Neurocomputing 458 (2021) 127–140

Contents lists available at ScienceDirect

Neurocomputing

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

Density saliency for clustered building detection and population capacity estimation

Kang Liu [a,b], Ju Huang [a,b], Mingliang Xu [c], Matjaž Perc [d], Xuelong Li [e,f,*]

[a] Shaanxi Key Laboratory of Ocean Optics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, PR China
[b] University of Chinese Academy of Sciences, Beijing 100049, PR China
[c] School of Information Engineering, Zhengzhou University, Zhengzhou 450001, PR China
[d] Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, Maribor SI-2000, Slovenia
[e] School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, PR China
[f] Key Laboratory of Intelligent Interaction and Applications (Northwestern Polytechnical University), Ministry of Industry and Information Technology, Xi'an 710072, PR China

ARTICLE INFO

Article history:
Received 21 March 2021
Revised 14 May 2021
Accepted 3 June 2021
Available online 5 June 2021
Communicated by Zidong Wang

Keywords:
Remote sensing
Clustered Building Detection (CBD)
Saliency heatmap
Deep Neural Network (DNN)
Population Capacity Estimation (PCE)

ABSTRACT

Building detection is a critically important task in the field of remote sensing and it is conducive to urban construction planning, disaster survey, shantytown modification, and emergency landing, it etc. However, few studies have focused on the task of the clustered building detection which is inescapable and challenging for some relatively low space resolution images. The appearance structures of those buildings are not clear enough for the single-building detection. Whereas, it has been found that the distributions of clustered buildings are mostly dense and cellular, while the backgrounds are not. This clue will be beneficial to the clustered building detection. Motivated by the fact above and other similar density estimation applications, this work mainly focuses on the information mining mechanism of dense and cellular structure. Firstly, we propose a concept of Clustered Building Detection (CBD), which contributes to develop clustered building detection techniques of remote sensing images. Secondly, a saliency estimation algorithm is proposed to mine the prior information for the clustered buildings. Thirdly and most notably, combining with the CBD and the density saliency map, a Population Capacity Estimation (PCE) method is presented. The PCE can be easily used to estimate the population carrying capacity of certain areas and future applied for national land resource management. Moreover, a Clustered Building Detection Dataset (CBDD) from Gaofen-2 satellite is annotated and contributed for the public research. The experimental results by the representative detection algorithms manifest the effectiveness for the clustered building detection.

1. Introduction

Object detection plays a critically significant role in the field of remote sensing [1–6]. As a typical representative, building detection particularly plays a paramount role among natural disaster survey, illegal construction surveillance, shantytown modification, population capacity estimation, anti-terrorism surveillance, emergency landing, *etc* [7–11]. There are a multitude of works about building detection of remote sensing images. These existing methods can be basically divided into hand-designed feature based methods and Deep Neural Network (DNN) based methods. The hand-designed feature based methods usually require much technical expertise and skills [12]. To obtain semantic information of the man-made objects, Morphological Building Index (MBI) is proposed to describe the characteristics of shape, spectral, geometric and contextual information [7]. But the original MBI method can not deal with the problem of multi-scale building objects. In order to overcome the inherent defects of MBI, multi-scale morphological attribute index is presented to automatically extract buildings [13]. Candidate building pixels are extracted to the Maximum Stable Extremum Region (MSER) which will be fused into Independent Component Analysis (ICA). And then the geometric features are used to choose final buildings [9]. Inspired by the discrimination of the observed geometric features, Geometric Building Index (GBI) is also proposed for accurate building detection [14]. Combining statistical method with Plateau Objective Function (POF), space statistical optimization technique is utilized to extract building footprint [10]. Although solving some problems of the building extraction/detection, these methods still have many limitations for

* Corresponding author at: School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, PR China.
E-mail address: i@nwpu.edu.cn (X. Li).

https://doi.org/10.1016/j.neucom.2021.06.002
0925-2312/

high-resolution remote sensing images. The reason is that the scale variety, illumination intensity, different shapes, and complex backgrounds of the remote sensing images are challenging to be dealt with perfectly.

Reaping huge fruits from the theory and practice of DNN based methods, and in order to overcome the limitations of the traditional building detection models, a multitude of DNN based studies on building detection have emerged [15–17]. Shahzad et al. [18] adopted a Fully Convolutional Network (FCN) to solve the problem of automatic building detection. To obtain the mid-level semantic information, Li et al. [19] proposed a cascaded DNN architecture incorporating Hough transform algorithm. To represents the building's high level features in the task of rural building detection and positioning, Sun et al. [20] utilized a two-phase Convolutional Neural Network (CNN) model to mine hierarchical features. And Xu et al. [21] captured the multiscale representation through a Feature Pyramid Network (FPN) which attempts to hierarchically learn much more discriminative features by combining the global semantic structures and local attention details. And the work [22] takes full advantage of the multi-resolution features to distinguish objects and cluttered backgrounds. To obtain a good universality and robustness, Dong et al. [23] designed suitable object scale features for the CNN model. To manage the problem of scale variety, Hamaguchi et al. [24] integrated several CNN models to a robust unified model. Each of the CNN model is a specialist for a certain size buildings. A multi-branch conditional Generative Adversarial Network (GAN), the first GANs-based data augmentation, is proposed to release the problem of sample diversity [25]. Alidoost et al. [26] investigated the ability of CNN model for building detection. The CNN model is also used to verify the roof identification with a single aerial image. For detecting small and dense buildings, Shu et al. [27] proposed a center point guided method which is an end-to-end model for both training stage and testing stage. This method aims at finding possible so-called center point proposals for subsequence refinement module. Whereas, Ji et al. [28] presented a Siamese weight-shared U-Net network and contributed a multi-source dataset for building detection, especially conducive to large-size buildings. In order to detect buildings in arbitrary direction, Yang et al. [29] proposed a U-Rotation Detection Network (U-RDN) to detect the bounding boxes. To solve the regional mismatching problem, Bai et al. [30] employed a Density Residual Network (DRNet) and Region of Interest (RoI) to align the texture information. In [31], Yang et al. proposed a clustered object detection (ClusDet) network which is an end-to-end framework. The key components of ClusDet include a cluster proposal sub-network, a scale estimation sub-network, and a detection network. To alleviate the computational time cost, Xie et al. [32] presented a locally constrained YOLO [33] framework which is a one-stage algorithm and the computation approximates real time. Integrating with saliency map, Du et al. [34] proposed a saliency-guided single shot detector to suppress clutter in complex scenes. Applying the condition random field (CRF) and visual saliency, a hierarchically coarse-to-fine model with a coarse layer and a fine layer is proposed to detect the airport of remote sensing images [35]. Based on the heatmap saliency, a Target Heatmap Network (THNet) is proposed to address the problems of huge storage and time consumption [36].

The building detection works reviewed above are mostly about the buildings which have clear appearance structure. We can seldomly find any detection works about the clustered buildings. The remote sensing objects are not only sparse and non-uniform, but also tend to be highly clustered in certain regions. Take buildings for example, the remote sensing images consist of sparse buildings and clustered buildings. For sparse buildings, the appearance structures are clear so that the objects can be distinguished from the surrounding background. While for the clustered buildings, the appearance structures are not as clear as the sparse buildings. Especially, for some low space resolution images, the buildings are very small in terms of pixels, making them hard to be distinguished from the surrounding background. However, the Clustered Building Detection (CBD) is equally significant for the applications of the remote sensing images. The CBD also possesses wide applications among urban planning, natural disaster survey, illegal construction surveillance, anti-terrorism surveillance, and emergency landing. Typically for some remote sensing platforms, the spatial resolution is not clear enough for extracting structure information of single buildings. Take Gaofen-2 satellite as an example, some new challenging problems occur when performing building detection on the 4-meter multi-spectral images.

