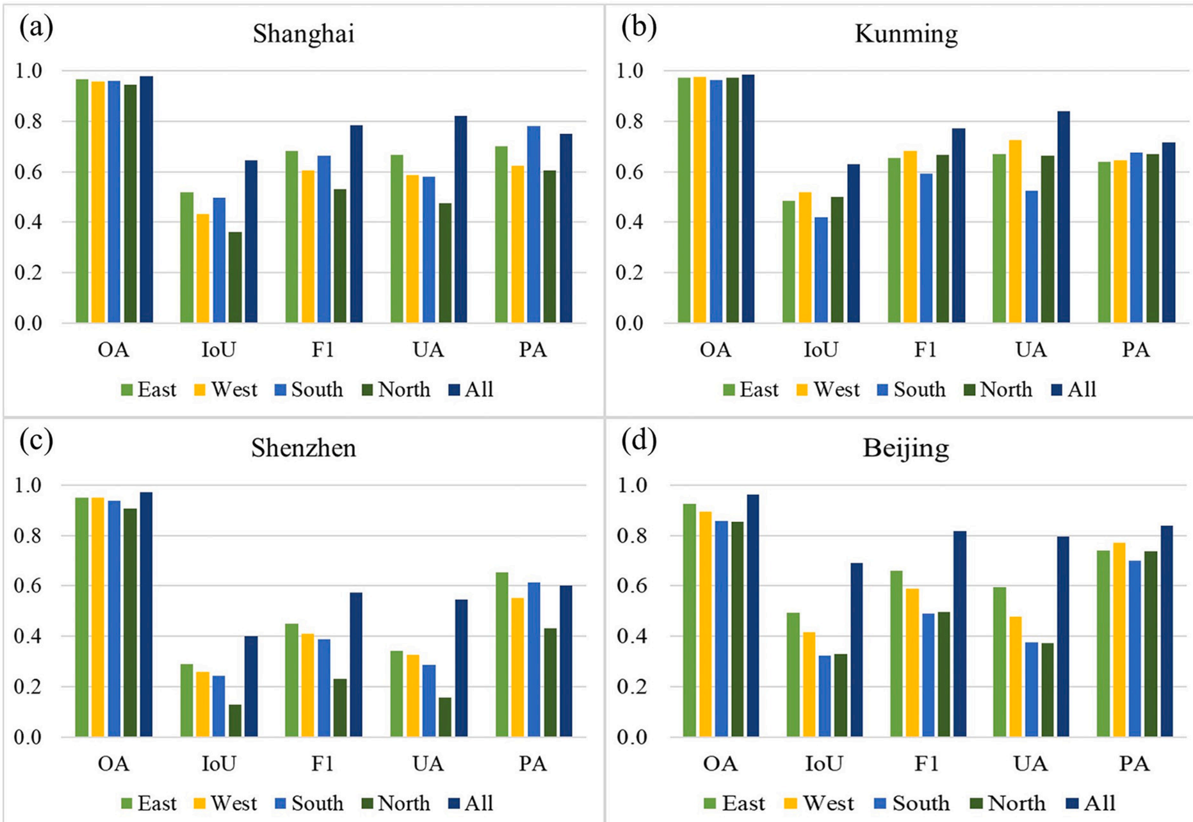

Fig. 17. Accuracies of building change detection using crowdsourced building labels from eastern, western, southern, northern, and all regions in four test cities, i.e., Shanghai (a), Kunming (b), Shenzhen (c), and Beijing (d).

aware pseudo label generation; and 3) full-level fused building change detection.

We collected the building detection dataset, which contains ZY-3 images and the corresponding crowdsourced building labels from the 27 major cities in China. This dataset is merely used for pretraining the building extraction network. In this study, we collected bi-temporal ZY-3 images of Beijing and Shanghai for testing the proposed method. The experimental results showed that our method can identify changed buildings more effectively and better balance false alarms and omissions, compared with the existing state-of-the-art methods. Through further analysis, we summarized five major conclusions:

