ISPRS Journal of Photogrammetry and Remote Sensing 188 (2022) 157–176

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images

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

Keywords:
Weakly supervised learning
Green plastic cover segmentation
Image-level label
High-resolution remote sensing

ABSTRACT

Green plastic cover (GPC) is a kind of green plastic fine mesh primarily used for covering construction sites and mitigating large amounts of dust during construction. Accurate GPC detection is vital for monitoring urban environment and understanding urban development. Convolutional neural network (CNN)-based segmentation methods are widely used for detecting object extents, while they rely on high-quality pixel-level labels with high acquisition cost. In this regard, weakly supervised learning can achieve pixel-level segmentation using only image-level labels, by first generating the class activation map (CAM) to obtain initial pixel-level labels and then applying the CNN-based segmentation methods to detect object extents. However, these initial labels are usually incomplete and noisy, caused by the local high response property of CAM. Moreover, the CNN-based segmentation methods often lead to blurry object boundaries due to the gradual down-sampling of feature maps, and meanwhile suffer from the class imbalance problem in real scenarios. Given these problems, we introduce weakly supervised learning into GPC detection to lower the label acquisition cost. Furthermore, to improve the completeness and correctness of initial labels and mitigate the blurry boundary problem, we propose a coarse-to-fine weakly supervised segmentation method (called CFWS), consisting of three steps: 1) object-based label extraction; 2) noisy label correction; and 3) boundary-aware semantic segmentation. Moreover, to alleviate the class imbalance problem, we propose a classification-then-segmentation strategy and integrate it into the CFWS to detect GPC. We test the CFWS on two datasets from Google Earth and Gaofen-2 high-resolution images, respectively. The results show that the CFWS obtains more complete GPCs and effectively retains boundaries on both datasets compared to existing state-of-the-art methods. In real scenarios, the classification-then-segmentation strategy significantly reduces a large number of false alarms generated by direct segmentation. These findings confirm that the CFWS holds great potentials for large-scale GPC detection and urban environmental monitoring. The source code will be available at https://github.com/lauraset/Coarse-to-fine-weakly-supervised-GPC-segmentation.

1. Introduction

Globally, urban areas are expanding, mainly influenced by population growth, urban–rural migration, and wealth growth (van Vliet et al., 2017). For example, since the reform and opening up, China is experiencing rapid urban development, accompanied by a large amount of urban infrastructure construction (Li et al., 2017; Yu, 2021). Urban construction usually produces large amounts of bare land, resulting in serious urban dust. Urban dust contains inhalable suspended particles, such as PM10 and PM2.5, which seriously pollute the air and thus endanger human health (Jiang et al., 2018; Yang et al., 2020). In this context, the environmental protection department requires that construction sites should be covered with green plastic cover (GPC). GPC is a kind of green plastic fine mesh and its color is environmentally friendly. The use of GPC can effectively alleviate urban dust pollution and meet environmental requirements. Therefore, accurate GPC detection is vital for monitoring urban environment and understanding urban development. However, to the authors' knowledge, few studies focus on GPC detection.

High-resolution remote sensing images, e.g., WorldView, IKONOS, ZY-3, and GF-2, offer an effective tool for GPC detection, and their rich spatial details make fine-scale urban observations possible (Bellens

* 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.isprsjprs.2022.04.012
Received 27 December 2021; Received in revised form 31 March 2022; Accepted 14 April 2022
Available online 21 April 2022

et al., 2008; Huang et al., 2020; Taubenböck et al., 2018). Recently, convolutional neural networks (CNNs), a kind of deep learning networks, can automatically learn multi-level features from high-resolution images, and have been widely applied to urban-related studies (Cao and Huang, 2021; Srivastava et al., 2019; Volpi and Tuia, 2016). For semantic segmentation tasks, i.e., assigning a category label to each pixel, the performance of CNNs heavily relies on high-quality pixel-level labels to optimize the network parameters. However, acquiring pixel-level labels is generally time-consuming and labor-intensive. One possible solution is to use weak labels, such as image-level labels (i.e., one image corresponds to one or multiple category labels) (Kolesnikov and Lampert, 2016), point labels (Bearman et al., 2016), scribbles (Lin et al., 2016), and bounding boxes (Khoreva et al., 2017). Since image-level labels are easier to obtain compared to other ones, this study focuses on weakly supervised semantic segmentation using image-level labels.

Image-level labels only provide information about the class of objects contained in the image, without specifying the spatial location of the objects, making it difficult for weakly supervised segmentation to achieve the same results as fully supervised segmentation with pixel-level pixels. The image-level weakly supervised segmentation method usually consists of two steps: 1) training image classification networks (e.g., CNNs) with image-level labels to obtain the class activation map (CAM) with the capability of object localization (Zhou et al., 2016), and then generating pseudo labels by the CAM thresholding; 2) using the pseudo labels to train a conventional semantic segmentation network to obtain object regions (Chan et al., 2021). The key of this method is how to obtain complete objects. In this regard, plenty of approaches have been widely explored in the computer vision domain (Ahn et al., 2019; Huang et al., 2018; Jo and Yu, 2021; Kolesnikov and Lampert, 2016; Wang et al., 2020). Although these methods perform well on natural images, they cannot be blindly applied to high-resolution remote sensing images (Chan et al., 2021), since the latter exhibit larger spectral and spatial heterogeneity, have more ambiguous boundaries, and are more sensitive to atmosphere and solar illumination. Therefore, it is essential to design weakly supervised semantic segmentation methods suited for high-resolution remote sensing images.

Given the characteristics of images and research objects, existing studies have successfully applied image-level weakly supervised segmentation techniques to high-resolution remote sensing images (Ali et al., 2020; Chen et al., 2020; Li et al., 2021; Zhang et al., 2020, 2021). For instance, Nivaggioli and Randrianarivo (2019) successfully adapted the AffinityNet and the random walk algorithm (Ahn and Kwak, 2018) for expanding initial labels from CAM thresholding to land cover classification. J. Chen et al. (2020) designed a super-pixel pooling layer to retain building boundaries, and then trained a boundary-aware building segmentation network by using initial labels from CAM thresholding. Li et al. (2021) obtained the initial building labels by CAM thresholding and CRF (conditional random field) segmentation (Krähenbühl and Koltun, 2011), and then applied them to optimize a building segmentation network. Although these methods are effective in detecting relevant objects, they cannot be directly used to green plastic cover (GPC) detection. Specifically, in this study, GPC is sparsely distributed, irregularly shaped, and usually green in color. These characteristics significantly distinguish GPC from other classes such as buildings and residential areas. However, few studies have focused on GPC extraction. Moreover, when we apply the aforementioned weakly supervised methods into GPC extraction, they exhibit three limitations:

1) *The local high response property of CAM and the potential noisy labels*. At the stage of obtaining initial labels, CAM generally has a high response to the most discriminative regions (Kolesnikov and Lampert, 2016; Zhou et al., 2016), making it difficult to identify complete objects. The direct use of the initial labels from CAM can introduce noise and reduce the generalization performance of networks, due to the fact that deep networks have the powerful learning ability and can even fit corrupted labels (Song et al., 2020). Thus, some researchers have used the noisy label learning technique to mitigate the impact of label noise, such as noise-robust loss function (Oh et al., 2021; Zhang and Sabuncu, 2018), sample selection via multi-network learning (Han et al., 2018), label correction (Dong et al., 2021; Yi and Wu, 2019). Although the noisy label learning can alleviate the impact brought by label noise to some extent, it relies heavily on the quality of initial labels and hence obtaining complete objects is still difficult.
2) *The blurry boundary problem in the semantic segmentation.* Commonly used semantic segmentation networks are based on the encoder-decoder structure, such as U-Net (Ronneberger et al., 2015), SegNet (Badrinarayanan et al., 2017), and DeepLabv3+ (Chen et al., 2018). However, the gradual down-sampling of feature maps in the encoder can lose high-frequency spatial details, leading to blurry object boundaries. In this regard, a common coping strategy is to explicitly guide the model to focus on boundaries using a reference boundary map (Jung et al., 2021; A. Li et al., 2021; Marmanis et al., 2018). Although this strategy can improve the object boundary accuracy, it usually relies on high-quality pixel-level boundary labels, which are difficult to obtain in the case of weakly supervised learning with only coarse labels. For the green plastic cover (GPC), owing to its irregular shape, the loss of spatial details and coarse labels (e.g., image-level labels) increase the difficulty of accurately identifying GPC.
3) *The class imbalance problem*. In real application scenarios, semantic segmentation models can be affected by the class imbalance problem. Taking GPC for example, it is mainly distributed in the inner city and the urban–rural zone with diverse and complex backgrounds, and merely occupies relatively small areas. These properties easily lead to numerous false alarms and omissions in the segmentation model, which significantly increases the intensity of manual post-processing. To mitigate the class imbalance problem, most of existing algorithms focus on the model training stage, such as under-sampling, oversampling, and cost sensitive learning (Buda et al., 2018; Johnson and Khoshgoftaar, 2019; Kellenberger et al., 2018; Yessou et al., 2020). Note that these methods can be directly applied to real scenarios once the model training is completed, which may cause numerous false alarms and omissions when the distributions of the training set and the real scenarios are obviously different. However, this issue is less considered in existing literature. In this study, the sparse spatial distribution of GPC makes it necessary to consider the class imbalance problem.

Given these problems, we propose a coarse-to-fine weakly supervised segmentation method (called CFWS), and apply it to GPC extraction using high-resolution remote sensing images. We present an effective way for the environmental monitoring of urban construction sites. The CFWS mainly consists of three steps: 1) we generate the initial pixel-level GPC labels by using a classification network with image-level labels (i.e., coarse labels) and an unsupervised image segmentation technique; 2) subsequently, we perform noisy label correction on the initial labels to remove potential label noise and obtain the refined labels (i.e., fine labels); 3) we use the fine labels to train a GPC segmentation network with a boundary-aware joint loss function. Moreover, for real scenarios, we design a classification-then-segmentation strategy, i.e., the classification network is first used to obtain the candidate GPC regions, and semantic segmentation is then performed on the candidate GPC regions for detecting the pixel-level GPC extents. The main contributions of this study are summarized below:

- For the first time, the weakly supervised semantic segmentation technique is applied to GPC recognition, lowering the acquisition cost of pixel-level labels.
- A coarse-to-fine pixel-level label generation method is proposed to alleviate the local high response property of CAM and the potential label noise problem.

- A boundary-aware semantic segmentation network is proposed to mitigate the boundary ambiguity.
- The classification-then-segmentation strategy is designed to reduce the effect of the class imbalance problem in real scenarios.

The remainder of this paper is organized as follows. Section 2 describes the collected dataset. Section 3 introduces the proposed method. Then, the experimental results and discussions are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper.

2. Dataset description

To test the proposed CFWS sufficiently, we created two datasets, namely, Google Earth green plastic cover detection dataset (GE-GPC) and Gaofen-2 green plastic cover detection dataset (GF2-GPC) (see Table 1 and Figs. 1-2). Each dataset contains three types of samples:

1) Classification samples. Each classification sample corresponds to a single category, i.e., GPC or non-GPC. For each category, we randomly divided classification samples into training, validation and test sets at a ratio of 6:2:2. To reduce the spatial autocorrelation of the classification samples (Chen and Wei, 2009), we set the minimum spatial distance between samples to the width of the samples.
2) Segmentation samples. Each pixel in segmentation samples is assigned as an individual label. The segmentation samples consist of 80 GPC samples randomly selected from the test set of classification samples. We manually annotated them to obtain pixel-wise GPC labels for validating the semantic segmentation model.
3) Large-area test samples. We collected two images with wide coverage, and they exhibit different geographic landscapes and urban development levels. We manually interpreted the corresponding pixel-wise GPC labels. The large-area samples were used considering that they have a larger coverage and are more complete, compared to the cropped classification samples, and allow the test of the proposed method in real scenarios with class imbalance problem.

2.1. GE-GPc

GE-GPC was acquired from Google Earth images (GE) (https://www.google.com/earth), with the spatial resolution of 1 m and containing red, green, and blue (RGB) visible bands (Fig. 1). The classification samples consist of 3,000 GPC and 30,000 non-GPC. The size of each sample is 512 × 512 pixels. All samples are randomly distributed in urban areas of China, and exhibit distinct spectral, spatial, and contextual differences. For the two large-scale test samples, one is located in Yishui County, Linyi City, Shandong Province, with an image size of 10,583 × 10,570 pixels, and the other is located in Zhangjiakou City, Hebei Province, with an image size of 10,583 × 10,595 pixels. All the GE images used above were accessed in September 2021. We used the open source QGIS software and python language to batch download the GE images, and in order to facilitate further research, the source code will be available at https://github.com/lauraset/Coarse-to-fine-weakly-supervised-GPC-segmentation.

Table 1
Composition of GE-GPC and GF2-GPC datasets.

Dataset	Classification samples	Segmentation samples	Large-area test samples	Spatial resolution
GE-GPC	3,000 GPC 30,000 non-GPC	80	Yishui: 10,583 × 10,570 pixels Zhangjiakou: 10,583 × 10,595 pixels	1 m
GF2-GPC	2,042 GPC 19,000 non-GPC	80	Beijing: 27,988 × 27,248 pixels Tianjin: 27,976 × 27,220 pixels	1 m

2.2. GF2-GPc

GF2-GPC was obtained from the Gaofen-2 images (GF-2) (Fig. 2). The GF-2 satellite is China's first sub-meter civil high-resolution optical satellite (Zhou et al., 2021), which was launched in August 2014 and carried a 1-m panchromatic camera and a 4-m multispectral camera (providing four spectral bands in RGB and near infrared). We collected 21 scenes of GF-2 images with less than 10% cloud coverage from the China Resources Satellite Application Center (https://www.cresda.com). Each scene consists of one panchromatic image and one multispectral image. All scenes cover 10 large and medium-sized Chinese cities, and they were acquired between 2015 and 2018 (Table 2). We preprocessed all the images with radiometric correction, atmospheric correction, *ortho*-rectification, and image-to-image registration, and then enhanced the spatial resolution of the multispectral images using the panchromatic images to generate 1-m multispectral images. We used the NNDiffuse pan-sharpening method, since it shows satisfactory effect on GF-2 images (Zhang et al., 2019). A total of 2,042 GPC and 19,000 non-GPC samples were cropped from the 19 GF-2 images and were used as the classification samples. The size of each sample was set to 256 × 256 pixels. The remaining two GF-2 images were used as large-area test samples, one in Beijing with an image size of 27,988 × 27,248 pixels, acquired on February 12, 2015, and the other in Tianjin with an image size of 27,976 × 27,220 pixels, acquired on June 20, 2015.

3. Methodology

The proposed CFWS consists of three steps: 1) object-based label extraction (Section 3.1); 2) noisy label correction (Section 3.2); and 3) boundary-aware GPC segmentation (Section 3.3). The workflow is shown in Fig. 3. For real scenarios, we develop a classification-then-segmentation strategy (Section 3.4). Details are provided below.

3.1. Object-based label extraction

We proposed an object-based label extraction method for obtaining initial pixel-level labels, to mitigate the local high response property of CAM. The method consists of the following three steps: 1) pixel-level CAM extraction, 2) object-based CAM generation, and 3) adaptive thresholding.