We have observed that the distributions of clustered buildings are mostly dense cellular, while the backgrounds not. This clear clue will be beneficial to the detection of clustered buildings. The knowledge of dense and cellular structure can be discovered to predict the clustered buildings of remote sensing images. Another clue and inspiration is from the works of the density estimation applications introduced in Section 2, such as crowd counting estimation, traffic jams, calculating cells and bacteria from microscopic images, and animal population estimates for ecological surveys, etc. These scenarios have similar clustered aggregation characteristics with the clustered buildings. Motivated by the facts mentioned above, this work mainly focuses on the information mining mechanism of dense and cellular structure. The **contributions of the work are fourfold** and summarized as follows:

1) We firstly propose the concept of Clustered Building Detection (CBD). The task of the CBD is a medium task between the single building detection and the semantic classification. The **major challenge** of CBD is how to detect the clustered buildings without clear appearance structures from relatively low space resolution images. The concept of CBD contributes to develop the detection techniques for clustered objects.
2) A density saliency algorithm is proposed to mine the information of dense and cellular structure for the CBD. This algorithm is beneficial to reduce the computation time in detection stage, because the density saliency with low probabilities can be regarded as the backgrounds.
3) A new Population Capacity Estimation (PCE) method is proposed. This method can easily predict the population capacity of certain rural areas. The PCE combines with the CBD and the density saliency map which can be used for the land resource monitoring in ther future. This is a significant application exploration for domestic satellites.
4) A Clustered Building Detection Dataset (CBDD) is contributed for the public research. The dataset contains totally 1564 samples and is manually annotated. The clustered buildings mainly contains rural clustered buildings. The images of CBDD are from the Gaofen-2 satellite which is China's first independently developed civilian optical remote sensing satellite with a sub-meter spatial resolution.

The remainder sections of this paper are organized as follows. Some related works are reviewed in Section 2. The proposed methodology is introduced in Section 3 which includes three subsections. The flow architecture of the proposed method is showed in Fig. 1. Combining with the flow architecture, we detailly introduce the three subsections. The experiments and results are explained in Section 4. Finally, we make the conclusion in Section 5.

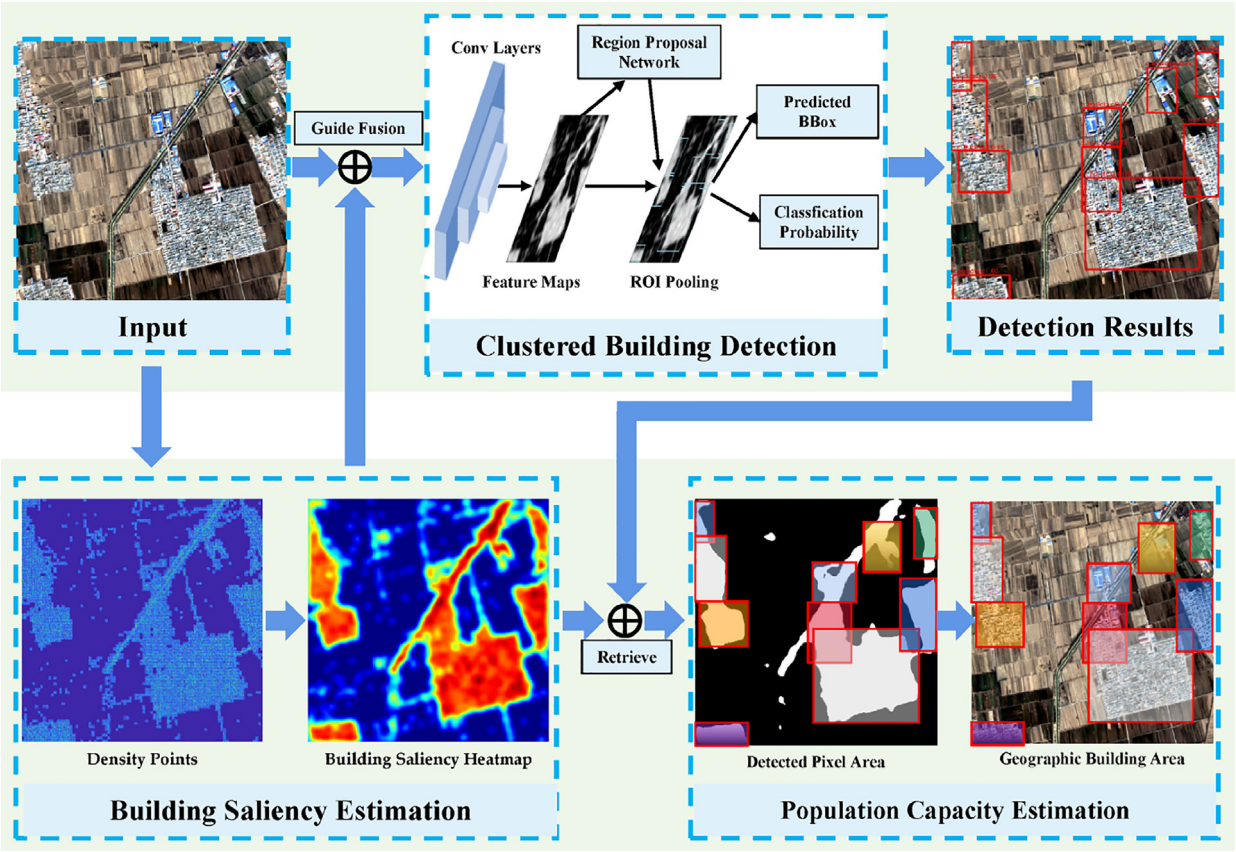

Fig. 1. Flowchart of the proposed methodology. The proposed methodology mainly contains three components. The first component is building saliency estimation which consists of the generation of density points and the estimation of saliency heatmap. The second component is Clustered Building Detection (CBD) which is conducted on the input image with the guided fusion of the saliency heatmap. The third component is Population Capacity Estimation (PCE), and in which the detection results retrieve the geographic building areas from the saliency heatmap.

2. Related works

The clustered building detection of the proposed methodology in this paper are based on the density saliency estimation and object detection methods. Therefore, the typical density estimation methods and object detection methods are reviewed in this section.

2.1. Density estimation

It is becoming a hot topic that the density estimation is used for the crowd counting estimation. The crowd count estimation can be used for mass estimation in wide applications, such as political rallies, civil unrest, sports activities, *etc* [37–39]. In addition, crowd counting methods also have great potential to handle the similar tasks in other areas, such as estimating the number of vehicles in traffic jams, calculating cells and bacteria from microscopic images, and animal population estimates for ecological surveys, *etc* [40–42].

In existing datasets for crowd counting, the person is usually annotated as a point. The annotated point usually is the center of the person head. Through a Gaussian kernel, the annotated point maps will be converted into density maps, and these heatmaps will be regarded as ground truth for training density map generating models. The common model of the crowd counting estimation is defined as in Eq. 1.

$$\mathbf{D}(x_m) = \sum_{n=1}^{N} N(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2\times 2}), \tag{1}$$

where $\mathbf{D}(\mathbf{x}_m)$ denotes the density map, the term $\mathbf{x}_m$ is a vector of two-dimension coordinate of image pixel, the term $\mathbf{z}_n$ is a vector of two-dimension coordinate of annotated points. The $\mathcal{N}(\mathbf{x}_m; \mathbf{z}_n, \sigma^2 \mathbf{1}_{2\times 2})$ is the 2D Gaussian distribution. $\sigma^2 \mathbf{1}_{2\times 2}$ is the covariance matrix. The loss function is usually defined as Eq. 2.

$$Loss = \sum_{m=1}^{M} F(\mathbf{D}^{gt}(x_m) - \mathbf{D}^{est}(x_m)), \tag{2}$$

where the term $\mathbf{D}^{gt}(x_m)$ is the density map of the ground truth, and $\mathbf{D}^{est}(x_m)$ is the density map of the estimated density map. The $F(\cdot)$ is a distance function.