1) The noisy label correction module can effectively remove noise (e.g., false alarms and omissions) contained in crowdsourced building labels, thus improving the accuracy of building change pseudo labels and building change detection.
2) For the uncertainty-aware pseudo label generation method, the object-to-pixel comparison not only can reduce the pseudo changes introduced by the pixel-based comparison, but also can alleviate the difficulty of the object-based comparison in identifying changes within dense building areas. In addition, the uncertainty-aware analysis can enhance the generalization performance of the network using only reliable labels.
3) The full-level fused building change detection method can simultaneously mitigate the data distribution differences between time-series images at the input, feature, and decision levels.
4) Pseudo labels hold great potential to be applied in situations where true labels are completely lacking or limited, thus alleviating the high acquisition cost of true labels.
5) By considering the spatial heterogeneities, samples from different geographic regions has a significant influence on change detection, and results show that it is necessary to consider samples from multiple cities to improve the performance of change detection.

There still exist three limitations in this study. The first point is the data quality. The data available in this study include crowdsourced building labels and ZY-3 images, and they have the advantage of wide coverage and low acquisition cost. However, the former may contain noise, e.g., false alarms, omissions, and mismatches, while the latter, limited by the spatial resolution (2.5 m), cannot delicately portray the building boundaries within dense building areas like aerial images or unmanned aerial vehicle image (UAV) images. These factors constrain the further improvement of building change detection accuracy. A possible strategy is to employ the subpixel mapping techniques (He et al., 2021) for improving the spatial resolution of predicted results and the edge-enhanced network modules (Xie et al., 2020) for optimizing the boundaries of predicted results. The second point is the uncertainty of pseudo labels. Due to the lack of true labels, pseudo labels were used to supervise the change detection network, which may lead to error accumulation. For example, for the missed buildings in the pseudo labels, the change detection network may also ignore them. In this regard, domain adaptation techniques (Tsai et al., 2018; Wang et al., 2022a) can be explored to alleviate this issue. The third point is the change dimension. This study focuses on the planar changes of buildings, such as demolition and expansion, while does not explore their vertical changes, e.g., height variation. Building heights reflect the vertical form of the city and can be obtained from Light Detection and Ranging (LiDAR), radar, and stereo or multi-view optical imagery (Cao and Huang, 2021; Esch et al., 2022). Note that the ZY-3 satellite can provide both multispectral and multi-view images of the same area, and this property allows us to detect both planar and height changes of buildings at a low cost of image acquisition. In future work, we plan to further