Step 1. pixel-level CAM extraction. We trained the image classification network using image-level labels, i.e., GPC and non-GPC classification samples, to generate pixel-level CAM that can indicate the discriminative but coarse GPC regions. Specifically, image classification networks, such as CNNs, perform layer-wise feature extraction on the input image, and then judge the corresponding category. In this study, we used RegNetY-4.0GF from the RegNet family (Radosavovic et al., 2020) as the classification network (Fig. 4), since it can significantly save computational resources while maintaining a high classification accuracy. RegNet is composed of three parts: a stem (containing a stride-two 3 × 3 convolution layer with 32 output channels), the body of the network for performing computations, and a head (containing average global pooling and fully connected layers) for predicting the output category (see Fig. 4). The body consists of a series of progressively downsampled stages, and each stage contains a series of blocks. For RegNet-Y series, each block is composed of standard residual bottleneck blocks with group convolution (Xie et al., 2017) and SE (Squeeze-and-Excitation) (Hu et al., 2018) attention mechanism. In this study, we kept the stem and the body unchanged, but added a dropout layer (with a dropout rate of 0.2) before the fully connected layer of the head to prevent the network from overfitting (Srivastava et al., 2014). Since this study focuses on GPC classification, i.e., binary classification, we set the number of channels of the prediction score (denoted by S) from the

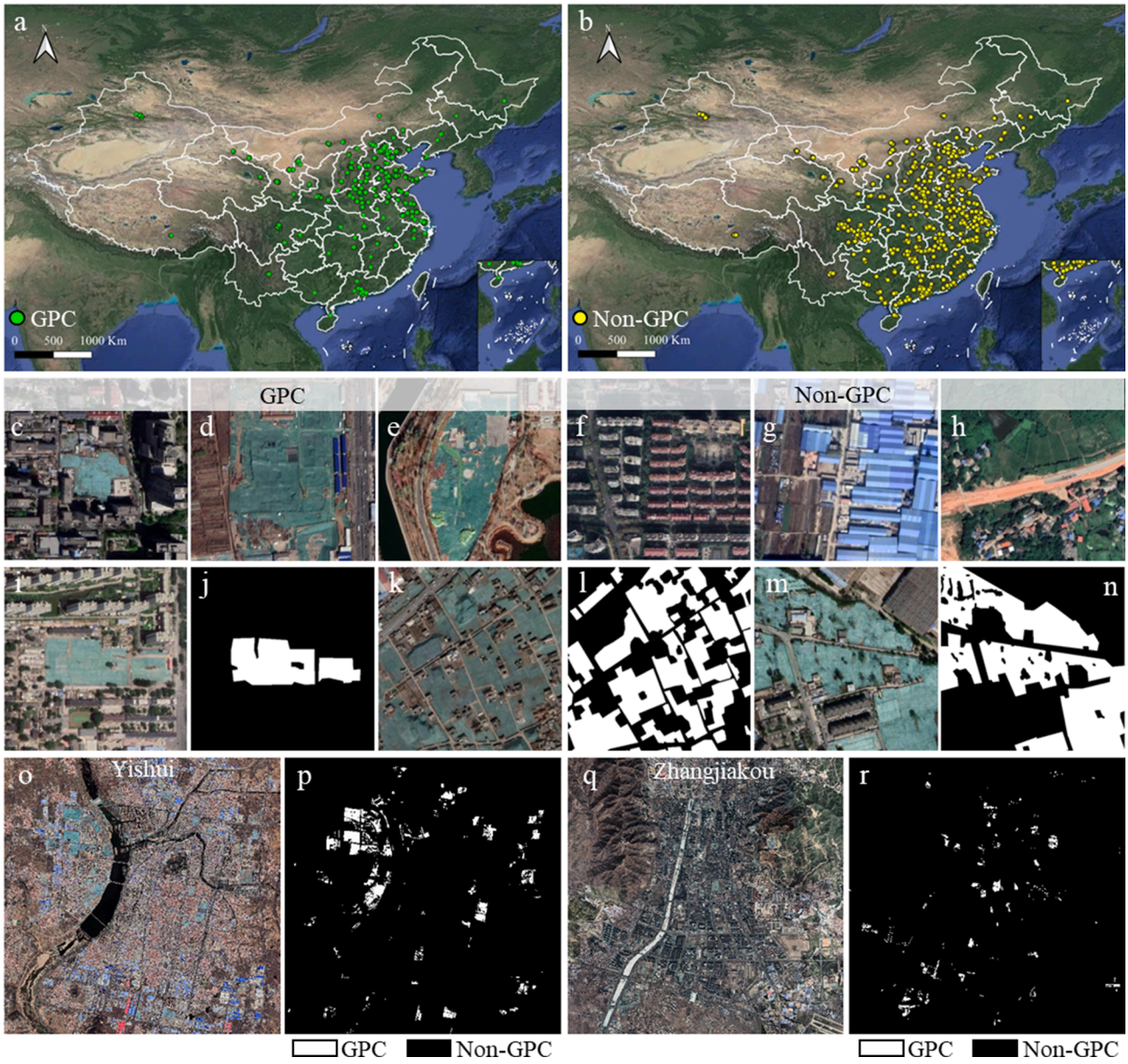

Fig. 1. The GE-GPC dataset. (a) and (b) are the spatial distribution of all GPC and non-GPC samples, respectively (Base map: google images). (c-e) are image-level GPC samples. (f-h) are image-level non-GPC samples. (i-j), (k-l), and (m-n) denote GPC samples and their corresponding pixel-level labels. (o-p) and (q-r) are the large-area test samples containing Google Earth images and their corresponding pixel-level labels.

network output to 2, and then normalized S using the Softmax activation function to obtain the probability p_c that the input image belongs to category c:

$$p_c = \frac{\exp(S_c)}{\sum_{c=1}^{C} exp(S_c)} \tag{1}$$

where for binary classification, $C = 2$ and exp is the exponential function. Under the supervision of image-level labels, we used the cross-entropy loss function L_{CE} to optimize the network parameters:

$$L_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} q_{n,c}\cdot log\left(p_{n,c}\right) \tag{2}$$

where $q_{n,c}$ denotes the n-th reference label (one-hot encoding) of category c. For this study, $q_{n,c}$ is equal to [0, 1] for GPC, and [1, 0] for non-GPC. N is the total number of samples and log denotes the logarithmic function. Using L_{CE}, we trained RegNetY-4.0GF with the Adam optimizer and the pretrained weights from ImageNet (Jia Deng et al., 2009). To achieve the image normalization, the mean of each band is subtracted and then divided by the standard deviation of each band. The number of epochs was set to 80. The learning rate was initially set to 0.001 and decreased by 0.1 at the 40th and 60th epochs. To avoid overfitting, we used the data augmentation, including horizontal or vertical flipping, random rotation, and random grid shuffle (i.e., the image was divided evenly into 4 blocks and then the order of each block was randomly shuffled). The data augmentation approach remains the same for all subsequent network training.

Grad-CAM++ (Chattopadhay et al., 2018) was generated from the well-trained RegNetY-4.0GF, and was used as the CAM in this paper to indicate the spatial location of GPC. The reason why Grad-CAM++ was chose is that it has a better localization ability than Grad-CAM (Selvaraju et al., 2017), and it can be applied to any CNN-based network compared to the original CAM (Zhou et al., 2016). At the spatial location (i, j), the Grad-CAM++ for class c (denoted by L_{ij}^{c}) is computed as:

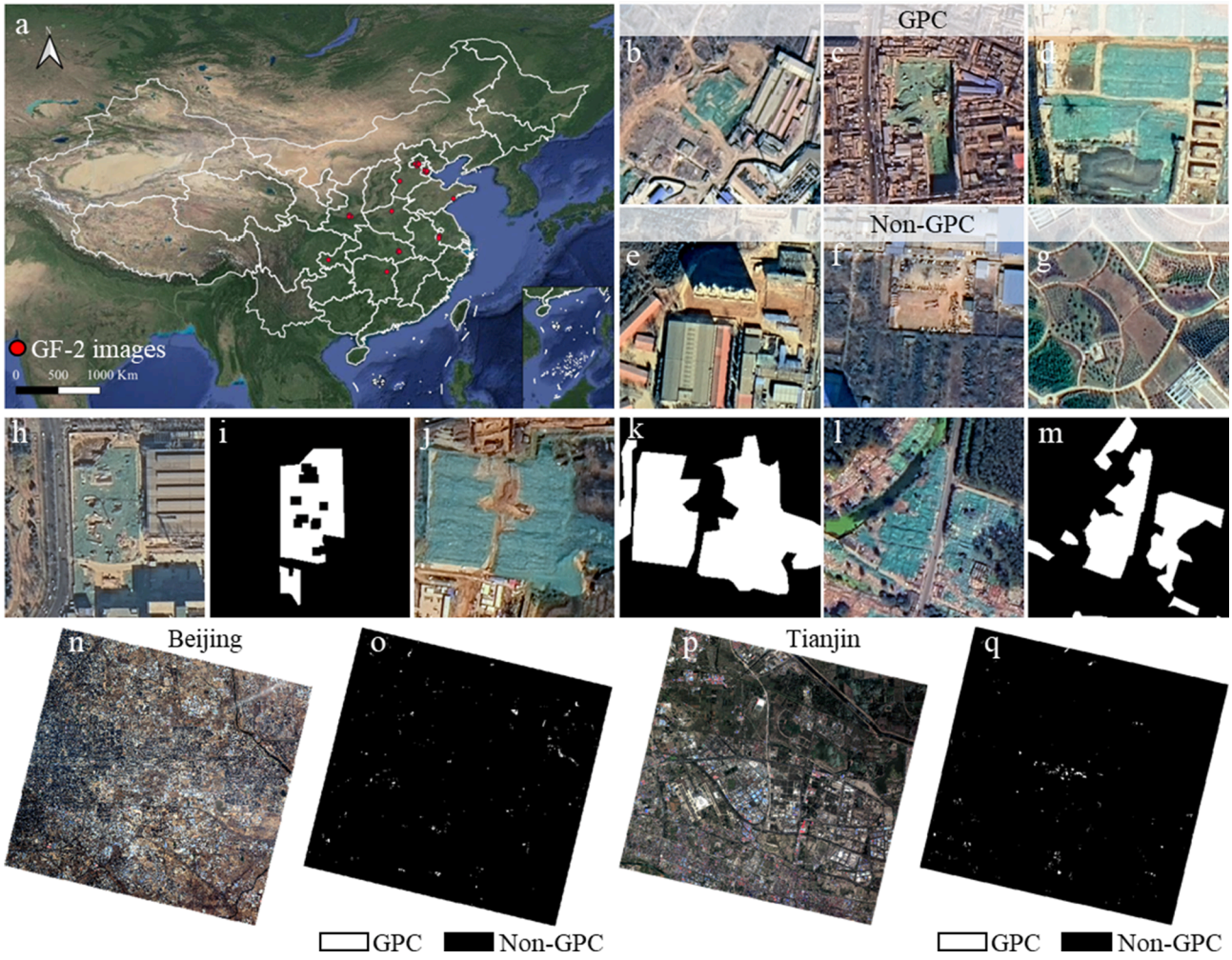

Fig. 2. The GF2-GPC dataset. (a) is the spatial distribution of GF-2 images (Base map: google images). (b-d) are image-level GPC samples. (e-g) are image-level non-GPC samples. (h-i), (j-k), and (l-m) denote GPC samples and their corresponding pixel-level labels. (n-o) and (p-q) are the large-area test samples containing GF-2 images and their corresponding pixel-level labels.

Table 2
The GF-2 images used in this study.

City	Date	Number
Zhengzhou	20,180,416	1
Shijiazhuang	20,150,304	1
Tianjin	20,150,620	3
Changsha	20,180,322	1
Nanjing	20,171,009	2
Qingdao	20,180,419	1
Chongqing	20,180,402	1
Beijing	20,150,217	4
	20,150,212	1
Wuhan	20,160,901	2
	20,161,208	1
Xian	20,161,130	1
	20,170,308	2

$$L_{ij}^{c} = \mathrm{ReLU}\left(\sum_{k=1}^{K} w_k^c \cdot A_{ij}^k\right) \forall\{i,j|i \in [1,H], j \in [1,W]\} \tag{3}$$

where A_{ij}^k denotes the feature map of the last convolutional layer and k denotes the k-th channel. H, W and K denote the height, width and total number of channels of the feature map, respectively. ReLU is rectified linear unit ($ReLU(x) = max(0, x)$), and can highlight features with a positive impact on the class of interest. The weight w_k^c is calculated as:

$$w_k^c = \sum_{i=1}^{H}\sum_{j=1}^{W} \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial Y_c}{\partial A_{ij}^k}\right) \tag{4}$$

where $\frac{\partial Y_c}{\partial A_{ij}^k}$ denotes the gradient of the prediction score Y^c, relative to the feature map A_{ij}^k. Here, Y^c needs to be a smooth function, so we apply the exponential function on the prediction score S from the fully connected layer to obtain the differentiable function $Y_c = exp(S_c)$. The weight α_{ij}^{kc} is given by:

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 Y_c}{\left(\partial A_{ij}^k\right)^2}}{2\frac{\partial^2 Y_c}{\left(\partial A_{ij}^k\right)^2} + \sum_{a=1}^{H}\sum_{b=1}^{W} A_{ab}^k \left\{\frac{\partial^3 Y_c}{\left(\partial A_{ij}^k\right)^3}\right\}} \tag{5}$$

Notice that the obtained Grad-CAM++ is relatively coarse and has the same size as the feature map of the last convolution layer. We upsampled this activation map by bilinear interpolation to make it the same size as the input image for subsequent processing.