For crowded person density estimation in aerial images, the work [38] proposed a Bayesian linear regression method to learn a mapping function from local features to crowd density. Using the local features of invariant color components, the work [40] formed a Probability Density Functions (PDF) for sequence images to capture density information of people. Based on the former idea, the work [42] made a crowd motion estimation for the aerial images. The work [41] developed a real-time monitoring method for crowd counting combining video surveillance and GIS. The method firstly obtained crowd counting models for each camera utilizing statistics regression methods. And then the monitoring system captured, analyzed, and presented all crowd counting information integrating a large number of cameraes and GIS. The work [39] contributed a big synthetic dataset without manpower annotations. This dataset can give a more changeable environment, larger range number of people for the crowd counting study. In order to overcome the limitation of the ground-truth density

map, the work [43] proposed a Bayesian loss from the point annotations, and formed a density contribution probability model.

The evaluation metrics of the crowd counting estimation are usually Mean Absolute Error (MAE) and Mean Square Error (MSE). In certain space resolution of remote sensing images, the clustered buildings have similar appearances with the crowd scene. Hence, these density estimation works can provide some inspirations for the preliminary estimation of the Clustered Building Detection (CBD).

2.2. Object detection methods

Object detection mainly focuses on the positioning and classification of objects in discrete images or sequential videos. This task is a fundamental problem in the field of computer vision. The related techniques can be widely used in the video surveillance, robot navigation, industry detection, and so on [44]. The traditional object detectors are mainly based on the hand-crafted feature [45–50]. The VJ detector [45], HOG detector [46], DPM detector [49] are typical representatives of the traditional detection methods. The VJ detector [45] is firstly regarded as the real-time method for human face detection. Motivated by the requirement of pedestrian detection and scale-invariant feature transform, the HOG detector [46] is a great improvement of feature representation for the task object detections. The DPM detector [49] achieved the milestone of the traditional approaches of object detection. The DPM is an extension of HOG detector and the core idea is "divide and conquer". The inference of DPM obtains the final results by assembling the results of different object components. However, these hand-crafted detectors need lots of expert experiences. The generalization ability of these classifiers can not meet the need of wide-range applications.

As the development of the Deep Neural Networks (DNN), a magnitude of DNN-based methods of object detection emerged [51]. These methods can be divided into two-stage methods and one-stage methods. Usually, the two-stage methods find the region proposals at the beginning, and then predict the location and category of proposal objects in the next step. The R-CNN [52] firstly utilized the selective search method to generate region proposals. Secondly, the AlexNet to used to extract the features from the fix-resized image. Finally, the features from the candidates are predicted by SVM classifier. In order to overcome the limitation of the R-CNN [52], the Fast R-CNN [53] which supports the bounding box regression is proposed. However, the computation is still time consuming. Ren et al. [54] combines the region proposal network (RPN) to propose a Faster R-CNN. The Faster R-CNN is the first end-to-end detector of deep network. For generic instance segmentation, He et al. added a segmentation branch to the Faster R-CNN. This architecture is called Mask R-CNN [55].

In some works, the object detection is regarded as a regression problem and implemented in one stage. These one-stage methods predict the object location and category in an unified model, such as YOLO [33], SSD [56], *etc*. Different from the two-stage methods, the computational time of YOLO [33] method is relatively faster. The YOLO method includes feature extraction and bounding box detection. First of all, the YOLO divides the input image into $S \times S$ grids. And next the bounding box and category probability of each grid will be separately predicted. Whereas, the SSD [56] method can balance the precision and the computational speed by combining the anchor-mechanism and regression theory. In order to settle the imbalanced problem between the negative samples and positive samples, the RetinaNet [57] designed a Focal Loss, and achieved the nearly mean average presicion with the two-stage methods.

The anchor-based methods above usually select certain points with certain step length in images. And then based on the center of each anchor, multiple bounding boxes are designed with fixed height and width. These bounding boxes are used for predicting the potential objects. However, a big problem of anchor-based methods is how to design proper anchors and bounding boxes.

Recently years, a large number of anchor-free methods are proposed, and these methods are mainly focus on the key points of the objects. The CornerNet [58] formed a heatmap for each bottom-right corner. The idea is that the detected paired corner points have similar predicted embedding vectors. Some similar appearances can generate similar embedding vectors, so this hypothesis is not always useful. Adding a center point to the CornerNet [58], the CenterNet [59] utilized a triplet to represent keypoints. However, in the situation of dense objects, this method can not conduct perfectly. The CentripetalNet [60] attempted to avoid the shortcomings of the CornerNet [58] and CenterNet [59], combining with the centripetal shift of the corner keypoints.

The object detection methods have been successfully applied in the natural scenes. With the evolution of huge demands, object detection equally plays an important role in the remote sensing field, such as airplane detection, ship detection, airport detection, building detection, *etc*.

3. The proposed methodology

The most existing works do not focus on the problem of Clustered Building Detection (CBD) for remote sensing images. However, this problem is inescapable and challenging for some low-space resolution remote sensing images. It has been found that the distributions of clustered buildings are mostly dense and cellular, while the backgrounds are not. This clue is beneficial for the detection of clustered buildings. Digging the knowledge of dense and cellular structure can help predicting the clustered buildings of the remote sensing images. Motivated by the fact mentioned, this work mainly focuses on the information mining mechanism of dense and cellular structure.

In the following subsections, we present the framework of the proposed CBD and Population Capacity Estimation (PCE). The flow architecture of the proposed method is showed in Fig. 1. The proposed methodology mainly contains three components which are introduced as following three subsections. The first component is building saliency estimation which consists of the generation of density points and the estimation of saliency heatmap. The second component is the CBD which is conducted on the input image fused by the saliency heatmap. The third component is the PCE in which the detection results retrieve the geographic building areas from the saliency heatmap.

3.1. Building saliency estimation

3.1.1. Density points

Observing the input image, the texture of the clustered building is obviously different from the backgrounds. The appearance of the clustered buildings are cellular structure, while the backgrounds are not. Besides, the goal of this work is to find an explainable way to the clustered building detection. In order to obtain the building saliency heatmap of the cellular structure which is conducive to the building detection, we firstly need to obtain the map of density points which are dense in the clustered building areas. The map of density points can be converted into building saliency heatmap based on the statistical bins and Gaussian kernel. In order to obtain the map of density points, the edge detection is performed and then the edge map E is obtained and based on Eq. (3).

$$E(x,y) = \alpha \times |E_X(x,y)| + (1-\alpha) \times |E_Y(x,y)| \tag{3}$$

where the $|E_X|$ and $|E_Y|$ is gray value of horizontal edge detection and vertical edge detection for the image separately, the term α is a weighted factor and it is set as 0.5 in the paper. In order to calculate the saliency map, we change the edge map E into the binary image E_T using the Eq. (4)

$$E_T(x,y) = \mathrm{sgn}(E(x,y) - T), \tag{4}$$

where the term T is the threshold for the binary image, $\mathrm{sgn}(\cdot)$ is sign function.