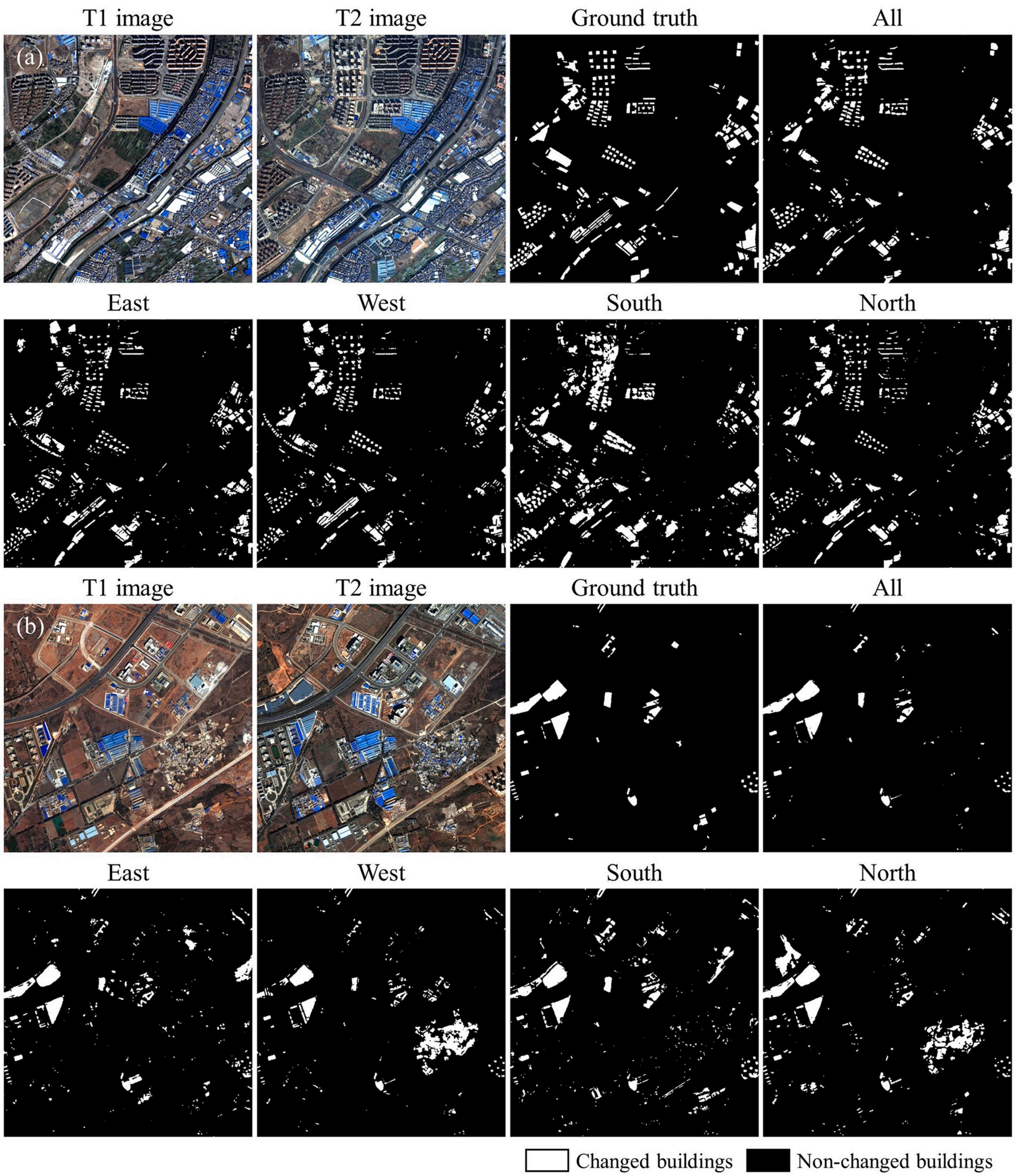

Fig. 18. Results of building change detection using crowdsourced building labels from eastern, western, southern, northern, and all regions. (a) and (b) show dense and sparse building areas in Kunming, respectively. Each image has a spatial extent of 1024 × 1024 pixels.

improve the accuracy of building change detection and estimate the height change of buildings using time-series multi-view ZY-3 images.

CRediT authorship contribution statement

Yinxia Cao: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Xin Huang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision.

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

The authors do not have permission to share data.

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), and the Special Fund of Hubei Luojia Laboratory (under Grant 220100031).