Step 2. Object-based CAM generation. Considering the low resolution and local high response property of CAM (Step 1), we generated object-based CAM by the unsupervised image segmentation technique (UIS) (Kanezaki, 2018).The basic criteria of UIS include: 1) pixels with similar features are desired to belong to the same class; 2) spatially contiguous pixels are likely to belong to the same class; and 3) the number of

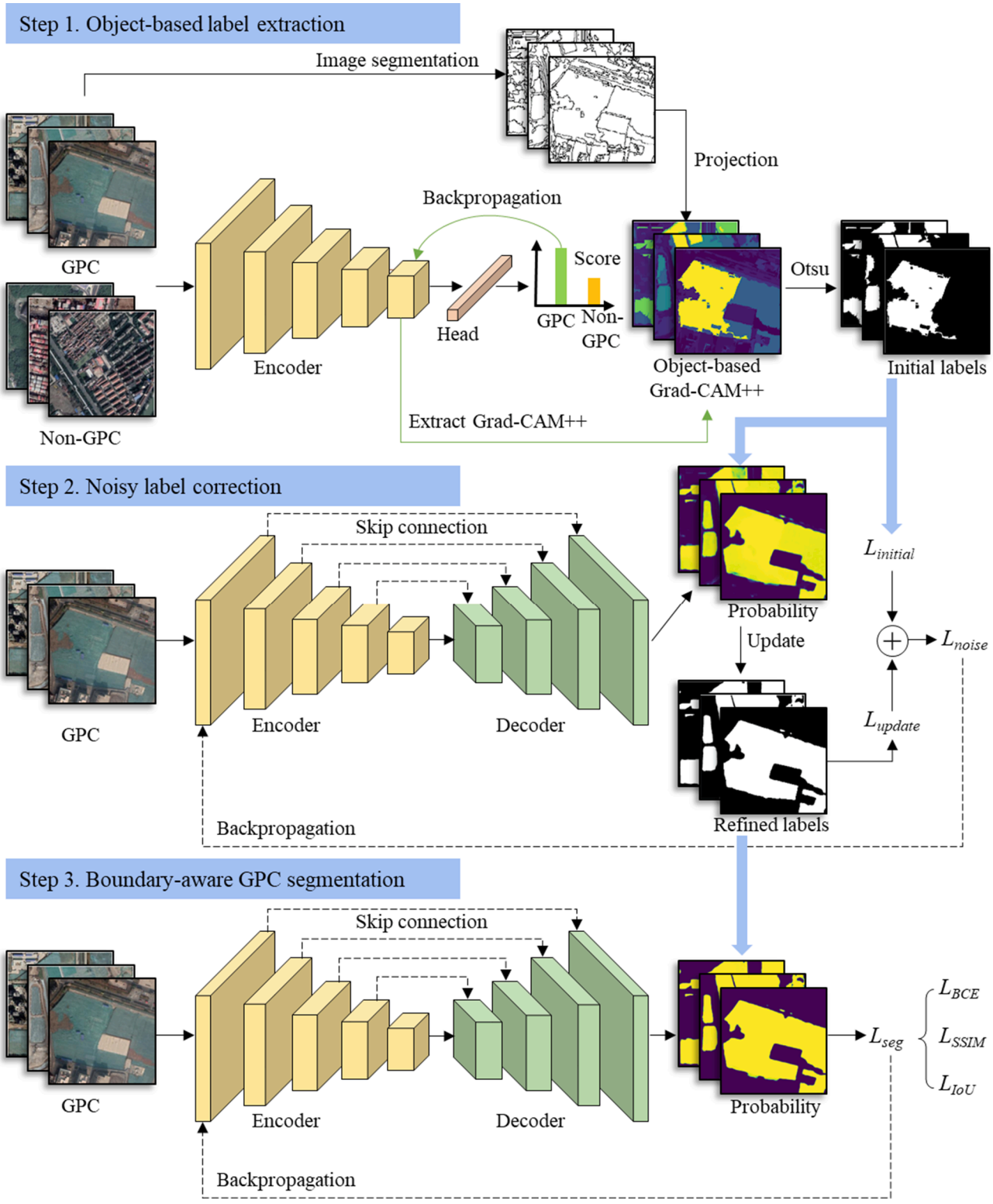

Fig. 3. The workflow of the proposed CFWS.

segmented objects should be large enough. Under the basic criteria, UIS consists of three steps: 1) firstly, the input image is clustered to generate super-pixels; 2) then, CNN network is applied on the input image to obtain the predicted labels; 3) finally, the reference label for each super-pixel is assigned as the predicted label with the most occurrences, and is used to calculate loss function. This process is iterated to update the network parameters. UIS was applied on each image to generate the segmentation map, with which we obtained the CAM value of each object by calculating the average value within each object.

Step 3. Adaptive thresholding. For the object-based CAM of each classification sample, we calculated the binarization threshold by the adaptive Otsu method (Otsu, 1979). If the CAM value of an object is greater than the threshold, this object is assigned as target (i.e., GPC), and as background (non-GPC) otherwise. All GPC classification samples

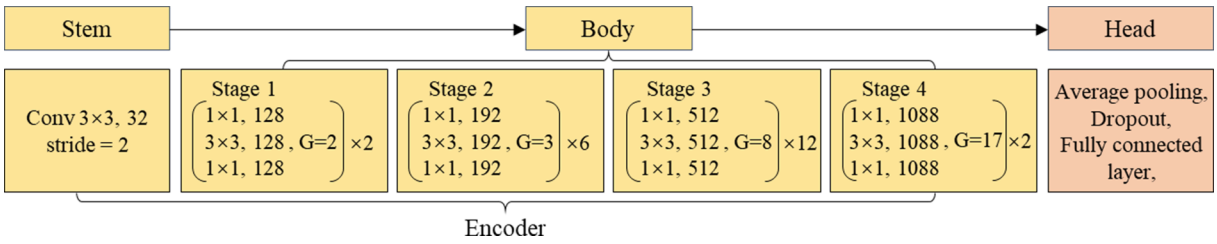

Fig. 4. The structure of the RegNetY-4.0GF network. Each convolution layer is denoted as (filter size, #output channels). Each stage consists of a sequence of identical residual bottleneck blocks with group convolution. "G" represents the number of groups.

were binarized to generate the initial pixel-level labels. Examples of object-based label extraction is illustrated in Fig. 5, and its validity is analyzed in Section 5.1.

3.2. Noisy label correction

Noisy label correction was designed to reduce the potential noise in the initial labels (Section 3.1) and thus mitigate the impact of noisy labels on the segmentation network. The method is easily scalable since it does not need auxiliary clean datasets or prior knowledge of noisy labels, and includes two steps:

Step 1. Initial network training. Although deep networks have powerful feature learning ability, they can easily fit random noise, which will significantly degrade the network performance. However, an interesting phenomenon is that deep networks tend to learn correctly labeled samples in the early stage and start learning mislabeled samples only in the later stage (Arazo et al., 2019). Moreover, when maintaining a high learning rate, deep networks do not easily fit the incorrect samples (Tanaka et al., 2018). In this context, we trained the U-Net (Fig. 6) using initial pixel-level labels with a fixed high learning rate and a few epochs, and used it as the initial segmentation network. U-Net is a widely-used encoder-decoder structure (Ronneberger et al., 2015). The encoder progressively compresses feature maps to extract high-level semantic features, while the decoder recovers the spatial information of feature maps, and finally generates the prediction result of equal size to the input. To enhance the spatial details of the prediction result, the feature maps of the encoder are added to the decoder by skip connection. In this study, RegNetY-4.0GF was used as the encoder (consistent with Section 3.1), and the decoder remained the original U-Net. The learning rate was fixed at 0.001 and the number of epochs was set to 10. The number of channels of the prediction result (denoted by S) was set to 1. The prediction result was normalized to the range [0,1] by the Sigmoid function ($p = \frac{1}{1+exp(-S)}$). We used Binary Cross Entropy (BCE) as the loss function:

$$L_{BCE} = -\frac{1}{N \times H \times W} \sum_{n=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(q_{n,ij} \cdot log p_{n,ij} + \left(1 - q_{n,ij}\right) \cdot log\left(1 - p_{n,ij}\right)\right) \tag{6}$$

where $q_{n,ij}$ is the reference label (i.e., initial labels) of the n-th sample at the spatial location (i, j), which is equal to 1 for GPC and 0 for non-GPC. H and W denote the height and width of each sample, respectively, while N denotes the total number of samples. Note that Eq. (2) is an extension of Eq. (6) over multiple classes, and in this paper, the former is used for image-level classification while the latter for pixel-level segmentation.

Step 2. noisy label correction. We used the initial network (Step 1) as the training starting point, and performed both network updating and initial label correction, of which the criterion is the predicted probability

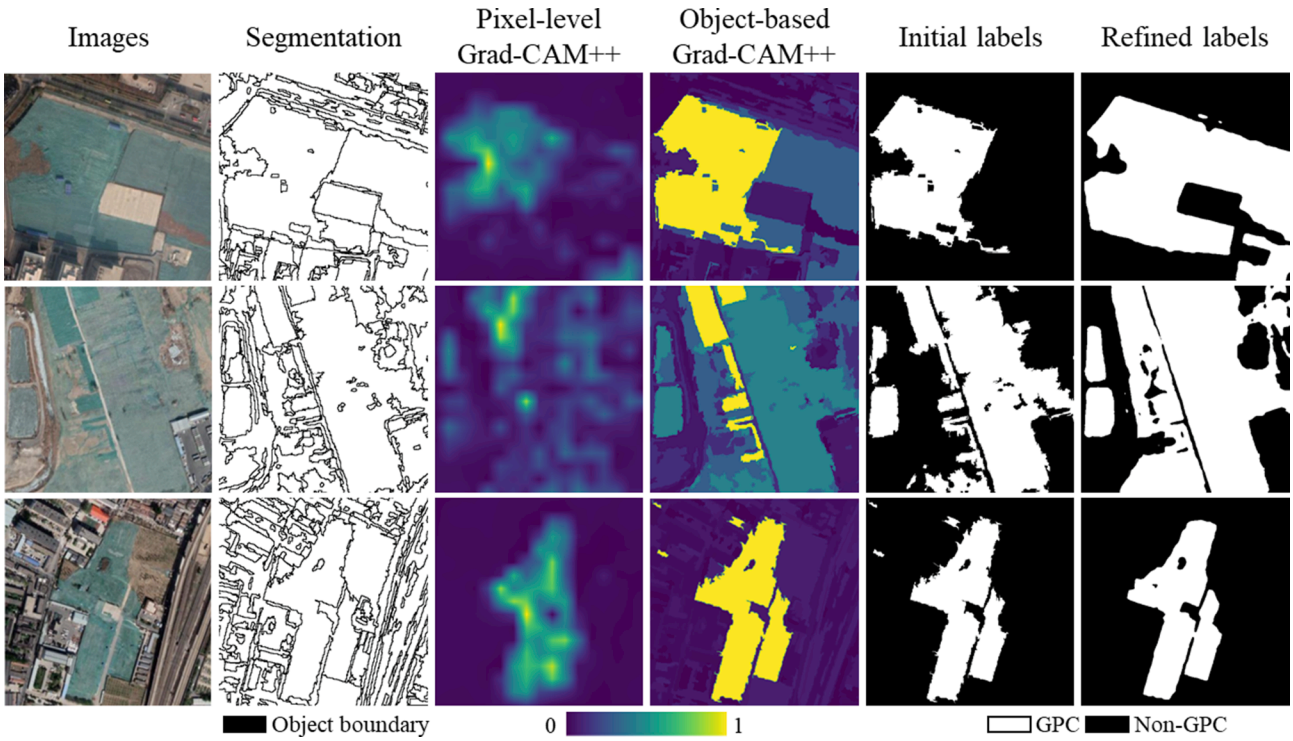

Fig. 5. Illustration of object-based label extraction and noisy label correction.

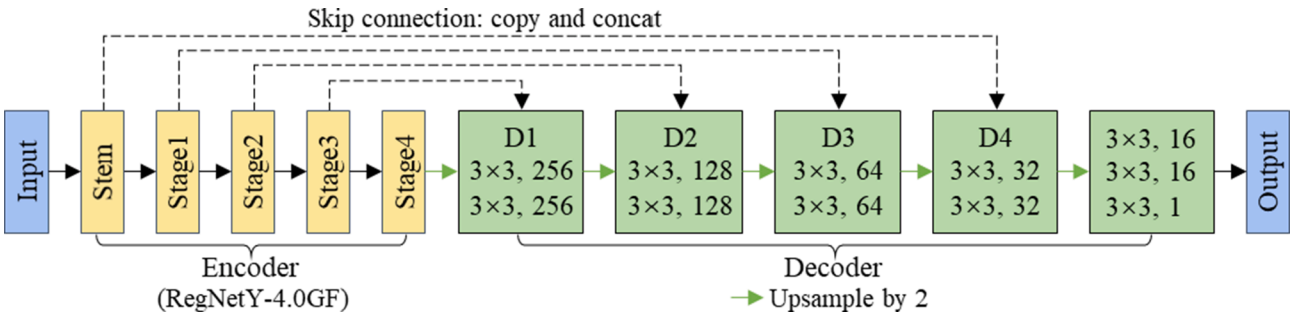

Fig. 6. The structure of the U-Net used in this study.

of the network. Specifically, for each pixel, if its predicted probability is less than a threshold, the corresponding initial label is corrected to the predicted label; otherwise, it remains unchanged. In this study, the threshold was set to 0.5. Besides, considering that most of initial labels are correct, if we discard them directly, the predicted results of the network may completely deviate from the initial labels (Yi and Wu, 2019). Therefore, we proposed a joint loss function, taking into account both initial and updated labels:

$$L_{noise} = L_{update}(q,p) + \lambda \cdot L_{initial}(\widehat{q},p) \tag{7}$$

where $\widehat{q}$ denotes the initial label q after correction, and p is the probability predicted by the network. $L_{initial}$ and L_{update} denote the initial and the updated loss functions, respectively, both of which use BCE (Eq. (6)). λ is the balanced factor, and was set to 0.2. We trained U-Net with Eq. (7) for a total of 10 epochs at a fixed learning rate of 0.001, to prevent it from fitting on noisy labels. Finally, we obtained refined labels for further segmentation. Examples of the refined labels can be found in Fig. 5. In-depth analysis is presented in Section 5.2.

3.3. Boundary-aware GPC segmentation

We designed a boundary-aware joint loss function to alleviate the blurry boundary problem caused by the gradual down-sampling of feature maps. The loss function consists of binary cross entropy (BCE) (cf. Eq. (6)), structural similarity (SSIM) (Wang et al., 2004) and intersection over union (IoU) (Zhou et al., 2019), and is formulated as:

$$L_{seg} = L_{BCE} + L_{SSIM} + L_{IoU} \tag{8}$$

SSIM loss can capture local structural changes in the image, and it gives higher weights to boundaries, which helps improve the network's ability to detect boundaries:

$$L_{SSIM} = 1 - \frac{(2\mu_p\mu_q + C_1)(2\sigma_{pq} + C_2)}{\left(\mu_p^2 + \mu_q^2 + C_1\right)\left(\sigma_p^2 + \sigma_q^2 + C_2\right)} \tag{9}$$

where μ denotes the mean value and σ is the standard deviation. p represents the probability of being predicted as GPC, and $p \in [0,1]$. q is the reference label which is equal to 1 for GPC and 0 for non-GPC. $C_1 = 0.01^2$ and $C_2 = 0.03^2$ are used to avoid 0 in the denominator. Generally, SSIM is computed on each local window, and then the average value of all windows is taken as the SSIM of the whole image.

IoU loss is often used to measure the similarity of two arbitrary shapes, which encodes the shape attributes of the target (e.g., length and width) into area attributes, facilitating the enhancement of the network's ability to perceive the range of GPC. It is expressed as:

$$L_{IoU} = 1 - \frac{p \cdot q}{p + q - p \cdot q} \tag{10}$$

We used Eq. (8) to optimize U-Net (Fig. 6), and refined labels (Section 3.2) were used as reference. We used the pretrained weight from ImageNet (Jia Deng et al., 2009) to initialize the network parameters. For each dataset, the number of epochs was set to 35. The learning rate was initially set to 0.001 and decreased by 0.1 for every 15 epochs. The performance of the boundary-aware loss function is analyzed in Section 5.3.

3.4. Classification-then-segmentation strategy in real scenarios

For real scenarios, we propose a classification-then-segmentation strategy to alleviate a large number of false alarms that are generated by direct segmentation, and thus reduce the intensity of manual post-processing. Firstly, we used the well-trained GPC classification model (Section 3.1) to predict the whole image. Limited by memory, the classification model can only process small image blocks. Thus, we performed sliding half-overlapped window prediction throughout the whole image (Fig. 7(a)). Specifically, one image window (e.g., 512 × 512 pixels) is predicted at a time, the window is moved by half of the window width, and the average of the overlapping regions is taken as the final prediction value of that region. This strategy ensures the boundary continuity of prediction results. To reduce omissions, we used a low threshold (T) to obtain the potential GPC region. Specifically, if the probability (P) of the region being predicted as GPC exceeds T, the region is marked as the candidate GPC region, and as non-GPC otherwise. Then, we applied the well-trained GPC segmentation model (Section 3.3) to detect the pixel-wise GPC ranges on the candidate GPC regions (Fig. 7(b)). In this study, the threshold was set to 0.2 and its sensitivity analysis is presented in Section 5.4.