Thus, the density points are estimated. Compared to the crowd counting dataset, these density points of clustered buildings can be directly obtained only using texture estimation method without manual annotations.

3.1.2. Saliency heatmap

The map of density points is utilized for generating the saliency heatmap. Corresponding to the E_T, let $C = \{C_{1,1}, C_{1,2}, \ldots, C_{P,Q-1}, C_{P,Q}\}$ denotes the two dimensional statistical bins,

$$C_{i,j} = \sum_{x=1}^{M}\sum_{y=1}^{N} E_{T_{i,j}}(x,y), \tag{5}$$

where $P,\ Q$ are separately statistical bin number of the horizontal axis and the vertical axis, and M, N are the height and width of single bin. Through the operation of Gaussian kernel, the filtered image F is calculated as in Eq. 6. The pseudo color saliency heatmap is also shown in Fig. 2.

$$F(x,y) = \frac{1}{\sum_{i=-r}^{r}\sum_{j=-r}^{r} G(i,j)} \sum_{i=-r}^{r}\sum_{j=-r}^{r} C(x+i, y+j)G(i,j), \tag{6}$$

where the term r is the radius of the Gaussian kernel $G(i,j)$ which is defined in the following Eq. (7).

$$G(i,j) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(i-i_0)^2 + (j-j_0)^2}{2\sigma^2}\right), \tag{7}$$

where the (i_0, j_0) is center point of the certain sliding window, σ^2 is variance of the $G(i,j)$.

Three examples with the saliency heatmaps for the clustered building images are showed in Fig. 2. The heatmap can reasonably express the visual saliency of the clustered buildings. Although some pixels are mistakenly identified as saliency of high probability, most of them can be excluded in the object detection stage.

3.2. Clustered building detection

3.2.1. Guide filter

The original images consist of complicated backgrounds which may interfere the object detection and reduce the precision of the detection metrics. Therefore, the input image is filtered by the Gaussian heatmap H. Here we design two strategies: Soft Guide Filter (SGF) and Hard Guide Filter (HGF). The SGF only uses the probabilities of the heatmap to filtered the original image in Eq. (8).

$$I_o(x,y) = I(x,y) \times H(x,y), \tag{8}$$

where $H(x,y) = \frac{F(x,y)}{\max(F) - \min(F)}$. However, this strategy may be not beneficial to reserving the color information of the clustered building. Therefore, we also design the second strategy HGF in Eq. (9). The original image is filtered in binary mask obtained with a threshold on the heatmap. In this case, the color information of the clustered building is reserved. In addition, the possibly mistaken detections can be reduced.

$$I_o(x,y) = I(x,y) \times \mathrm{sgn}(H(x,y) - T_H), \tag{9}$$

where the term T_H is the threshold of the heatmap, $\mathrm{sgn}(\cdot)$ is a sign function.

3.2.2. Detection stage

The object detection for the clustered buildings is conducted and the detection algorithms are performed on the guide fusion map. In this work, the Faster R-CNN [54] is adopted as the primary object detection tool. So the main framework of the Faster R-CNN [54] is shown in Fig. 1. Through convolution layers, the feature maps are obtained. And then the Region Proposal Network (RPN) is utilized to extract region candidates from the feature maps. Applying ROI pooling, the proposal feature maps are acquired, and they are used to calculate the classification information and localization information. By means of the full connection and softmax in the final layer of DNN, the classification probability can be calculated. The classification score denotes the recognition probability of certain category. The localization of objects is usually regarded as a regression problem. The total optimization loss is the sum of classification softmax loss and box regression loss.

3.3. Population Capacity Estimation

In 2016, Jean et al. [61] utilized deep learning framework to estimate country's socio-economic indicators, such as the consumption expenditure and family asset. The work adopted the high-resolution remote sensing satellite images, urban nightlight data as well as the family annual income and expenditure data from the World Bank. Five African countries, such as Nigeria, Tanzania, Uganda, Malawi, and Rwanda, are selected to demonstrate the effectiveness and feasibility. Inspired by this case and in order to explore a promising application based on the Gaofen-2 satellite images, a population capacity estimation (PCE) model of certain area is established. The population capacity is related to the building areas and the population density. Therefore, the Geographic Building Area (GBA) firstly needs to be estimated.

Rather than using rectangle predicted boxes, the saliency heatmap are more reasonable to calculate the GBA. The rectangle predicted boxes contains many non-building areas which will cause the results be far away from the real value. So the GBA can not be directly calculated on the predicted pixel boxes. Whereas, the saliency heatmap can be assisted to estimate the GBA. First, the components of the saliency heatmap will be selected by using the predicted boxes of detection results. The pixel areas of these selected clustered buildings will be calculated. In order to calculate the pixel areas for the population estimation, the building saliency heatmap are utilized with the threshold T_H which is defined in Eq. 9. The term T_H is set according to the parameter experiments. Therefore, the pixel areas are calculated by statistical method on the building saliency heatmap with the threshold T_H. Then the GBA can be obtained through a conversion coefficient. Let A_{avg} is the average living area and defined in Eq. (11). Thus, the Population Capacity Estimation (PCE) can be calculated as Eq. (10).

$$PCE = \sum_{i=1}^{L} \frac{A_i \times f \times PR_i}{A_{avg}}, \tag{10}$$

where the term A_i is the pixel area of the i-th detected clustered buildings. The term f is the conversion coefficient which converts the pixel contour area to the geographic area, and it is decided by optical parameter of the remote sensing satellite. The PR_i is the Plot Ratio of the i-th geographic clustered building, and which is the proportion of the real building areas for the geographic area. The term L is the connected domain number of the detected clustered buildings.