References

Aguirre-Gutiérrez, J., Seijmonsbergen, A.C., Duivenvoorden, J.F., 2012. Optimizing land cover classification accuracy for change detection, a combined pixel-based and object-based approach in a mountainous area in Mexico. Appl. Geogr. 34, 29–37. https://doi.org/10.1016/j.apgeog.2011.10.010.
Ahmed, N., Rahman, R.M., Adnan, M.S.G., Ahmed, B., 2021. Dense prediction of label noise for learning building extraction from aerial drone imagery. Int. J. Remote Sens. 42, 8906–8929. https://doi.org/10.1080/01431161.2021.1973685.
Anniballe, R., Noto, F., Scalia, T., Bignami, C., Stramondo, S., Chini, M., Pierdicca, N., 2018. Earthquake damage mapping: an overall assessment of ground surveys and VHR image change detection after L'Aquila 2009 earthquake. Remote Sens. Environ. 210, 166–178. https://doi.org/10.1016/j.rse.2018.03.004.
Arazo, E., Ortego, D., Albert, P., O'Connor, N.E., McGuinness, K., 2019. Unsupervised label noise modeling and loss correction. In: 36th International Conference on Machine Learning, ICML 2019. PMLR, pp. 465–474.
Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.
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.
Bovolo, F., Bruzzone, L., 2007. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. IEEE Trans. Geosci. Remote Sens. 45, 218–236. https://doi.org/10.1109/TGRS.2006.885408.
Bovolo, F., Bruzzone, L., 2006. A novel theoretical framework for unsupervised change detection based on CVA in polar domain. Int. Geosci. Remote Sens. Symp. 45, 379–382. https://doi.org/10.1109/IGARSS.2006.102.
Bruzzone, L., Prieto, D.F., 2000. Automatic analysis of the difference image for unsupervised change detection. IEEE Trans. Geosci. Remote Sens. 38, 1171–1182. https://doi.org/10.1109/36.843009.
Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: a case study of 42 chinese cities. Remote Sens. Environ. 264, 112590 https://doi.org/10.1016/j.rse.2021.112590.
Caye Daudt, R., Le Saux, B., Boulch, A., 2018. Fully convolutional siamese networks for change detection. In: Proceedings - International Conference on Image Processing, ICIP. IEEE, pp. 4063–4067. https://doi.org/10.1109/ICIP.2018.8451652.
Celik, T., 2009. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. IEEE Geosci. Remote Sens. Lett. 6, 772–776. https://doi.org/10.1109/LGRS.2009.2025059.
Chen, G., Hay, G.J., Carvalho, L.M.T., Wulder, M.A., 2012. Object-based change detection. Int. J. Remote Sens. 33, 4434–4457.
Chen, H., Li, W.Y., Shi, Z.W., 2022. Adversarial instance augmentation for building change detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3066802.
Chen, Hao, 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.
D'Amour, C.B., Reitsma, F., Baiocchi, G., Barthel, S., Güneralp, B., Erb, K.H., Haberl, H., Creutzig, F., Seto, K.C., 2017. Future urban land expansion and implications for global croplands. Proc. Natl. Acad. Sci. U. S. A. 114, 8939–8944. https://doi.org/10.1073/pnas.1606036114.
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.
Esch, T., Brzoska, E., Dech, S., Leutner, B., Palacios-Lopez, D., Metz-Marconcini, A., Marconcini, M., Roth, A., Zeidler, J., 2022. World settlement footprint 3D - a first three-dimensional survey of the global building stock. Remote Sens. Environ. 270, 112877 https://doi.org/10.1016/j.rse.2021.112877.
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.
Fang, S., Li, K., Shao, J., Li, Z., 2022. SNUNet-CD: a densely connected siamese network for change detection of VHR images. IEEE Geosci. Remote Sens. Lett. 19, 16–20. https://doi.org/10.1109/LGRS.2021.3056416.
Gamba, P., Dell'Acqua, F., Stasolla, M., Trianni, G., Lisini, G., 2011. Limits and challenges of optical very-high-spatial-resolution satellite remote sensing for urban applications. Urban Remote Sens. Monit. Synth. Model. Urban Environ. 35–48. https://doi.org/10.1002/9780470979563.ch3.
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.
Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., Briggs, J.M., 2008. Global change and the ecology of cities. Science 319, 756–760. https://doi.org/10.1126/science.1150195.
He, D., Zhong, Y., Wang, X., Zhang, L., 2021. Deep convolutional neural network framework for subpixel mapping. IEEE Trans. Geosci. Remote Sens. 59, 9518–9539. https://doi.org/10.1109/TGRS.2020.3032475.
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.
He, S., Yu, S., Li, G., Zhang, J., 2020. Exploring the influence of urban form on land-use efficiency from a spatiotemporal heterogeneity perspective: evidence from 336 chinese cities. Land Use Policy 95, 104576. https://doi.org/10.1016/j.landusepol.2020.104576.
Hosseinpour, H., Samadzadegan, F., Javan, F.D., 2022. CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. ISPRS J. Photogramm. Remote Sens. 184, 96–115.
Huang, X., Cao, Y., Li, J., 2020. 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., 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., Zhang, L., 2012. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE JSel. Top. Appl. Earth Obs. Remote Sens. 5, 161–172. https://doi.org/10.1109/JSTARS.2011.2168195.
Huang, X., Zhang, L., Zhu, T., 2014. Building change detection from multitemporal high-resolution remotely sensed images based on a morphological building index. IEEE JSel. Top. Appl. Earth Obs. Remote Sens. 7, 105–115. https://doi.org/10.1109/JSTARS.2013.2252423.
Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57, 574–586. https://doi.org/10.1109/TGRS.2018.2858817.
Deng, Jia, Dong, Wei, Socher, R., Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255. https://doi.org/10.1109/cvprw.2009.5206848.
Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. IEEE Trans. Geosci. Remote Sens. 55, 6054–6068. https://doi.org/10.1109/TGRS.2017.2719738.
Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. In: 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.
Laben, C.A., Brower, B.V., 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening.
Lai, Y., Tang, B., Chen, X., Zheng, X., 2021. Spatial determinants of land redevelopment in the urban renewal processes in Shenzhen, China. Land Use Policy 103, 105330. https://doi.org/10.1016/j.landusepol.2021.105330.
Leichtle, T., Geiß, C., Wurm, M., Lakes, T., Taubenböck, H., 2017. Unsupervised change detection in VHR remote sensing imagery – an object-based clustering approach in a dynamic urban environment. Int. J. Appl. Earth Obs. Geoinf. 54, 15–27. https://doi.org/10.1016/j.jag.2016.08.010.
Li, D., Wang, M., Pan, J., 2006. Auto-dodging processing and its application for optical RS images. Geomatics Inf. Sci. Wuhan Univ. 31, 753–756.
Li, L., Xia, M., Liu, C., Li, Liang, Wang, H., Yao, J., 2020. Jointly optimizing global and local color consistency for multiple image mosaicking. ISPRS J. Photogramm. Remote Sens. 170, 45–56. https://doi.org/10.1016/j.isprsjprs.2020.10.006.
Li, Q., Shi, Y., Huang, X., Zhu, X.X., 2020. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF). IEEE Trans. Geosci. Remote Sens. 58, 7502–7519. https://doi.org/10.1109/TGRS.2020.2973720.
Li, X., Luo, M., Ji, S., Zhang, L., Lu, M., 2020. Evaluating generative adversarial networks based image-level domain transfer for multi-source remote sensing image segmentation and object detection. Int. J. Remote Sens. 41, 7327–7351. https://doi.org/10.1080/01431161.2020.1757782.
Li, X., Zhou, Y., Zhu, Z., Liang, L., Yu, B., Cao, W., 2018. Mapping annual urban dynamics (1985–2015) using time series of landsat data. Remote Sens. Environ. 216, 674–683. https://doi.org/10.1016/j.rse.2018.07.030.
Li, Y., Yin, H., Yao, J., Wang, H., Li, L., 2022. A unified probabilistic framework of robust and efficient color consistency correction for multiple images. ISPRS J. Photogramm. Remote Sens. 190, 1–24. https://doi.org/10.1016/j.isprsjprs.2022.05.009.
Liu, C., Huang, X., Zhu, Z., Chen, H., Tang, X., Gong, J., 2019. Automatic extraction of built-up area from ZY3 multi-view satellite imagery: analysis of 45 global cities. Remote Sens. Environ. 226, 51–73.
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.