3.5. Accuracy assessment

Five accuracy metrics were selected to evaluate the performance of the proposed method on identifying GPC extents, including overall accuracy (OA), precision, recall, F1-score and intersection over union (IOU), since they are widely used to assess the accuracy of object segmentation (Ali et al., 2020; Guo et al., 2021). They are calculated as:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

where TP (true positive) represents the number of pixels correctly predicted as GPC, FP (false positive) represents the number of pixels incorrectly predicted as GPC, TN (true negative) represents the number of pixels correctly predicted as non-GPC, and FN (false negative) rep-

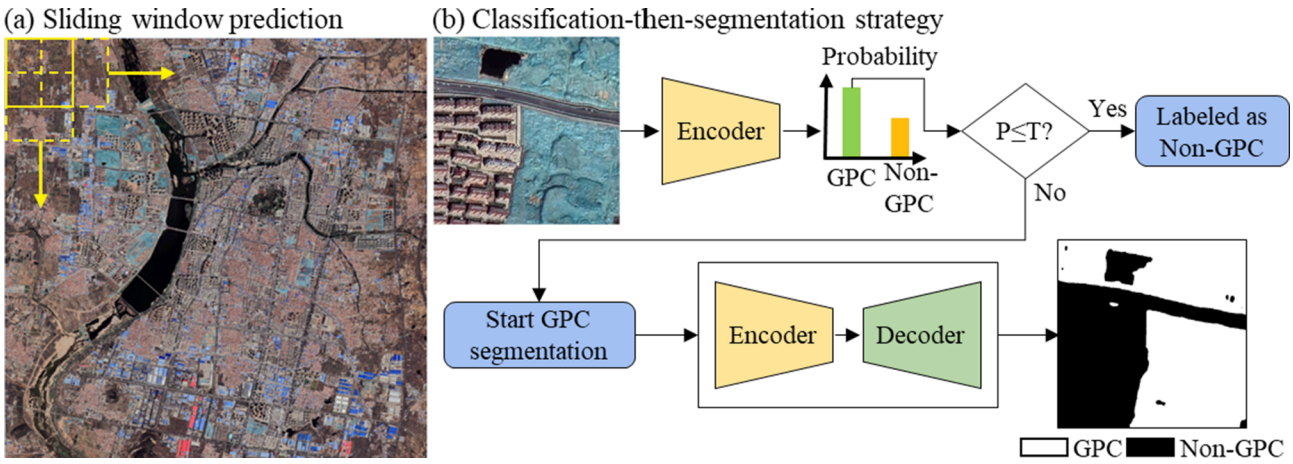

Fig. 7. Illustration of the sliding window prediction (a) and the classification-then-segmentation strategy (b).

resents the number of pixels incorrectly predicted as non-GPC. OA denotes the classification accuracy for all categories (i.e., GPC and non-GPC). The precision and recall of GPC correspond to the false alarm and omission rates, respectively. F1-score is the harmonic mean of precision and recall, and is a better measure than OA in the case of class imbalance. IoU is the ratio between the intersection of the predicted and the actual GPC extents and their union, and can measure the quality of GPC segmentation.

4. Results

4.1. Comparison with other methods

To validate the performance of the proposed CFWS, we compared three state-of-the-art weakly supervised semantic segmentation methods, namely SEC (seed, expand and constrain) (Kolesnikov and Lampert, 2016), IRNet (inter-pixel relation network) (Ahn et al., 2019) and TSWS (two-step weakly supervised segmentation) (Li et al., 2021). SEC and IRNet are initially designed for natural images, while TSWS is

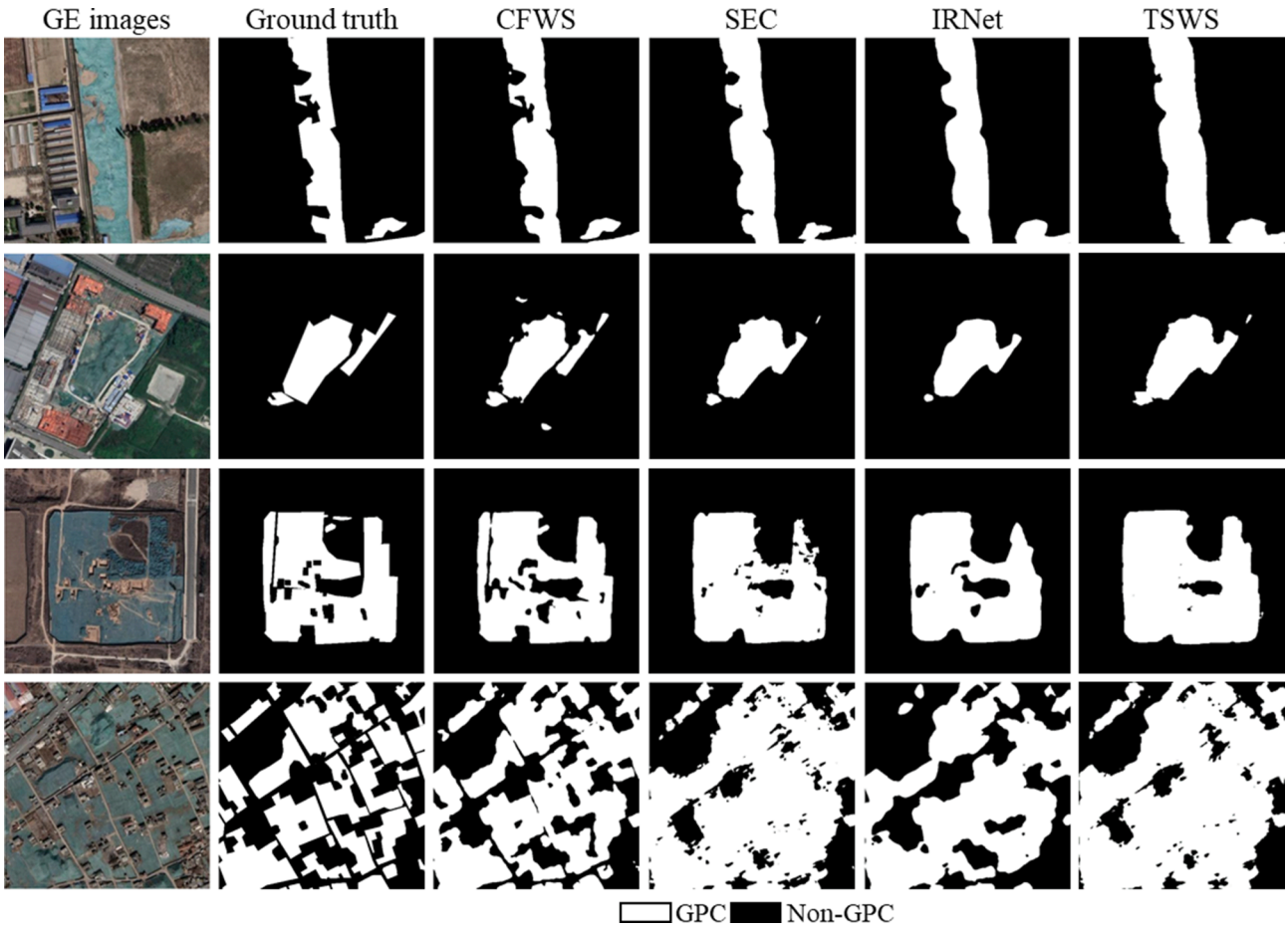

Fig. 8. Results of different weakly supervised segmentation methods on the segmentation samples of GE-GPC.

applied to building extraction with high-resolution remote sensing images. Specifically, SEC introduced three loss functions, namely seeding, expansion, and boundary constraints, into the semantic segmentation network to detect object extents. IRNet utilized the semantic affinity between pixels (defined as the similarity between the embedding vectors of pixels), to recognize the object extent. Li et al. (2021) fused the results of CAM thresholding and CRF (conditional random field) segmentation (Krähenbühl and Koltun, 2011) to generate initial building labels, and then combined the classification loss and the CRF loss to obtain relatively complete building regions. For fair comparison, classification (RegNetY-4.0GF) and segmentation (U-Net) networks used in SEC, IRNet, and TSWS were consistent with our method. We tested these methods on the segmentation samples of GE-CPC and GF2-GPC (Section 2).

In Fig. 8 and Fig. 9, we can observe that these GPCs have irregular shapes, and exhibit distinct spectral and contextual differences. Notice that the GF-2 image was fused by a 1-m panchromatic band and a 4-m multispectral band, and shows a lower quality than the GE image that is directly composed of a 1-m multispectral band, leading to the higher spectral heterogeneity for the former. These properties significantly increase the difficulty of accurately identifying the GPC extent, especially when only image-level labels are available. However, the results of Figs. 8-9 show that, our CFWS method extracts more complete GPCs on both datasets using only image-level labels, while effectively retaining boundaries, compared to other methods.

As shown in Table 2, among all the methods, our CFWS method achieves the optimal accuracy for GPC detection on both datasets, with F1-score close to 90% and IoU over 80%. Other methods obtain F1-scores of 86.2% to 87.1% and IoU of 75.7% to 77.1% on GE-GPC, while they exhibit lower accuracies on GF2-GPC, with F1-scores of 68.2% to 79.1% and IoU of 51.8% to 65.5%. It can be said that our method significantly outperforms other ones on GF2-GPC, indicating the robustness of our method to different images.

4.2. Test in real scenarios

For real scenarios, we designed the classification-then-segmentation strategy (i.e., the two-step method) (Section 3.4). To verify its effectiveness, we compared the direct segmentation strategy (i.e., the one-step method) that only the segmentation model is used to predict GPC extents. Notice that the two-step method used both GPC and non-GPC classification samples to optimize the classification network, and then adopted refined GPC labels to train the segmentation network. For fair comparison, the one-step method used both refined GPC labels and non-GPC classification samples (of which the reference pixel-level labels are all non-GPCs) to construct the segmentation network, which ensures that both methods use the same data.

Fig. 10 and Fig. 11 display the predicted results for the large-area test samples of GE-GPC and GF2-GPC, respectively. In Fig. 10, the distribution of GPC is dense in Yishui and sparse in Zhangjiakou. The results show that our method produces fewer false alarms in non-GPC regions and successfully detects more complete GPCs, compared to the one-step method (see Fig. 10(a–d)). Similar results can be observed in Fig. 11. Notice that, compared to the Google Earth images (Fig. 10), the GF-2 images (Fig. 11) have a wider coverage and a smaller proportion of GPCs, which may lead the algorithm to generate more false alarms. However, as shown in Fig. 11(a–d), the proposed two-step method effectively suppresses false alarms on the GF-2 images while identifying more complete GPCs. For shadow-free GPCs, our approach identified relatively complete extents. As for shadow-contaminated GPCs (Fig. 10

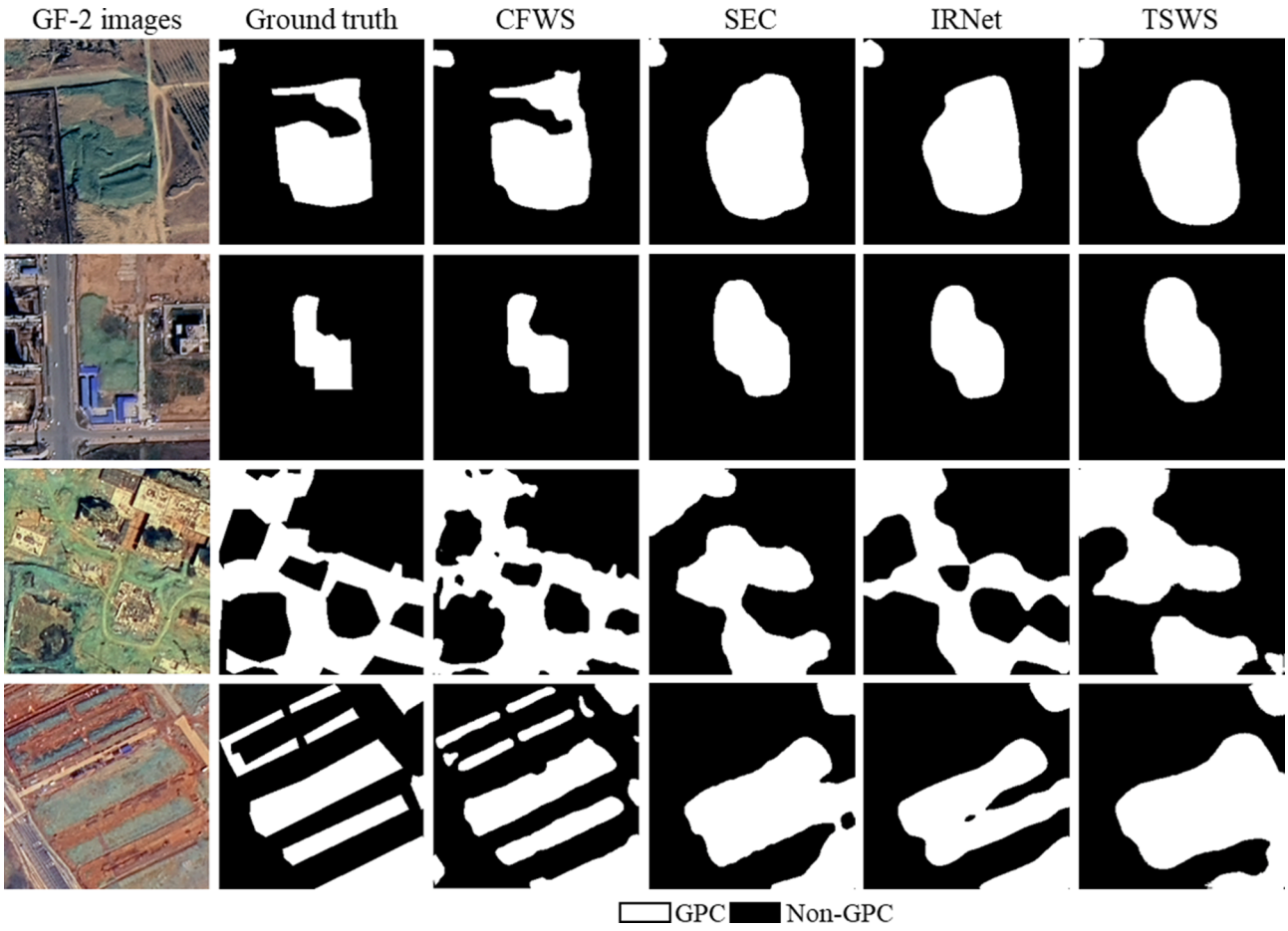

Fig. 9. Results of different weakly supervised segmentation methods on the segmentation samples of GF2-GPC.