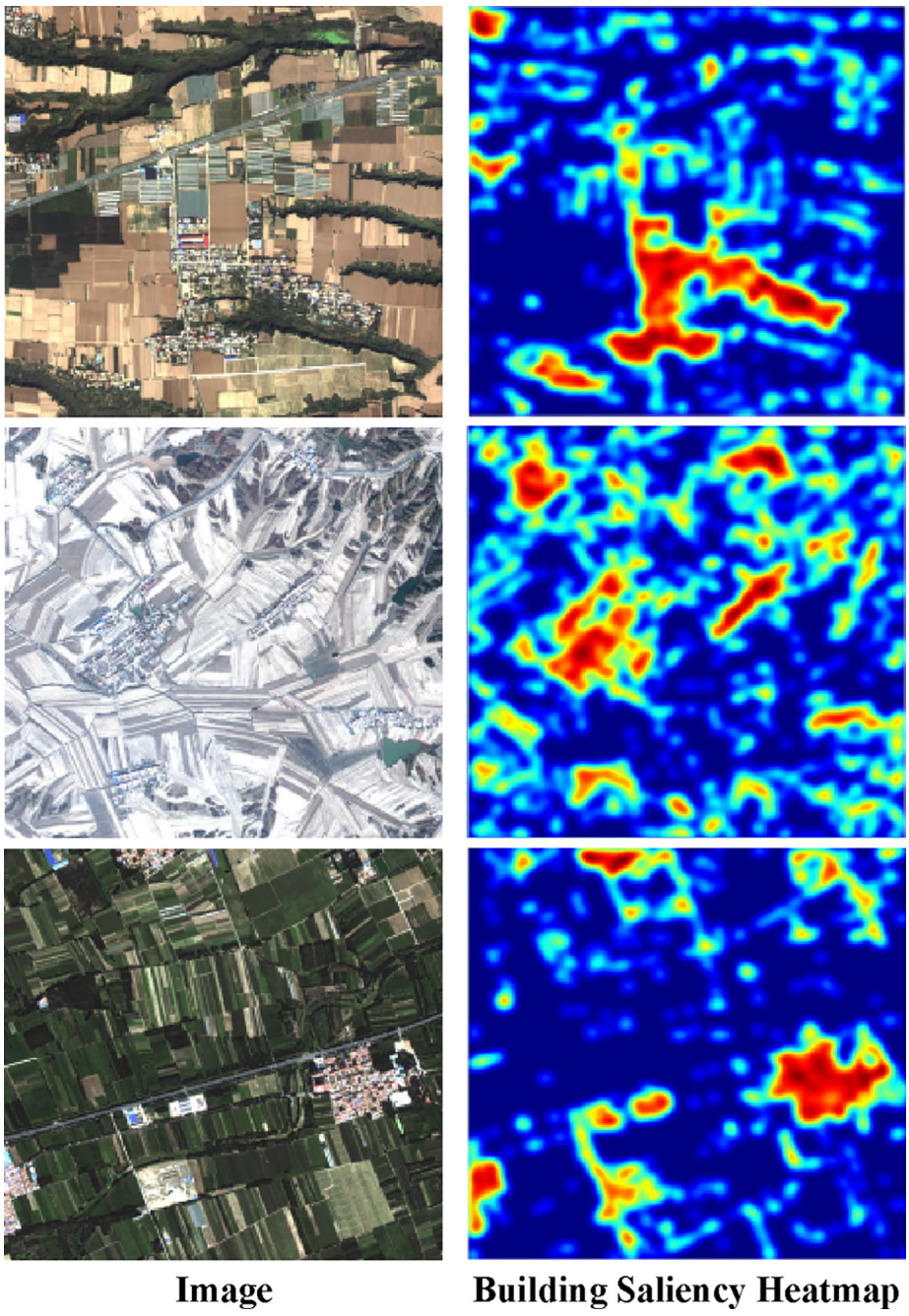

Fig. 2. Examples of saliency heatmap of clustered building images. The clustered building images are in the left column, and the relative saliency heatmaps are in the right column.

According to the statistics data form the National Bureau of Statistics of China,[1] the average living area in urban area is 39.8 m^2, and that in rural area is 48.9 m^2 in the year 2019. Thus, the A_{avg} is defined as following Equation.

$$A_{avg} = \begin{cases} 39.8 & if\ RC = 1 \\ 48.9 & otherwise \end{cases}, \tag{11}$$

where the RC denotes regional category, and which can be divided into urban area and rural area. Let $RC \in \{1, 0\}$ denotes the set of regional category. The elements in this set separately denote urban area and rural area.

4. Experiments and results

4.1. Dataset

In order to verify the Clustered Building Detection algorithms, one publicly available Clustered Building Detection dataset (CBDD) is contributed and can be downloaded from Baidu Netdisk.[2] The images of the CBDD are sampled and annotated from the Gaofen Image Dataset (GID) [62]. The GID is a large-scale dataset for land use classification, and of which the images are from Gaofen-2 satellite. The Gaofen-2 satellite belongs to the High-Definition Earth

[1] http://www.stats.gov.cn/.

[2] https://pan.baidu.com/s/1w8zcX14Q-wJABh2ePSf6aw (password: 4aiv).

Table 1
Technical payload specification of the Gaofen-2 satellite.

Payload	Spectrum Range (um)	Spatial Resolution (m)	Scanning Range (km)	Side Swing Angle	Revisit Time (day)
Panchromatic	0.45 ∼ 0.90	1	45 (Two Cameras)	±35°	5
Multispectral	0.45 ∼ 0.52 (blue) 0.52 ∼ 0.59 (green) 0.63 ∼ 0.69 (red) 0.77 ∼ 0.89 (near infrared)	4	45 (Two Cameras)	±35°	5

Table 2
The building statistics of CBDD. The scale division follows the COCO.

Scale	Instance Number	Percentage	Pixel Number
small	6	0.0007	$< 32^2$
medium	5354	0.5860	$32^2 \sim 96^2$
large	3776	0.4133	$> 96^2$

Observation System (HDEOS) which is promoted by China National Space Administration (CNSA). As the second satellite of the HDEOS, the Gaofen-2 is the first sub-meter remote sensing satellite of China. The main technical specifications of Gaofen-2 satellite is shown in Table 1. Gaofen-2 carries two panchromatic and multispectral (PMS) sensors. On the board of Gaofen-2 satellite, the scanning range is 45 km with 2 combined swath. The resolution of panchromatic sensor is 0.8 m in the sub-satellite point, and that is 3.24 m of the multi-spectral sensor. The panchromatic sensor is with effective spatial resolution of 1 m, and the multi-spectral sensor is with effective spatial resolution of 4 m. The field of view of single camera is 2.1°. The global covering observation of Gaofen-2 satellite is within 69 days, and its repeat observation is within 5 days. Since its launch in 2014, Gaofen-2 has played a significant role in a host of applications, such as land resource survey, environmental monitoring, crop estimation, and construction planning, *etc*.

The proposed CBDD is sampled from the GID, and the space resolution of CBDD is 680 ×720. To achieve the goal of CBD, the experiment dataset CBDD is manually labeled in the form of the PASCAL VOC.[3] The dataset contains 1564 images and the same number annotation files. To facilitate the experimental comparison, the format of the CBDD is also converted into another version in the form of COCO.[4] The GID has 150 samples from Gaofen-2 satellite. Each of them is with a spatial resolution of 6800 × 7200 pixels fully covered 506 km^2 geographic area. The spatial resolution of GID is 4 m, and the images are acquired from the multispectral sensors of Gaofen-2 satellite. Besides, these images are much diverse from not less than 60 China's cities.

The building statistics of CBDD is in Table 2. The pixel numbers of building objects range from hundreds to hundreds of thousands. Particularly, the scale division of building instances follows the COCO, thus the instance number of small clustered building is 6 and the percentage is 0.0007. Therefore, the experiments in this paper mainly focus on the medium scale instances and large scale instances. The small clustered buildings do not be taken into consideration. The samples consist of mountain, great plain, river, gobi, desert, *etc*. The backgrounds of the samples differ tremendously, and which are also showed in Fig. 3.

4.2. Metrics and setups

4.2.1. Metrics

The COCO metrics are adopted to quantitatively evaluate the proposed method. In the field of object detection, the Intersection over Union (IoU, Eq. (12)) means the overlap ratio between the predicted boxes and the ground truth boxes.

$$IoU = \frac{B_{predict} \cap B_{gt}}{B_{predict} \cup B_{gt}}, \tag{12}$$

where the $B_{predict}$ is the area of predicted box, and the B_{gt} is the area of ground truth box. The COCO metrics are based on the *precision* and *recall* as Eq. 13 and Eq. (14).

$$precision = \frac{TP}{TP + FP}, \tag{13}$$

$$recall = \frac{TP}{TP + FN}, \tag{14}$$

where the term *TP* is True Positives, *FP* is False Positives, and *FN* is False Negatives. These terms are all calculated based on IoU.

The COCO metrics include $AP, AP_{50}, AP_{75}, AP_s, AP_m$, and AP_l. The *AP* is the average precision over different IoU thresholds. The AP_{50} is equal to the PASCAL VOC metric and the IoU threshold is 0.50. The AP_{75} is a much strict metric with IoU threshold of 0.75. The AP_s is for the small objects which areas are smaller than 32^2. The AP_m is for medium objects which areas are between 32^2 and 96^2. The AP_l is for large objects which areas are bigger than 96^2. The series of *AP* metrics are comprehensive indicators for the object detection.