Liu, Y., Pang, C., Zhan, Z., Zhang, X., Yang, X., 2021. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. IEEE Geosci. Remote Sens. Lett. 18, 811–815. https://doi.org/10.1109/LGRS.2020.2988032.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 3226–3229.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017b. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657. https://doi.org/10.1109/TGRS.2016.2612821.

Marin, C., Bovolo, F., Bruzzone, L., 2015. Building change detection in multitemporal very high resolution SAR images. IEEE Trans. Geosci. Remote Sens. 53, 2664–2682. https://doi.org/10.1109/TGRS.2014.2363548.

Matikainen, L., Hyyppä, J., Ahokas, E., Markelin, L., Kaartinen, H., 2010. Automatic detection of buildings and changes in buildings for updating of maps. Remote Sens. 2, 1217–1248. https://doi.org/10.3390/rs2051217.

Mertes, C.M., Schneider, A., Sulla-Menashe, D., Tatem, A.J., Tan, B., 2015. Detecting change in urban areas at continental scales with MODIS data. Remote Sens. Environ. 158, 331–347. https://doi.org/10.1016/j.rse.2014.09.023.

Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto (Canada). PhD Thesis.

Mnih, V., Hinton, G., 2012. Learning to label aerial images from noisy data. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, pp. 567–574.

Moghadam, N.K., Delavar, M.R., Hanachee, P., 2015. Automatic urban illegal building detection using multi-temporal satellite images and geospatial information systems. Int. Arch. Photogramm. Remote SensSpat. Inf. Sci. - ISPRS Arch. 40, 387–393. https://doi.org/10.5194/isprsarchives-XL-1-W5-387-2015.

Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. IEEE Trans. Image Process. 16, 463–478. https://doi.org/10.1109/TIP.2006.888195.