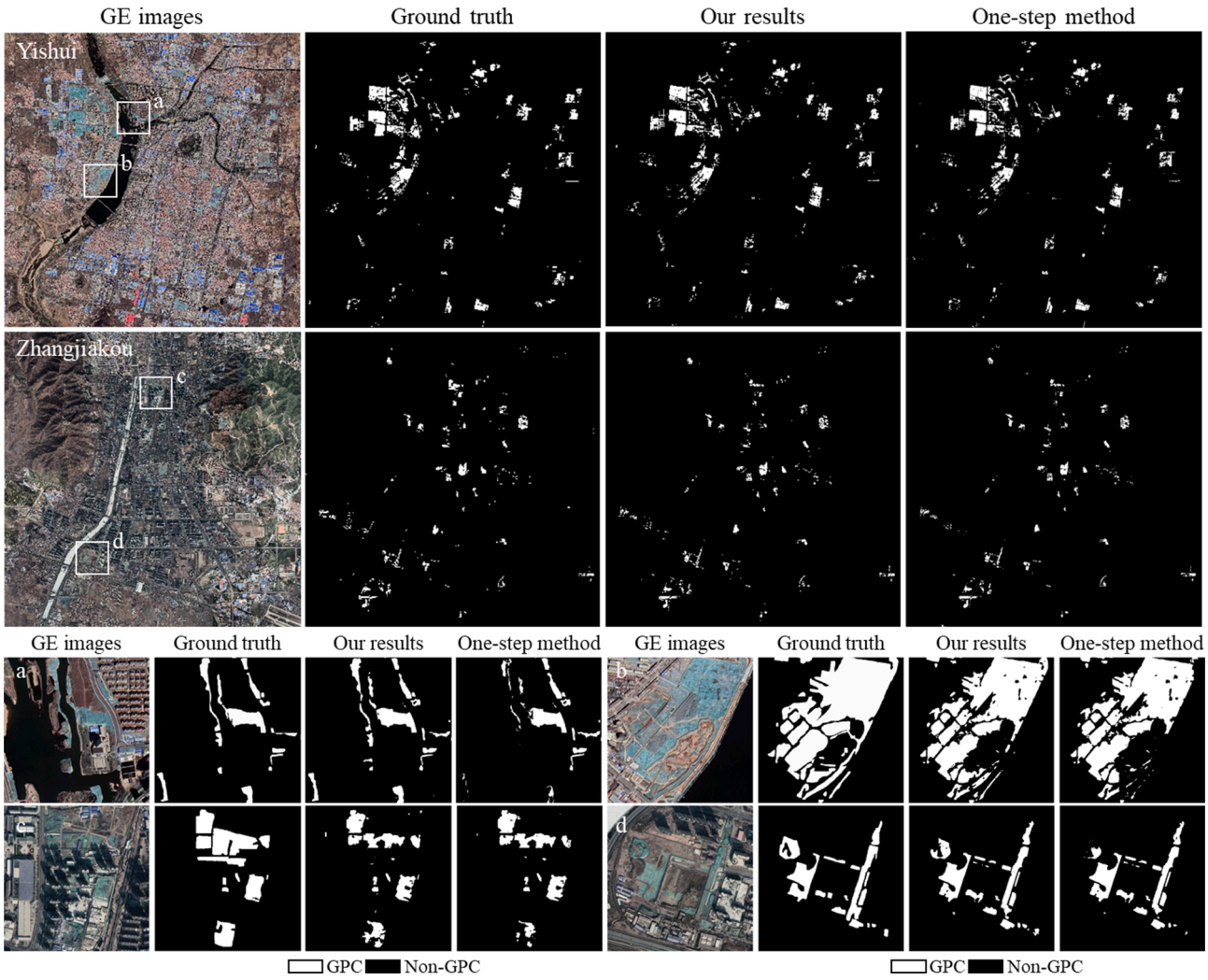

Fig. 10. Results of our two-step method and the one-step method on the GE-GPC test datasets. The last two rows show the zoomed-in views of regions a-d, where each region has a spatial extent of 1 km × 1 km.

(c)), our approach did not delineate their complete boundaries and led to a few omissions. We further analyzed the issue qualitatively and quantitatively in Section 5.5.1.

Table 3 shows the accuracies of our two-step method and the one-step method on large-area test samples of GE-GPC and GF2-GPC. In general, the two-step method outperforms the one-step method on the four large-area images. The accuracy difference between the two methods is primarily influenced by the sparse spatial distribution of GPC and the complex background, and this issue is especially obvious in Beijing and Tianjin, which have a wide image coverage. The two-step method can alleviate this issue to a certain extent, and obtains better accuracy than the one-step method, indicating the necessity of adopting the classification-then-segmentation strategy.

5. Discussions

5.1. Performance of object-based label extraction

To verify the effectiveness of object-based label extraction, we compared it with four widely-used initial label extraction methods. The first one is "pixel-based CAM", i.e., directly applying adaptive thresholding (the Otsu method in Section 3.1) on pixel-level CAM. The second one is "adaptive CAM + CRF", i.e., adaptively generating foreground (GPC) and background (non-GPC) thresholds for pixel-level CAM, and then using CRF (Krähenbühl and Koltun, 2011) for post-processing. The third one is "fixed CAM + CRF", which means that fixed foreground and background thresholds are selected for pixel-level CAM to obtain initial labels, and then the initial labels are post-processed using CRF. Here, the fixed thresholds were set the same as (Ahn et al., 2019), with the foreground threshold of 0.3 and the background threshold of 0.05. The fourth method is the widely-used multi-resolution segmentation method (MRS) (Blaschke, 2010), i.e., applying adaptive thresholding on object-level CAM obtained by the MRS method. A key parameter is the segmentation scale, and it is set to 50 considering the image spatial resolution and the size of target objects (i.e., green plastic cover) (Ma et al., 2017). We tested these methods on segmentation samples, and the quantitative and qualitative results are shown in Table 4 and Fig. 12, respectively. In Table 4, we can see that our method obtains the F1-score of 83.3% and the IoU of 71.4% on GE-GPC, and reaches the F1-score of 73.7% and the IoU of 58.4% on GF2-GPC, significantly outperforming other methods on both datasets. It can be observed that the MRS method only performs better than the pixel-based CAM. This phenomenon confirms that the object-based analysis is useful for mitigating the local high response property of CAM, but meanwhile, it's vital to choose appropriate segmentation method.

The visualization results in Fig. 12 show that our method can extract more complete GPCs, and better preserve boundaries, compared to other ones. We can observe that the "pixel-based CAM" approach suffers from the low resolution and the local high response property of CAM, which makes it difficult to obtain complete objects. The "adaptive CAM + CRF"

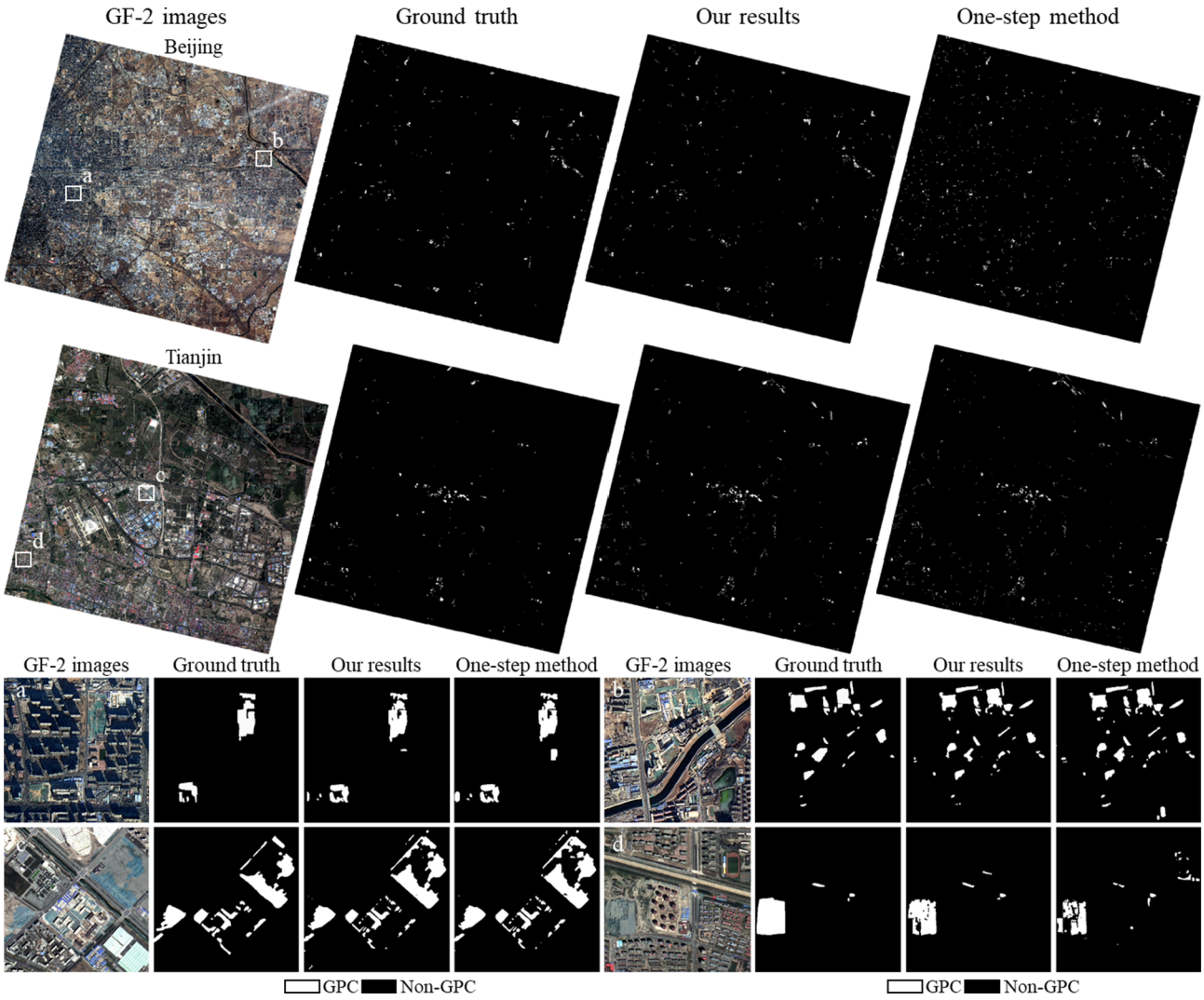

Fig. 11. Results of our two-step method and the one-step method on the GF2-GPC test datasets. The last two rows show the zoomed-in views of regions a-d, where each region has a spatial extent of 1 km × 1 km.

Table 3
Accuracies of different weakly supervised segmentation methods on the segmentation samples of GE-GPC and GF2-GPC.

Dataset	Method	OA	F1-score	Precision	Recall	IoU
GE-GPC	CFWS (Ours)	0.963	0.910	0.902	0.918	0.835
	SEC	0.947	0.871	0.871	0.871	0.771
	IRNet	0.945	0.862	0.891	0.834	0.757
	TSWS	0.942	0.863	0.837	0.891	0.759
GF2-GPC	CFWS (Ours)	0.954	0.893	0.835	0.959	0.806
	SEC	0.885	0.736	0.674	0.812	0.582
	IRNet	0.912	0.791	0.748	0.840	0.655
	TSWS	0.862	0.682	0.628	0.747	0.518

and the "fixed CAM + CRF" methods usually require manual selection of appropriate CRF parameters for post-processing. Similarly, the MRS method needs to choose the scale parameter in advance. However, it is challenging to obtain the optimal parameters adapted to all images, and such post-processing method with parameters usually leads to significant quality differences between the extraction results of different images. In contrast, our method, which segments images using the unsupervised image segmentation technique, is able to generate relatively complete GPC extents. The unsupervised segmentation technique can adapt to each image without manual adjustment of parameters and provide better initial pixel-level labels for subsequent semantic segmentation.

5.2. Performance of noisy label correction

To evaluate the performance of noisy label correction, we compared the results with and without noisy label correction on the segmentation samples. As shown in Table 5, the baseline method consists of object-based label extraction and GPC segmentation network with the binary cross entropy loss function (BCE, see Eq. (6)). It can be seen that the noisy label correction module significantly improves the F1-score and IoU values of the GPC detection, regardless of whether the segmentation network uses the BCE or the boundary-aware loss function (BAL, see Eq. (8)), indicating that the module can effectively correct noisy samples (Table 6).

To further analyze the effect of noisy label correction on the initial labels, we compared the initial labels before and after noisy label correction, and calculated the ratio of increased and decreased GPC pixels for each training sample. The change (increase or decrease) ratio of GPC pixels can reflect the proportion of the potential noisy labels as well as the intensity of noisy label correction. The results are displayed in Fig. 13. For GE-GPC, the ratio of GPC pixels over all training samples increases by 3.44% and decreases by 3.87%, with a total change ratio of 7.31%. For GF2-GPC, the ratio of GPC pixels over all training samples

Table 4
Accuracies of the two-step method (ours) and the one-step method on the GE-GPC and GF2-GPC test dataset.

Dataset	City/county	Method	OA	F1-score	Precision	Recall	IoU
GE-GPC	Yishui	Two-step (Ours)	0.993	0.898	0.932	0.866	0.815
		One-step	0.993	0.895	0.934	0.858	0.810
	Zhangjiakou	Two-step (Ours)	0.996	0.835	0.896	0.781	0.716
		One-step	0.996	0.815	0.913	0.735	0.687
GF2-GPC	Beijing	Two-step (Ours)	0.998	0.755	0.726	0.785	0.606
		One-step	0.996	0.563	0.449	0.754	0.392
	Tianjin	Two-step (Ours)	0.997	0.668	0.55	0.851	0.502
		One-step	0.997	0.613	0.499	0.795	0.442

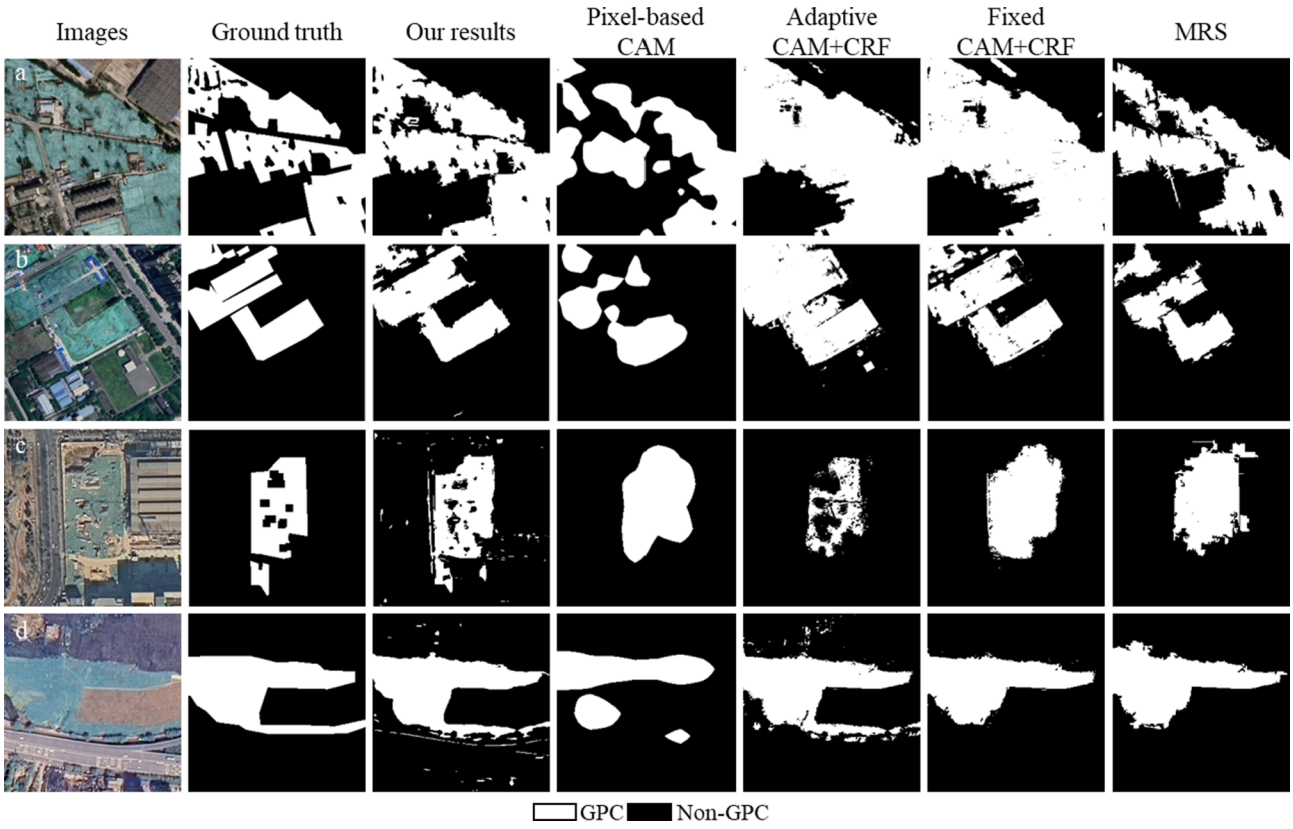

Fig. 12. Results of different initial pixel-level label extraction methods on the segmentation samples of GE-GPC (a-b) and GF2-GPC (c-d).