4.2.2. Setups

In the experiments, we divided the annotated dataset into train subset, validation subset, and test subset with the percentage 0.25, 0.25, and 0.5. In our experiments, the train subset and validation subset are together used for the training. In the experiments, 6 representatively different methods, including Faster R-CNN [54], YOLO [33], SSD [56], RetinaNet [57], CornerNet [58], CentripetalNet [60], are selected to perform the clustered building detection.

The competitive experiments are all conducted on the MMdetection platform[5] (The MMDetection is an open source object detection toolbox based on PyTorch. It is a part of the OpenMMLab project developed by Multimedia Laboratory, CUHK[6]). The models are trained with batch size 16 on 4 Nvidia GeForce GTX 1080 GPUs (4 images per GPU). The initial learning rate is set to 0.02 and weight decay is 0.0001. The optimizer is SGD. The maximum epoch is set as 200, because most of the models achieved convergence below epoch 50. The random flip ratio is 0.5. Most of the comparative methods on MMdetection are implemented with the COCO format, so the CBDD is also converted from the PASCAL VOC format to the COCO format.

In this paper, the conversion coefficient f of Eq. (10) is set as 10.33. The f is calculated from the spatial dimension (6800 × 7200 pixels) of the GID image and its geographic area (506 km^2). The PR_i in Eq. (10) of certain detected clustered building area can not be directly obtained from the detected results. However, the buildings of the courtyards in China's rural areas are

[3] http://host.robots.ox.ac.uk/pascal/VOC/.
[4] https://cocodataset.org/.
[5] https://github.com/open-mmlab/mmdetection/.
[6] http://mmlab.ie.cuhk.edu.hk/.

Fig. 3. Examples of the annotated dataset CBDD. The examples consist of mountain, great plain, river, gobi, desert, *etc*. The samples are diverse because their backgrounds differ tremendously.

Table 3
Parameter experiments of term T_H based on the Faster R-CNN [54].

Method	Backbone	T_H	AP	AP_{50}	AP_{75}	AP_m	AP_l
Faster R-CNN [54]	Resnet-50	0.1	0.409	0.796	**0.375**	0.389	0.435
		0.2	**0.410**	0.797	0.369	**0.397**	**0.447**
		0.3	0.408	**0.802**	**0.375**	0.390	0.440
		0.4	0.401	0.791	0.351	0.383	0.437
		0.5	0.398	0.786	0.347	0.381	0.430
		0.6	0.395	0.782	0.345	0.379	0.427
		0.7	0.393	0.779	0.345	0.378	0.425

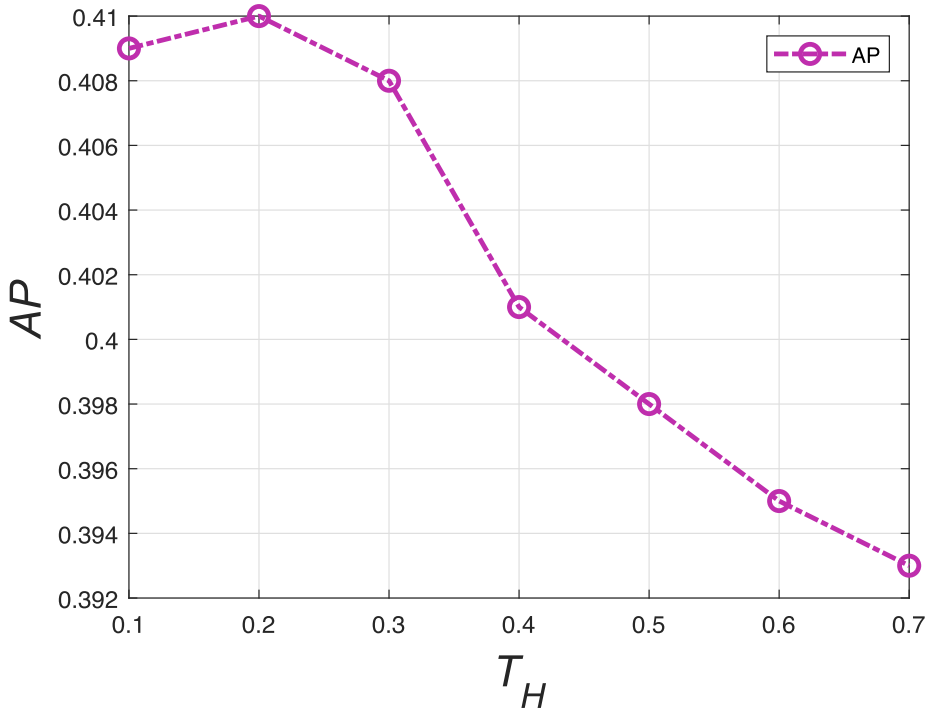

Fig. 4. The average precision (AP) curve as the heatmap threshold T_H. This experiment is based on Faster R-CNN [54] with the backbone Resnet-50.

mostly constructed in families and the roads are laying between neighbour buildings. Thus, the detected clustered buildings in this paper can be regarded as single-floor buildings, and the PR_i is set as 0.5.

4.3. Results and analysis

The specific values of the parameters in the paper are set by parameter experiments. The parameters include threshold T of

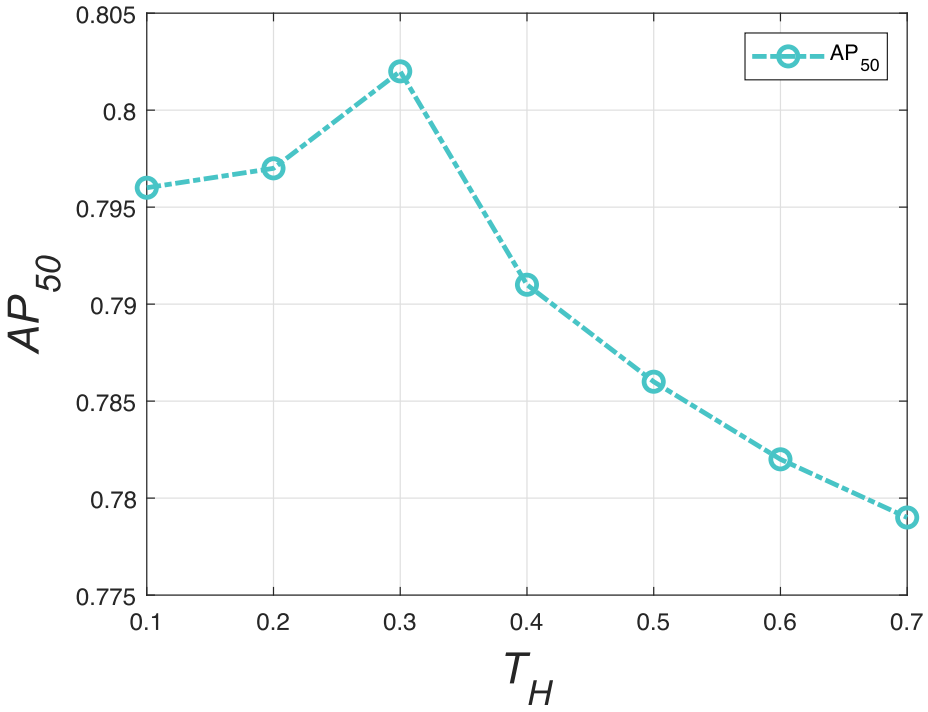

Fig. 5. The average precision (AP_{50}) curve as the heatmap threshold T_H. This experiment is based on Faster R-CNN [54] with the backbone Resnet-50.