Nielsen, A.A., Conradsen, K., Simpson, J.J., 1998. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: new approaches to change detection studies. Remote Sens. Environ. 64, 1–19. https://doi.org/10.1016/S0034-4257(97)00162-4.

Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113, 1276–1292. https://doi.org/10.1016/j.rse.2009.02.014.

Papadomanolaki, M., Vakalopoulou, M., Karantzalos, K., 2021. A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection. IEEE Trans. Geosci. Remote Sens. 1–18 https://doi.org/10.1109/TGRS.2021.3055584.

Peng, D., Bruzzone, L., Zhang, Y., Guan, H., DIng, H., Huang, X., 2021. SemiCDNet: a semisupervised convolutional neural network for change detection in high resolution remote-sensing images. IEEE Trans. Geosci. Remote Sens. 59, 5891–5906. https://doi.org/10.1109/TGRS.2020.3011913.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sens. 11, 1382. https://doi.org/10.3390/rs11111382.

Pesaresi, M., Corbane, C., Julea, A., Florczyk, A.J., Syrris, V., Soille, P., 2016. Assessment of the added-value of sentinel-2 for detecting built-up areas. Remote Sens. 8 https://doi.org/10.3390/rs8040299.

Qin, J., Fang, C., Wang, Y., Li, G., Wang, S., 2015. Evaluation of three-dimensional urban expansion: A case study of Yangzhou City, Jiangsu Province, China. Chin. Geogr. Sci. 25, 224–236. https://doi.org/10.1007/s11769-014-0728-8.

Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., 2014. ISPRS semantic labeling contest. ISPRS Leopoldshöhe Ger. 1, 4.

Saha, S., Bovolo, F., Bruzzone, L., 2019. Unsupervised deep change vector analysis for multiple-change detection in VHR images. IEEE Trans. Geosci. Remote Sens. 57, 3677–3693. https://doi.org/10.1109/TGRS.2018.2886643.

Schott, J.R., Salvaggio, C., Volchok, W.J., 1988. Radiometric scene normalization using pseudoinvariant features. Remote Sens. Environ. 26, 1–16. https://doi.org/10.1016/0034-4257(88)90116-2.

Seto, K.C., Güneralp, B., Hutyra, L.R., 2012. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. Proc. Natl. Acad. Sci. U. S. A. 109, 16083–16088. https://doi.org/10.1073/pnas.1211658109.

Shi, W., Zhang, M., Zhang, R., Chen, S., Zhan, Z., 2020. Change detection based on artificial intelligence: state-of-the-art and challenges. Remote Sens. https://doi.org/10.3390/rs12101688.

Shi, Y., Li, Q., Zhu, X.X., 2020. Building segmentation through a gated graph convolutional neural network with deep structured feature embedding. ISPRS J. Photogramm. Remote Sens. 159, 184–197.

Stokes, E.C., Seto, K.C., 2019. Characterizing urban infrastructural transitions for the sustainable development goals using multi-temporal land, population, and nighttime light data. Remote Sens. Environ. 234, 111430 https://doi.org/10.1016/j.rse.2019.111430.

Sun, C., Wu, J., Chen, H., Du, C., 2022. SemiSANet: a semi-supervised high-resolution remote sensing image change detection model using siamese networks with graph attention. Remote Sens. 14, 2801. https://doi.org/10.3390/rs14122801.

Sun, Y., Zhang, X., Huang, J., Wang, H., Xin, Q., 2022. Fine-grained building change detection from very high-spatial-resolution remote sensing images based on deep multitask learning. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2020.3018858.

Tang, X., Zhang, H., Mou, L., Liu, F., Zhang, X., Zhu, X.X., Jiao, L., 2022. An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning. IEEE Trans. Geosci. Remote Sens. 60 https://doi.org/10.1109/TGRS.2021.3106381.

Tang, Y., Huang, X., Zhang, L., 2013. Fault-tolerant building change detection from urban high-resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 10, 1060–1064. https://doi.org/10.1109/LGRS.2012.2228626.

Taubenböck, 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.

Tian, J., Cui, S., Reinartz, P., 2014. Building change detection based on satellite stereo imagery and digital surface models. IEEE Trans. Geosci. Remote Sens. 52, 406–417. https://doi.org/10.1109/TGRS.2013.2240692.

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481.