Table 5
Accuracies of different initial pixel-level label extraction methods on the segmentation samples of GE-GPC and GF2-GPC.

Dataset	Method	OA	F1-score	Precision	Recall	IoU
GE-GPC	Ours	0.932	0.833	0.836	0.830	0.714
	Pixel-based CAM	0.885	0.680	0.792	0.595	0.515
	Adaptive CAM + CRF	0.915	0.812	0.743	0.896	0.684
	Fixed CAM + CRF	0.918	0.785	0.846	0.731	0.645
	MRS	0.900	0.736	0.798	0.683	0.582
GF2-GPC	Ours	0.887	0.737	0.683	0.801	0.584
	Pixel-based CAM	0.834	0.487	0.631	0.396	0.322
	Adaptive CAM + CRF	0.873	0.595	0.808	0.471	0.424
	Fixed CAM + CRF	0.856	0.596	0.670	0.537	0.425
	MRS	0.845	0.552	0.648	0.481	0.381

increases by 6.87% and decreases by 7.25%, with a total change ratio of 14.12%. We can find that the total change ratio of GF2-GPC is significantly higher than that of GE-GPC. This finding indicates that the initial labels of GF2-GPC contain more noise, which is also reflected in the lower accuracy of initial labels (Table 4).

Fig. 14 presents the results of initial labels before and after noisy label correction. We can observe that the noisy label correction module successfully filters out the potential false alarms in initial labels (e.g., Fig. 14(a) and (h)), and meanwhile effectively identifies relatively complete GPCs (e.g., Fig. 14(e) and (f)). These results verify that noisy label correction can improve the quality of initial labels, and generate relatively clear pixel-level samples for subsequent semantic segmentation.

5.3. Performance of boundary-aware segmentation

To investigate the effectiveness of the boundary-aware loss function (BAL, see Eq. (8)), we compared it with the binary cross entropy loss function (BCE, see Eq. (6)), and the corresponding accuracies on the segmentation samples are shown in Table 6. It can be found that when we only use object-based label extraction and do not use noisy label correction, BAL significantly enhances the accuracy of GPC segmentation, where the F1-score and IoU values are improved by 1.6% and 2.5%, respectively, for GE-GPC, and 0.8% and 1.2%, respectively, for GF2-GPC. Moreover, BAL also outperforms BCE on GPC segmentation

Table 6
Ablation experimental results of the proposed method. OBLE: object-based label extraction. NLC: noisy label correction. BCE: binary cross entropy. BAL: boundary-aware loss.

Dataset	OBLE	NLC	Segmentation loss		OA	F1-score	Precision	Recall	IoU
			BCE	BAL					
GE-GPC	√		√		0.949	0.866	0.929	0.812	0.764
	√	√	√		0.961	0.907	0.898	0.916	0.829
	√			√	0.952	0.882	0.892	0.873	0.789
	√	√		√	0.963	0.910	0.902	0.918	0.835
GF2-GPC	√		√		0.950	0.868	0.918	0.823	0.767
	√	√	√		0.951	0.887	0.820	0.966	0.797
	√			√	0.948	0.876	0.834	0.922	0.779
	√	√		√	0.954	0.893	0.835	0.959	0.806

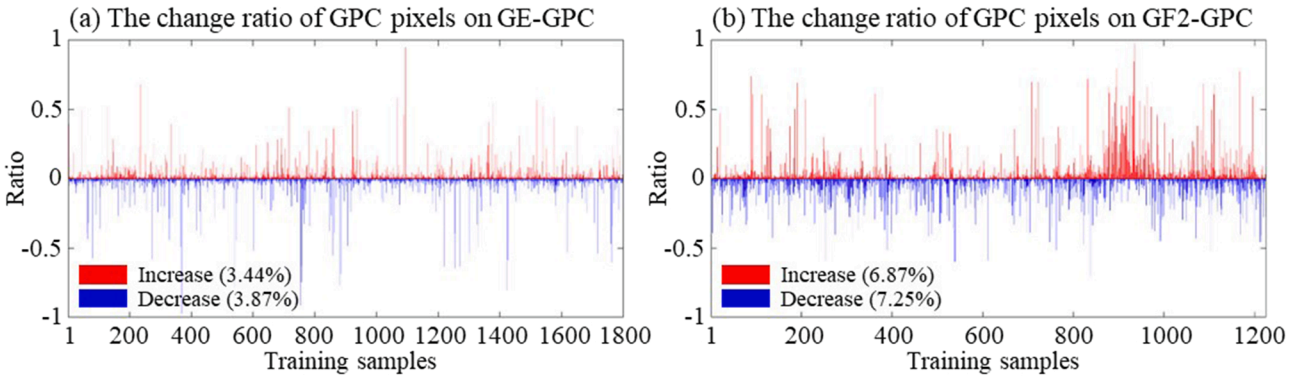

Fig. 13. The change ratio of GPC pixels on the training samples of GE-GPC (a) and GF2-GPC (b).

when both object-based label extraction and noisy label correction modules are used. With BAL, the F1-score and IoU values are improved by 0.3% and 0.6%, respectively, for GE-GPC, and 0.6% and 0.9%, respectively, for GF2-GPC. These results indicate that the boundary-aware loss function can improve the accuracy of GPC segmentation while being stable to noisy labels. Fig. 15 displays the segmentation results with the boundary-aware loss function and BCE, when both object-based label extraction and noisy label correction modules are used. It can be seen that the boundary-aware loss function generates fewer false alarms than BCE (e.g., Fig. 15(a) and (e)).

5.4. Threshold sensitivity analysis for real scenarios

For real scenarios, we proposed the classification-then-segmentation strategy (i.e., the two-step method) to alleviate a large number of false alarms that may arise from direct segmentation (i.e., the one-step method). In Section 4.2, we compared the two-step method with the one-step method both quantitatively and qualitatively, and found that the two-step method is significantly better than the one-step method. Note that the two-step method may suffer from the error accumulation, i.e., the GPC missed in the first step cannot be recognized in the second step. To deal with this issue, in the first step, we used a small threshold to obtain the potential GPC regions. To further analyze the sensitivity of thresholds in the two-step method, we tested the accuracy of GPC segmentation with different classification thresholds (denoted as T, see Section 3.4). The threshold starts from 0 and increases to 1.0 with an interval of 0.1. For the four test regions in this study, when the threshold exceeds 0.8, all regions will be predicted to be non-GPC. Therefore, we set the maximum value of the threshold as 0.7 in the sensitivity experiment. Fig. 16 shows the accuracies of GPC segmentation on the four test regions with different thresholds. We can observe that for the four test regions, the F1-score and IoU values show a trend of rising and then falling as the threshold increases. When the threshold is too low, most of the regions are predicted as potential GPC regions, which tends to introduce a large number of false alarms, resulting in low precision and high recall. On the contrary, when the threshold is too high, only a small portion of the regions are predicted as potential GPC regions, which may miss a large number of GPCs, leading to high precision and low recall. For the four regions, the F1-score and IoU values are relatively stable when the threshold is between 0.1 and 0.4, and this threshold range can provide a reference for practical applications. In this study, the threshold was set to 0.2, since this value can provide a trade-off between omissions and false alarms.

5.5. Effectiveness of the proposed method

We proposed the weakly-supervised deep learning method to identify GPCs, mainly considering the following two factors:

1) The powerful feature extraction ability.

GPCs in Figs. 1, 2, and 5 have a different spectral property against their surroundings, which is useful for GPC detection. However, we notice that GPCs in different areas exhibit different spectral, spatial, and contextual features, making it challenging to design a simple but effective extraction method. Compared to traditional classification methods, e.g., random forest (RF) and support vector machine (SVM) (Huang and Zhang, 2013; Rodriguez-Galiano et al., 2012), deep learning does not need specific domain knowledge, and can automatically learn discriminative and representative features from massive data.

2) The low acquisition cost of image-level weak labels.

For deep learning, the semantic segmentation methods heavily rely on large amounts of high-quality pixel-level labels (i.e., one pixel corresponds to one label) for optimizing parameters, which is usually time-consuming and labor-intensive. In addition, the traditional classifiers (e.g., RF and SVM) also need high-quality pixel-level labels. Therefore, in this study, we consider image-level weak labels (i.e., one image corresponds to one label), which have significantly lower acquisition cost.

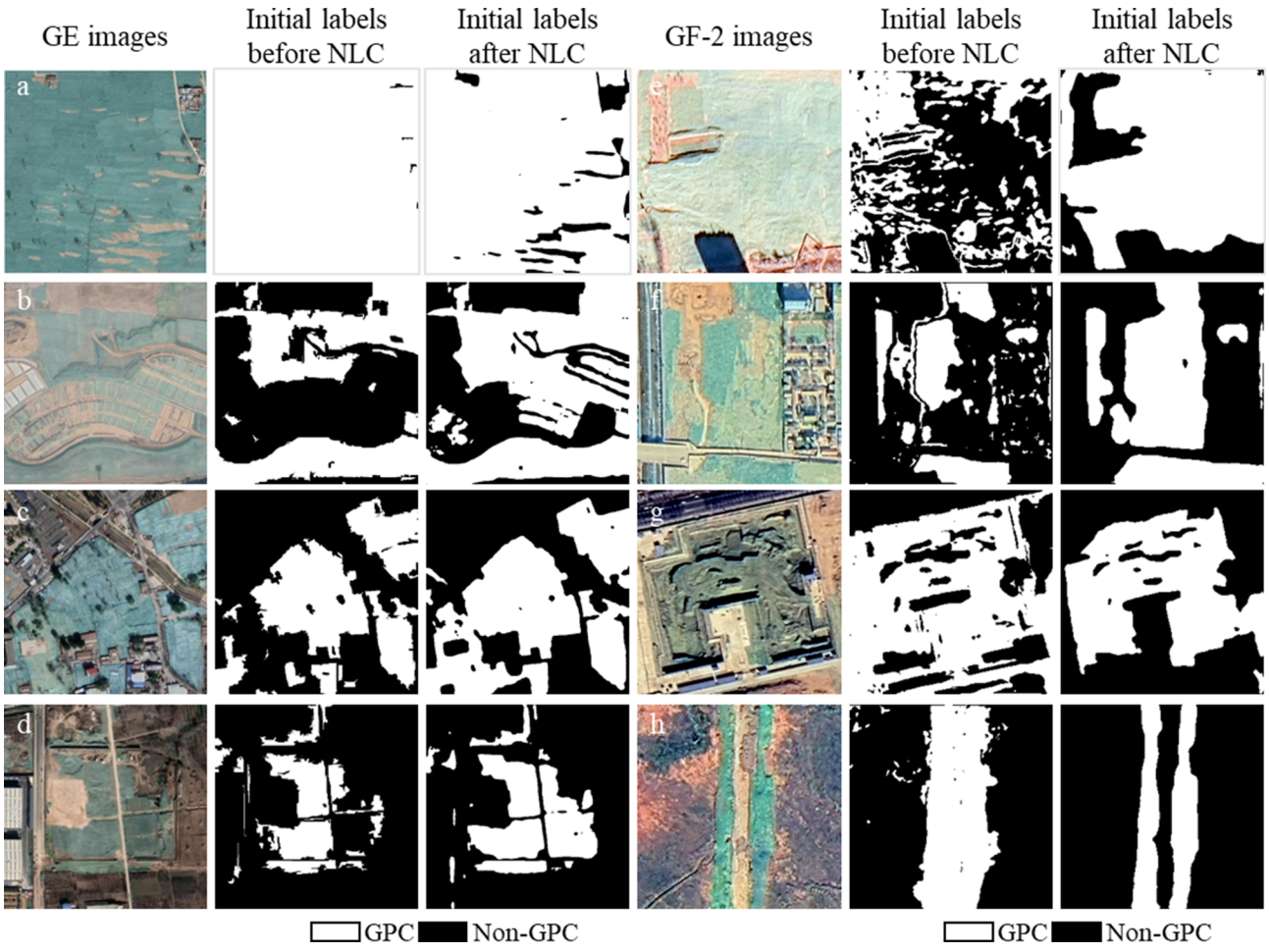

Fig. 14. Results of initial labels before and after noisy label correction (NLC) on the training sets of GE-GPC (a-d) and GF2-GPC (e-h).

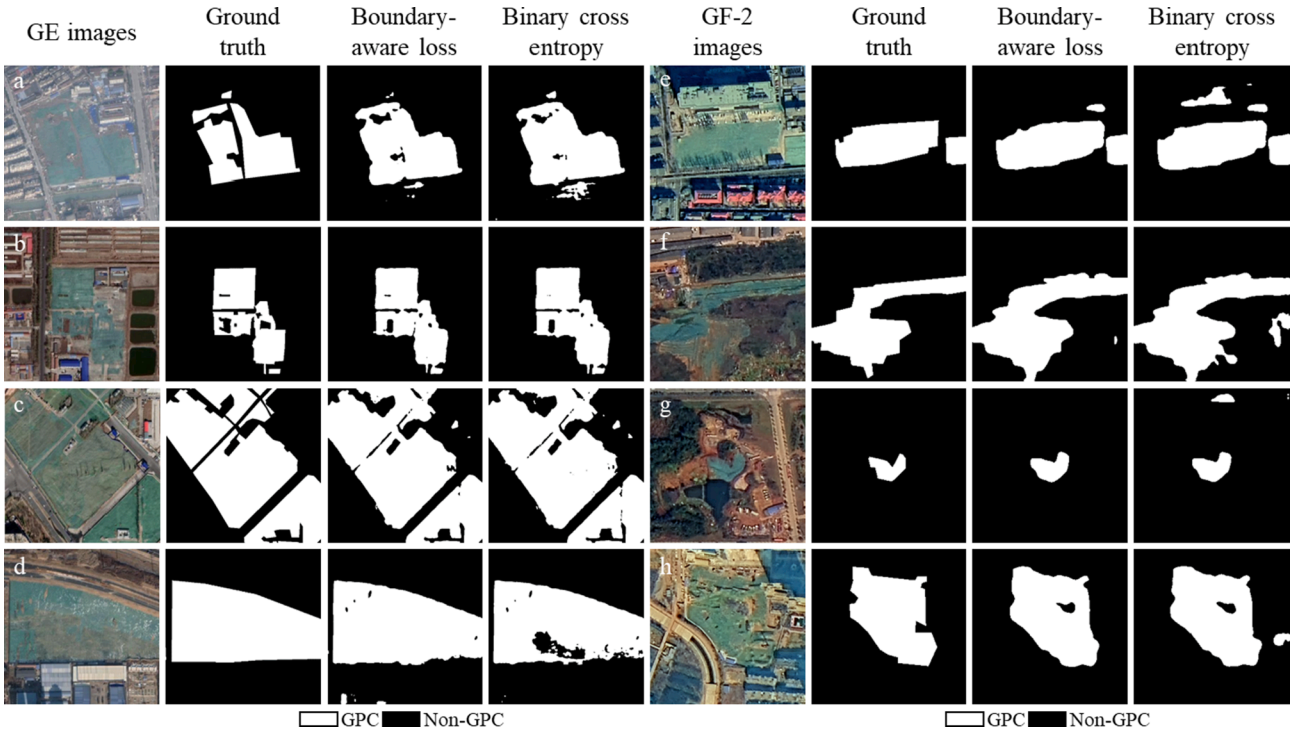

Fig. 15. Results of GPC segmentation with boundary-aware loss and binary cross entropy on the segmentation samples of GE-GPC (a-d) and GF2-GPC (e-h).