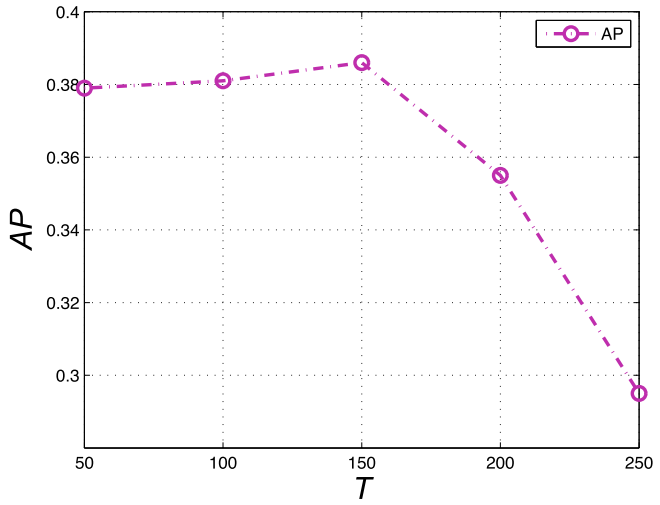

Fig. 6. The average precision (AP) curve as the threshold T of Eq. (4). This experiment is based on Faster R-CNN [54] with the backbone Resnet-50.

Eq. (4), radius r and sigma σ of Eq. (6). Some intelligent optimization algorithms [63–66] could be considered to optimize these parameters. However, these parameters are hyper-parameters in different stages, so the optimization problem is not a convex optimization problem. Therefore, we adopted alternating optimization strategy (fix other parameters) to find the appropriate parameter setting. The curves are separately showed in Figs. 6–8 in the revised manuscript.

The parameter experiment of term T_H is reported in Table 3. This experiment results will decide the setting of the term T_H. The Faster R-CNN [54] is adopted with backbone Resnet-50 to perform the parameter experiment. The number of the probability heatmap is floating, so the T_H is between 0.0 and 1.0. For more intuitive display, the primary indicator AP and AP_{50} are selected to plot in Figs. 4 and 5.

From the figures and the table, we can see the best metrics are achieved when the term T_H is set as 0.2 and 0.3. Therefore, the T_H can be set between 0.2 and 0.3. In the following experiments, the T_H is set as 0.25.

The parameter experiment of threshold T of Eq. (4) is showed in Fig. 6. The AP achieves the best result when the threshold T is set as 150. Therefore, the threshold T is set as 150 in the following experiments.

The parameter experiments of radius r of Eq. (6) and Sigma σ of Eq. (7) are separately showed in Figs. 7 and 8. Therefore, the term r is set as 24, and the term σ is set as 8 in the following experiments.

Table 4 shows the AP results of separately experimental strategies. The testing methods are based on different image scales and backbones which are using the presets for the COCO dataset. The Resnet-50 and Resnet-101 are used for the Faster R-CNN. Because the number of the small instances is small, and the experiment results of AP_s are all 0. The AP_s are not essential to report in the statistic tables. According to the preliminary experiment with the guide fusion of saliency heatmap. The method based Resnet-50 achieved better APs (the primary AP is 0.405) than that of Resnet-101 (the primary AP is 0.386). Then we add the guide fusion to the Faster R-CNN with Resnet-50. We test both of these two strategies, including soft guide fusion (SGF) and hard guide fusion (HGF). The results demonstrate the strategy of HGF can improve the AP scores of CBD, while the SGF can decrease the AP scores. The reason for this phenomenon is that the SGF can erode the available information of clustered buildings in color space. Moreover, we can find the resized scales can effect the APs, such as the testing results of YOLOv3 and SSD. Those methods which resized scales are close to the original 680×720 will have better APs.

In Fig. 9, some example pairs are displayed from column (a) to column (j). The detected results are on the top of the image pairs and the corresponding saliency heatmaps are on the bottom of the image pairs. On the whole, the detection results of the clustered buildings are reasonable and satisfactory. We can see that the actually detected objects have the most powerful energy areas

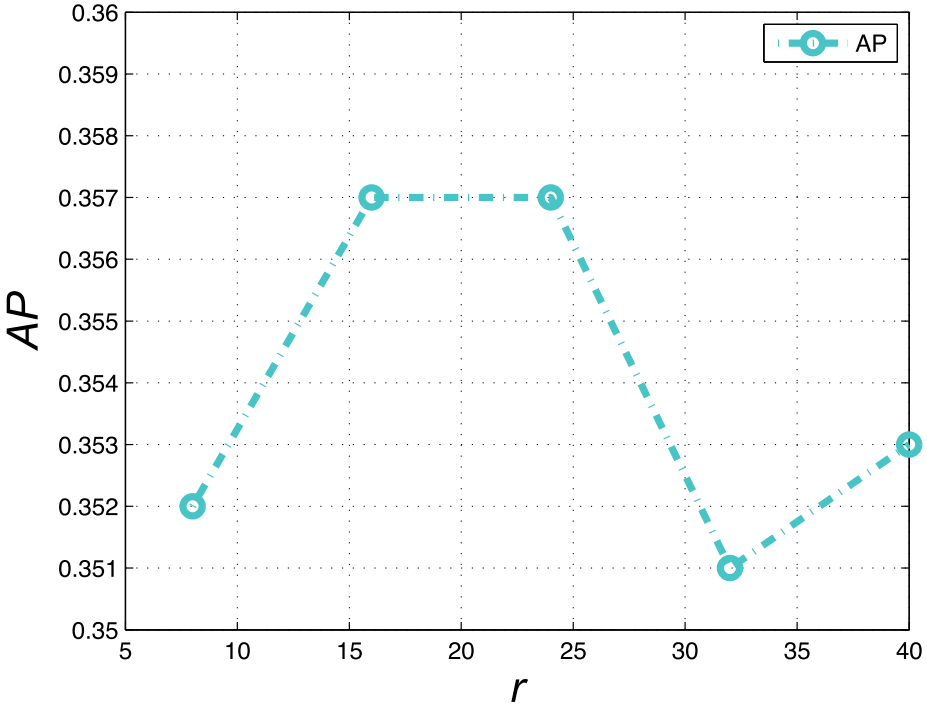

Fig. 7. The average precision (*AP*) curve as the radius r of Eq. (6). This experiment is based on Faster R-CNN [54] with the backbone Resnet-50.

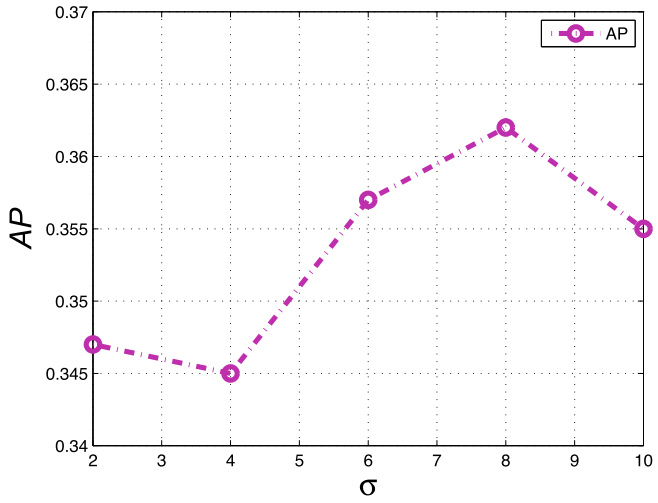

Fig. 8. The average precision (*AP*) curve as the Sigma σ of Eq. (7). This experiment is based on Faster R-CNN [54] with the backbone Resnet-50.

in the saliency heatmaps. These results can demonstrate that the saliency heatmaps of the clustered buildings can guide the detection task. First, we calculate the predicted pixel areas of the detected objects in the images. Second, we convert these predicted pixels into the GBA. Finally, we report the accordingly predicted population capacity estimation, that is predicted PCE. The predicted GBA and PCE of the examples are reported in Table 5, and also displayed in the example images of Fig. 9.