Van Etten, A., Lindenbaum, D., Bacastow, T.M., 2018. SpaceNet: A Remote Sensing Dataset and Challenge Series arXiv Prepr. arXiv1807.01232.

Voogt, J.A., Oke, T.R., 2003. Thermal remote sensing of urban climates. Remote Sens. Environ. 86, 370–384. https://doi.org/10.1016/S0034-4257(03)00079-8.

Wang, J., Ma, A., Zhong, Y., Zheng, Z., Zhang, L., 2022. Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: from airborne to spaceborne imagery. Remote Sens. Environ. 277, 113058 https://doi.org/10.1016/j.rse.2022.113058.

Wang, X., Li, P., 2020. Extraction of urban building damage using spectral, height and corner information from VHR satellite images and airborne LiDAR data. ISPRS J. Photogramm. Remote Sens. 159, 322–336. https://doi.org/10.1016/j.isprsjprs.2019.11.028.

Wang, Z., Zhang, Y., Luo, L., Wang, N., 2022. CSA-CDGAN: channel self-attention-based generative adversarial network for change detection of remote sensing images. Neural Comput. Appl. https://doi.org/10.1007/s00521-022-07637-z.

Weng, W., Zhu, X., 2021. INet: convolutional networks for biomedical image segmentation. In: IEEE Access. Springer, pp. 16591–16603. https://doi.org/10.1109/ACCESS.2021.3053408.

Xiao, X., Ma, L., 2006. Color transfer in correlated color space. In: Proceedings - VRCIA 2006: ACM International Conference on Virtual Reality Continuum and Its Applications, pp. 305–309. https://doi.org/10.1145/1128923.1128974.

Xie, Y., Zhu, J., Cao, Y., Feng, D., Hu, M., Li, W., Zhang, Y., Fu, L., 2020. Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering. IEEE JSel. Top. Appl. Earth Obs. Remote Sens. 13, 1842–1855.

Ye, S., Chen, D., Yu, J., 2016. A targeted change-detection procedure by combining change vector analysis and post-classification approach. ISPRS J. Photogramm. Remote Sens. 114, 115–124. https://doi.org/10.1016/j.isprsjprs.2016.01.018.

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64, 107–115. https://doi.org/10.1145/3446776.

Zhang, C., Yue, P., Tapete, D., Jiang, L., Shangguan, B., Huang, L., Liu, G., 2020. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. ISPRS J. Photogramm. Remote Sens. 166, 183–200. https://doi.org/10.1016/j.isprsjprs.2020.06.003.

Zhang, W., Lu, X., Li, X., 2018. A coarse-to-fine semi-supervised change detection for multispectral images. IEEE Trans. Geosci. Remote Sens. 56, 3587–3599. https://doi.org/10.1109/TGRS.2018.2802785.

Zhang, X., Xiao, P., Feng, X., Yuan, M., 2017. Separate segmentation of multi-temporal high-resolution remote sensing images for object-based change detection in urban area. Remote Sens. Environ. 201, 243–255. https://doi.org/10.1016/j.rse.2017.09.022.

Zhang, Y., Chen, G., Myint, S.W., Zhou, Y., Hay, G.J., Vukomanovic, J., Meentemeyer, R. K., 2022. UrbanWatch: a 1-meter resolution land cover and land use database for 22 major cities in the United States. Remote Sens. Environ. 278, 113106 https://doi.org/10.1016/j.rse.2022.113106.

Zhang, Z., Guo, W., Li, M., Yu, W., 2020. GIS-supervised building extraction with label noise-adaptive fully convolutional neural network. IEEE Geosci. Remote Sens. Lett. 17, 2135–2139. https://doi.org/10.1109/LGRS.2019.2963065.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.660.

Zheng, Z., Zhong, Y., Wang, J., Ma, A., Zhang, L., 2021. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: from natural disasters to man-made disasters. Remote Sens. Environ. 265, 112636 https://doi.org/10.1016/j.rse.2021.112636.

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.

Zhu, Q., Liao, C., Hu, H., Mei, X., Li, H., 2021. MAP-net: multiple attending path neural network for building footprint extraction from remote sensed imagery. IEEE Trans. Geosci. Remote Sens. 59, 6169–6181. https://doi.org/10.1109/TGRS.2020.3026051.