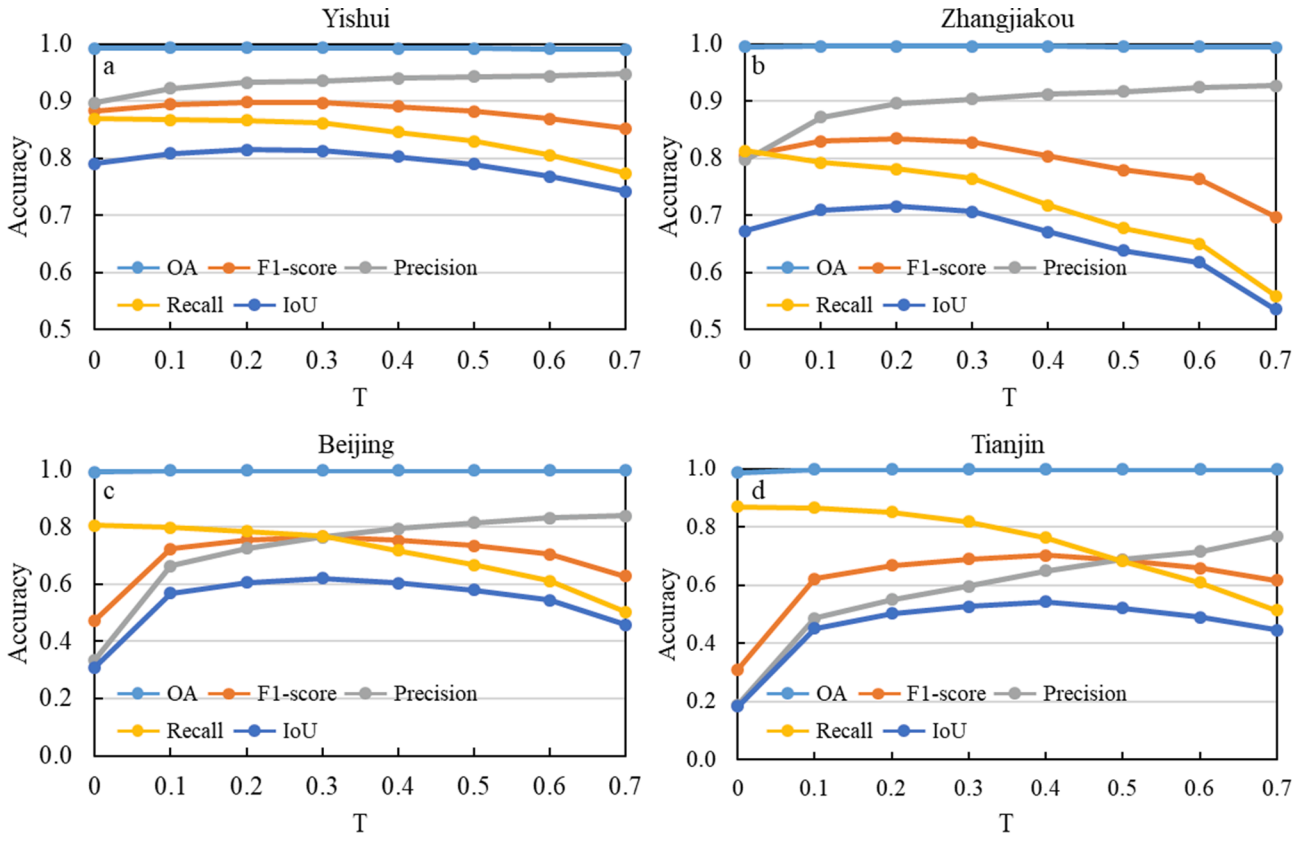

Fig. 16. Accuracies of GPC segmentation for GE-GPC (a-b) and GF2-GPC (c-d) with different classification thresholds (T).

However, the limited information of image-level labels makes it difficult for accurate GPC segmentation. Therefore, we propose a coarse-to-fine weakly supervised segmentation method, aiming to obtain satisfactory results with relatively low labelling cost.

We further evaluated the performance of the proposed method in three conditions including GPCs contaminated by shadows, GPCs mixed with cropland, and GPCs mixed with urban greenery below.

5.5.1. *GPCs contaminated by shadows*

We analyzed the performance of our approach on shadow-free/contaminated GPCs qualitatively and quantitatively. Table 7 shows the accuracies of GPC segmentation with different weakly supervised segmentation methods on the GE-GPC test datasets. Overall, our approach performs best. Furthermore, Fig. 17 compared the results of shadow-free/contaminated GPCs. Frankly, we can see that our approach did not delineate the complete boundaries of the contaminated GPCs. The same problem can be also observed in other methods. The issue is possibly caused by the loss of spectral and textural information in the shadow regions. Specifically, for the shadow-contaminated GPCs, their spectral and textural properties are significantly different from those of shadow-free GPCs. More importantly, our approach uses image-level weak labels (not pixel-level precise labels), which only indicate if the object of interest exists in the image, without specifying their spatial locations. This may weaken the ability of the method in detecting object boundaries. Therefore, the proposed method did not obtain satisfactory results for the shadow-contaminated GPCs.

Table 7
Accuracies of GPC segmentation with different weakly supervised segmentation methods on the GE-GPC test datasets. The highest score is marked in bold.

	OA	F1-score	Precision	Recall	IoU
CFWS (Ours)	**0.995**	**0.882**	**0.923**	**0.844**	**0.789**
SEC	0.993	0.844	0.912	0.785	0.730
IRNet	0.993	0.843	0.911	0.785	0.729
TSWS	0.993	0.851	0.870	0.833	0.741

We plan to address the issue of detecting shadow-contaminated GPCs in future work. Generally, the shadow effect can be alleviated by two methods, i.e., shadow classification and shadow correction. The former considers shadows as a single class and then classifies shadows with high-quality pixel-level labels (Jiao et al., 2020; Luo et al., 2020), while the latter detects shadow-contaminated areas and then attempts to recover their ground information (Luo et al., 2019; Ma et al., 2008). The performance of the two methods on identifying shadow-contaminated GPCs needs more detailed exploration in future work.

5.5.2. *GPCs mixed with cropland*

We analyzed these images containing both GPCs and cropland. For each image patch (512 × 512 pixels), we manually interpreted the spatial extent of cropland with the aid of Google Earth high-resolution images and 10-m global land cover product from ESA (Zanaga et al., 2021), and obtained the pixel-level cropland labels. We used the image patches that simultaneously contain GPCs and cropland for the experiment. Quantitative results are recorded in Table 8. It can be seen that our method achieves the highest F1-score and IoU values on GPC segmentation, compared to other weakly supervised methods, verifying that our approach can distinguish GPCs from cropland more effectively. The results shown in Fig. 18 further demonstrate the satisfactory performance of our approach in identifying GPCs and cropland. Although some cropland may have a similar spectral and texture property to GPCs, their spatial and contextual patterns can be well captured by the proposed model. For example, the spatial distribution of GPCs is usually scattered, while that of cropland is concentrated. The proposed method is able to describe and learn these discriminative features.

5.5.3. *GPCs mixed with urban greenery*

We investigated these images containing both GPCs and urban

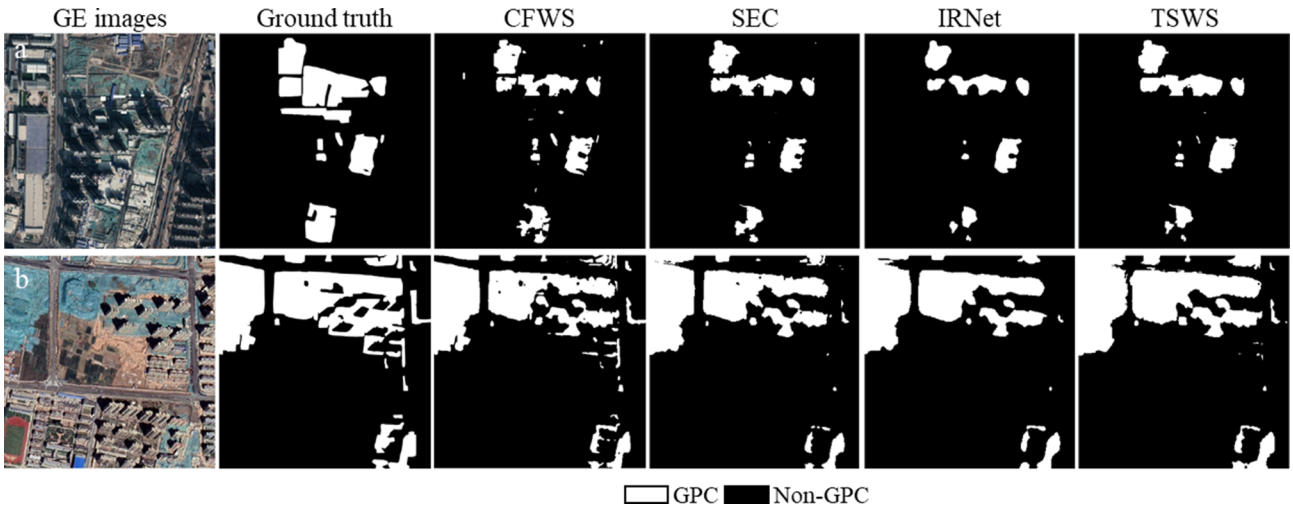

Fig. 17. Results of different weakly supervised segmentation methods on shadow-contaminated GPC regions (a-b).

Table 8
Accuracies of GPC segmentation by different weakly supervised segmentation methods on image patches containing both GPCs and cropland from the GE-GPC test datasets. The highest score is marked in bold.

	OA	F1-score	Precision	Recall	IoU
CFWS (Ours)	**0.984**	**0.859**	**0.906**	**0.817**	**0.753**
SEC	0.979	0.810	0.893	0.741	0.681
IRNet	0.980	0.828	0.868	0.791	0.706
TSWS	0.979	0.821	0.840	0.802	0.696

greenery. For each image patch (512 × 512 pixels), we manually interpreted the spatial extent of urban greenery with the aid of Google Earth high-resolution images, and obtained the pixel-level urban greenery labels. The image patches that contain both GPCs and urban greenery were focused on. Results are recorded in Table 9, and it can be seen that our approach significantly outperforms other ones. It indicates that our approach can better distinguish GPCs from urban greenery. This conclusion can be also supported by the visual inspection in Fig. 19.

Table 9
Accuracies of GPC segmentation with different weakly supervised segmentation methods on image patches containing both GPCs and urban greenery from the GE-GPC test dataset. The highest score is marked in bold.

	OA	F1-score	Precision	Recall	IoU
CFWS (Ours)	**0.981**	**0.900**	**0.925**	**0.876**	**0.818**
SEC	0.972	0.849	0.921	0.788	0.738
IRNet	0.973	0.854	0.920	0.797	0.746
TSWS	0.972	0.855	0.879	0.832	0.746

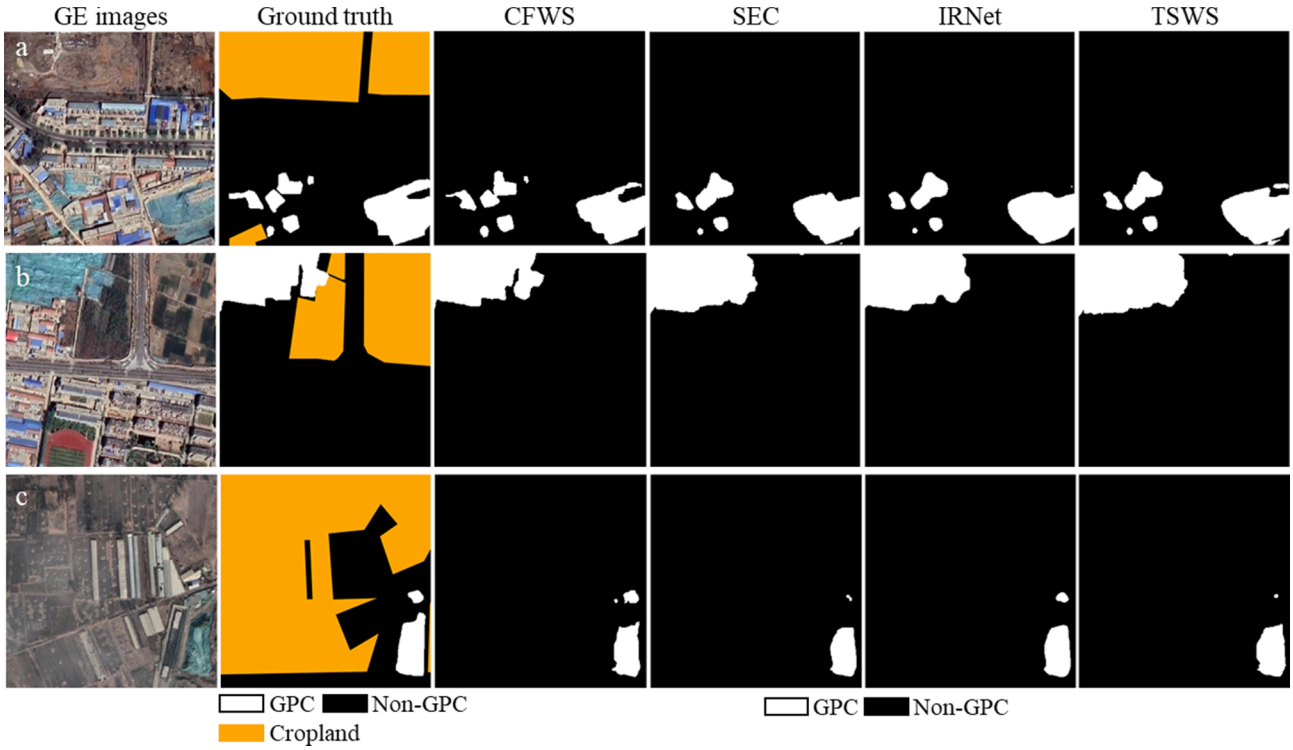

Fig. 18. Results of different weakly supervised segmentation methods on image patches (512 × 512 pixels) containing both GPCs and cropland (a-c).

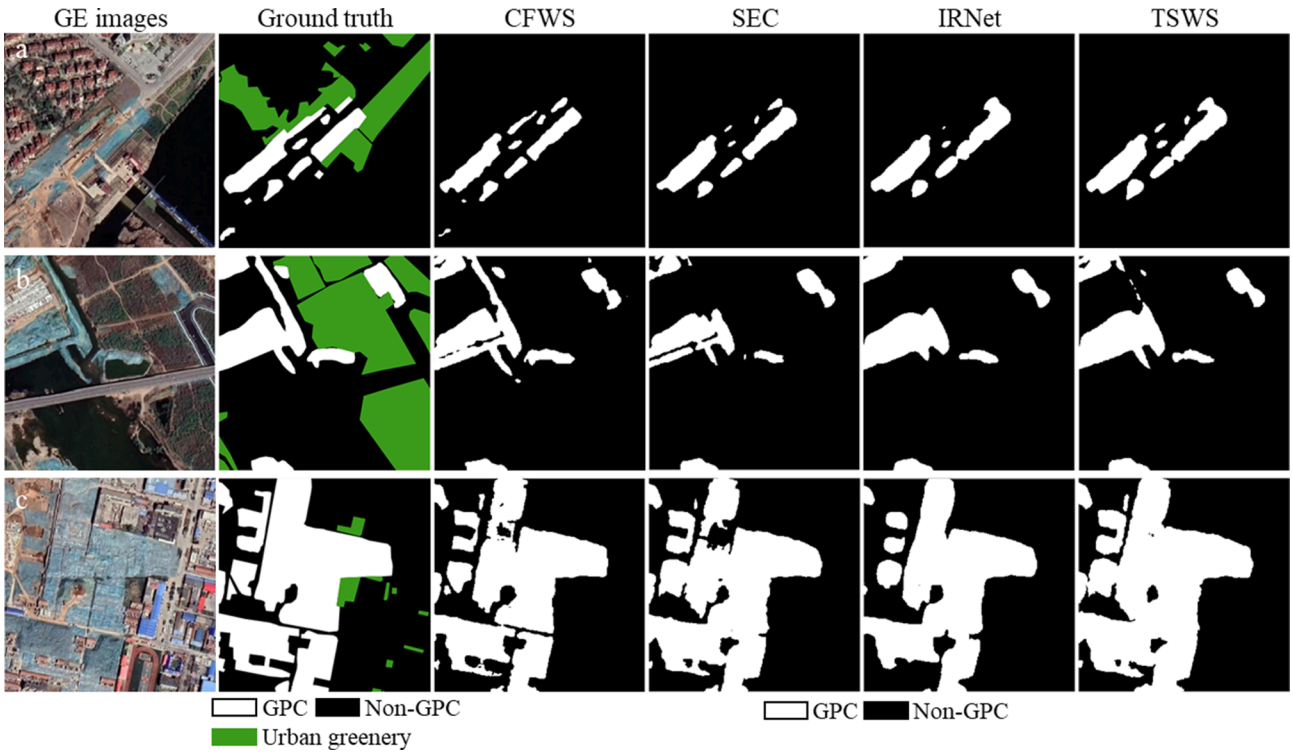

Fig. 19. Results of different weakly supervised segmentation methods on image patches (512 × 512 pixels) containing both GPCs and urban greenery (a-c).

6. Conclusions

In this study, we focused on GPC segmentation using high-resolution remote sensing imagery, which is important for monitoring urban environment and understanding urban development. Convolutional neural network (CNN)-based segmentation methods are widely used for detecting object extents, while they rely on high-quality pixel-level labels with high acquisition cost. In this regard, weakly supervised learning can achieve pixel-level GPC segmentation with only image-level labels. Existing studies on image-level weakly supervised segmentation have made great progress, but they are still limited by the local high response property of CAM and the potential label noise problem. Moreover, the widely-used encoder-decoder semantic segmentation models tend to produce blurry object boundaries due to the gradual down-sampling of feature maps, which poses a great challenge to weakly supervised segmentation where only coarse labels are available. When applied to real scenarios, semantic segmentation models are usually influenced by the class imbalance problem, which is less considered in existing research. Given these limitations, we proposed a coarse-to-fine weakly supervised learning method (called CFWS) for GPC detection. The CFWS consists of three components: 1) object-based label extraction; 2) noisy label correction; and 3) boundary-aware semantic segmentation. Furthermore, for real scenarios, we designed a classification-then-segmentation (i.e., the two-step approach) strategy to mitigate the class imbalance problem.

We tested the proposed method on the GE-GPC and GF2-GPC datasets. The results showed that the proposed CFWS extracted more complete GPCs on both datasets compared to existing state-of-the-art methods, while effectively retaining boundaries. In real scenarios, the classification-then-segmentation strategy significantly reduced a large number of false alarms generated by direct segmentation (i.e., the one-step method). Furthermore, we found that, object-based label extraction effectively obtained more complete initial GPC labels with better boundaries, compared to existing methods, and can adapt to each image without manual adjustment of parameters. Then, we analyzed the impact of noisy label correction and found that it can effectively filter out potential false alarms in the initial labels while reducing omissions and generating relatively clear pixel-level labels. In addition, we explored the effectiveness of the boundary-aware loss function and the results showed that it can improve the accuracy of GPC segmentation while being stable to noisy labels. Finally, we analyzed the threshold sensitivity in the two-step method. The experimental results showed that the proposed method holds great potentials for GPC detection in real scenarios, and provides an effective means for urban environmental monitoring.

There still exist some limitations in this study. The first one is the transferability of the proposed method. When new regions are significantly different from the training set, it may be difficult to obtain satisfactory results by directly using the well-trained network. In this study, we applied the well-trained network to four large regions and achieved good results with the classification-then-segmentation strategy. However, when new regions contain few GPCs, complex backgrounds may still cause the network to generate numerous false alarms even with the classification-then-segmentation strategy. One solution is to first apply prior knowledge to mask the background for obtaining the potential GPC region, and then perform GPC segmentation only on that region. Another way is to use transfer learning techniques, such as domain adaptation (Tuia et al., 2016) that can reduce the data distribution offset between the new regions (i.e., the target domain) and the training set (the source domain). The second limitation is the data availability. In this study, GF-2 images were collected from the China Resources Satellite Application Center and are not publicly available. In contrast, Google Earth high-resolution imagery is freely available, which provides a basic data support for large-scale GPC identification. In future research, we plan to explore transfer learning techniques and apply the proposed method to other regions. The code of this study will be available at https://github.com/lauraset/Coarse-to-fine-weakly-supervised-GPC-segmentation.

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.

Acknowledgments

The research was supported by the National Natural Science Foundation of China under Grant 41971295, and the Foundation for Innovative Research Groups of the Natural Science Foundation of Hubei Province under Grant 2020CFA003.

References

Ahn, J., Cho, S., Kwak, S., 2019. Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2209–2218.

Ahn, J., Kwak, S., 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4981–4990.

Ali, M.U., Sultani, W., Ali, M., 2020. Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery. ISPRS J. Photogramm. Remote Sens. 162, 115–124.

Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K., 2019. Unsupervised label noise modeling and loss correction. Int. Conf. Mach. Learning. PMLR 312–321.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision. Eur. Conf. Computer Vision. Springer 549–565.

Bellens, R., Gautama, S., Martinez-Fonte, L., Philips, W., Chan, J.-W., Canters, F., 2008. Improved classification of VHR images of urban areas using directional morphological profiles. IEEE Trans. Geosci. Remote Sens. 46 (10), 2803–2813.

Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogramm. Remote Sens. 65 (1), 2–16. https://doi.org/10.1016/j.isprsjprs.2009.06.004.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259.

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.

Chan, L., Hosseini, M.S., Plataniotis, K.N., 2021. A comprehensive analysis of weakly-supervised semantic segmentation in different image domains. Int. J. Comput. Vis. 129 (2), 361–384.

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N., 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 839–847.

Chen, D., Wei, H., 2009. The effect of spatial autocorrelation and class proportion on the accuracy measures from different sampling designs. ISPRS J. Photogramm. Remote Sens. 64 (2), 140–150.

Chen, J., He, F., Zhang, Y.i., Sun, G., Deng, M., 2020. SPMF-Net: Weakly supervised building segmentation by combining superpixel pooling and multi-scale feature fusion. Remote Sens. 12 (6), 1049. https://doi.org/10.3390/rs12061049.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 801–818.

Dong, R., Fang, W., Fu, H., Gan, L., Wang, J., Gong, P., 2021. High-Resolution Land Cover Mapping Through Learning With Noise Correction. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Guo, H., Shi, Q., Marinoni, A., Du, B.o., Zhang, L., 2021. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. Remote Sens. Environ. 264, 112589. https://doi.org/10.1016/j.rse.2021.112589.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M., 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. arXiv Prepr. arXiv1804.06872.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

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, 111802. https://doi.org/10.1016/j.rse.2020.111802.

Huang, X., Zhang, L., 2013. An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 51 (1), 257–272.

Huang, Z., Wang, X., Wang, Jiasi, Liu, W., Wang, Jingdong, 2018. Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7014–7023.

Jia Deng, Wei Dong, Socher, R., Li-Jia Li, Kai Li, Li Fei-Fei, 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.

Jiang, N., Dong, Z., Xu, Y., Yu, F., Yin, S., Zhang, R., Tang, X., 2018. Characterization of PM10 and PM2.5 Source Profiles of Fugitive Dust in Zhengzhou, China. Aerosol Air Qual. Res. 18 (2), 314–329.

Jiao, L., Huo, L., Hu, C., Tang, P., 2020. Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation. Remote Sens. 12 (12), 2001. https://doi.org/10.3390/rs12122001.

Jo, S., Yu, I.-J., 2021. Puzzle-CAM: Improved localization via matching partial and full features. arXiv Prepr. arXiv2101.11253.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6, 1–54.

Jung, H., Choi, H.-S., Kang, M., 2021. Boundary Enhancement Semantic Segmentation for Building Extraction From Remote Sensed Image. IEEE Trans. Geosci. Remote Sensing 60, 1–12.

Kanezaki, A., 2018. Unsupervised image segmentation by backpropagation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1543–1547.

Kellenberger, B., Marcos, D., Tuia, D., 2018. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. Remote Sens. Environ. 216, 139–153.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 876–885.

Kolesnikov, A., Lampert, C.H., 2016. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. Eur. Conf. Computer Vision. Springer 695–711.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. Adv. Neural Inf. Process. Syst. 24, 109–117.

Li, A., Jiao, L., Zhu, H., Li, L., Liu, F., 2021. Multitask Semantic Boundary Awareness Network for Remote Sensing Image Segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Li, C., Song, Y., Chen, Y., 2017. Infrastructure Development and Urbanization in China. China's Urban. Socioecon. Impact 91–107.

Li, Z., Zhang, X., Xiao, P., Zheng, Z., 2021. On the Effectiveness of Weakly Supervised Semantic Segmentation for Building Extraction From High-Resolution Remote Sensing Imagery. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 14, 3266–3281.

Lin, D., Dai, J., Jia, J., He, K., Sun, J., 2016. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3159–3167.

Luo, S., Li, H., Shen, H., 2020. Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset. ISPRS J. Photogramm. Remote Sens. 167, 443–457. https://doi.org/10.1016/j.isprsjprs.2020.07.016.

Luo, S., Shen, H., Li, H., Chen, Y., 2019. Shadow removal based on separated illumination correction for urban aerial remote sensing images. Signal Process. 165, 197–208. https://doi.org/10.1016/j.sigpro.2019.06.039.

Ma, H., Qin, Q., Shen, X., 2008. Shadow segmentation and compensation in high resolution satellite images. In: IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. II–1036.

Ma, L., Li, M., Ma, X., Cheng, L., Du, P., Liu, Y., 2017. A review of supervised object-based land-cover image classification. ISPRS J. Photogramm. Remote Sens. 130, 277–293. https://doi.org/10.1016/j.isprsjprs.2017.06.001.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172.

Nivaggioli, A., Randrianarivo, H., 2019. Weakly supervised semantic segmentation of satellite images. In: 2019 Joint Urban Remote Sensing Event (JURSE). IEEE, pp. 1–4.

Oh, Y., Kim, B., Ham, B., 2021. Background-Aware Pooling and Noise-Aware Loss for Weakly-Supervised Semantic Segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6913–6922.

Otsu, N., 1979. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man. Cybern. 9 (1), 62–66.

Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P., 2020. Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10428–10436.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Song, H., Kim, M., Park, D., Shin, Y., Lee, J.-G., 2020. Learning from noisy labels with deep neural networks: A survey. arXiv Prepr. arXiv2007.08199.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Srivastava, S., Vargas-Muñoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. Remote Sens. Environ. 228, 129–143. https://doi.org/10.1016/j.rse.2019.04.014.

Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K., 2018. Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5552–5560.

Taubenböck, H., Kraff, N.J., Wurm, M., 2018. The morphology of the Arrival City - A global categorization based on literature surveys and remotely sensed data. Appl. Geogr. 92, 150–167.

Tuia, D., Persello, C., Bruzzone, L., 2016. Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE Geosci. Remote Sens. Mag. 4 (2), 41–57.

van Vliet, J., Eitelberg, D.A., Verburg, P.H., 2017. A global analysis of land take in cropland areas and production displacement from urbanization. Glob. Environ. Chang. 43, 107–115.

Volpi, M., Tuia, D., 2016. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55 (2), 881–893.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. image Process. 13 (4), 600–612.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500.

Yang, S., Liu, J., Bi, X., Ning, Y., Qiao, S., Yu, Q., Zhang, J., 2020. Risks related to heavy metal pollution in urban construction dust fall of fast-developing Chinese cities. Ecotoxicol. Environ. Saf. 197, 110628. https://doi.org/10.1016/j.ecoenv.2020.110628.

Yessou, H., Sumbul, G., Demir, B., 2020. A comparative study of deep learning loss functions for multi-label remote sensing image classification. In: IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1349–1352.

Yi, K., Wu, J., 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7017–7025.

Yu, B., 2021. Ecological effects of new-type urbanization in China. Renew. Sustain. Energy Rev. 135, 110239. https://doi.org/10.1016/j.rser.2020.110239.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., 2021. ESA WorldCover 10 m 2020 v100.

Zhang, D.D., Xie, F., Zhang, L., 2019. Preprocessing and fusion analysis of GF-2 satellite Remote-sensed spatial data. In: Proceedings of 2018 International Conference on Information Systems and Computer Aided Education, ICISCAE 2018. pp. 24–29.

Zhang, J., Jia, X., Hu, J., 2021. SP-RAN: Self-paced Residual Aggregated Network for Solar Panel Mapping in Weakly Labelled Aerial Images. IEEE Trans. Geosci. Remote Sens. 1.

Zhang, L., Ma, J., Lv, X., Chen, D., 2020. Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. IEEE Geosci. Remote Sens. Lett. 17 (1), 117–121.

Zhang, Z., Sabuncu, M.R., 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In: 32nd Conference on Neural Information Processing Systems (NeurIPS).

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

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R., 2019. Iou loss for 2d/3d object detection. In: 2019 International Conference on 3D Vision (3DV). IEEE, pp. 85–94.

Zhou, D., Wang, G., He, G., Yin, R., Long, T., Zhang, Z., Chen, S., Luo, B., 2021. A large-scale mapping scheme for urban building from gaofen-2 images using deep learning and hierarchical approach. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 14, 11530–11545.