The evaluation of the PCE is not easy to make. For one thing, the true population statistics of areas are not easily accessible. Although a multitude of population statistics are possessed by the administrative department, the statistics of particular areas are not within easy reach. For another, the building areas do not have directly strong correlations to the real population. We clarify that the PCE is not the real population estimation of certain areas, but it is just an indicator of potential bearing capacity. Therefore, there is no available metric which can make an objective evaluation for the PCE. This study provides a feasibility exploration for the significant applications with Gaofen-2 and other domestic satellites.

Table 4
Clustered building detection results of competitive methods using COCO metric.

Method	Scale	Backbone	*AP*	AP_{50}	AP_{75}	AP_m	AP_l
YOLOv3 [33]	320 × 320	Darknet-53	0.291	0.724	0.162	0.293	0.300
YOLOv3 [33]	608 × 608	Darknet-53	0.294	0.733	0.161	0.292	0.296
SSD [56]	300 × 300	VGG-16	0.312	0.725	0.208	0.295	0.338
SSD [56]	512 × 512	VGG-16	0.346	0.765	0.260	0.344	0.355
RetinaNet [57]	680 × 720	Resnet-50	0.227	0.628	0.089	0.245	0.205
CornerNet [58]	511 × 511	HourglassNet	0.177	0.370	0.149	0.265	0.135
CentripetalNet [60]	511 × 511	HourglassNet	0.357	0.715	0.308	0.356	0.362
Faster R-CNN [54]	680 × 720	Resnet-101	0.386	0.758	0.352	0.378	0.422
Faster R-CNN [54]	680 × 720	Resnet-50	0.405	**0.797**	0.364	0.397	0.433
Faster R-CNN+SGF (**ours**)	680 × 720	Resnet-50	0.388	0.782	0.346	0.373	0.422
Faster R-CNN+HGF (**ours**)	680 × 720	Resnet-50	**0.410**	**0.797**	**0.369**	**0.398**	**0.447**

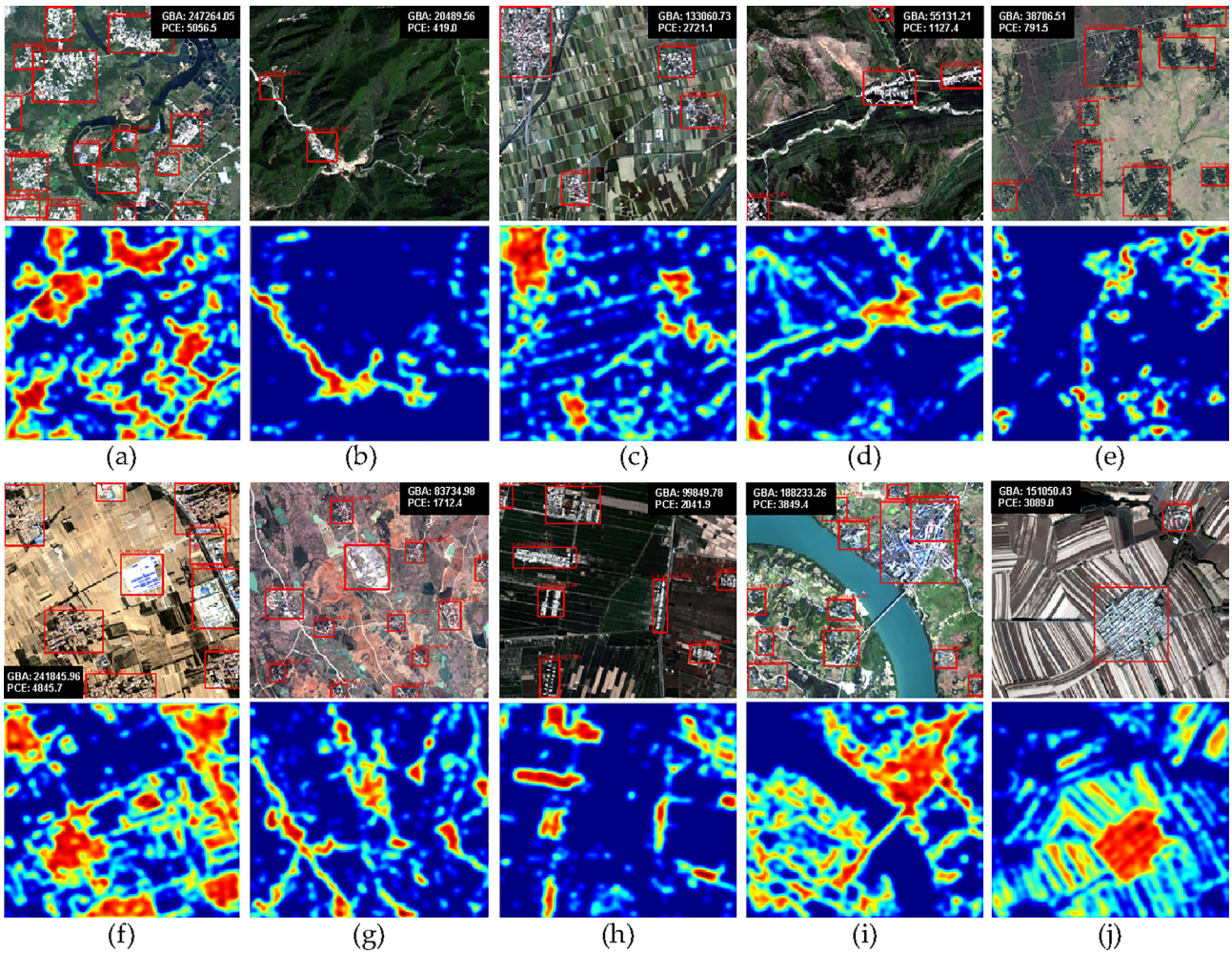

Fig. 9. Typical results of CBD by Faster R-CNN+HGF. Image pairs are showed from column (a) to column (j). The detected results are on the top of the image pairs, while the corresponding saliency heatmaps are on the bottom of the image pairs.

Table 5
The statistics of Population Capacity Estimation for some examples, and which are also displayed in Fig. 9.

Image	Pred. Pixels	Pred. GBA (m^2)	Pred. PCE
(a)	47873	247264.05	5056.5
(b)	3967	20489.56	419.0
(c)	25762	133060.73	2721.1
(d)	10674	55131.21	1127.4
(e)	7494	38706.51	791.5
(f)	46824	241845.96	4845.7
(g)	16212	83734.98	1712.4
(h)	19332	99849.78	2041.9
(i)	36444	188233.26	3849.4
(j)	29245	151050.43	3089.0

In order to show that the proposed method is practical or reasonable, the village population statistics of China is reported in Table 6. The originally statistical data is retrieved from the Ministry of Housing and Urban–Rural Development of the Peoples Republic of China.[7] The village population statistics are average population of administrative villages in different provinces and the national total. In Table 6, the average population of administrative villages are from hundreds to thousands. The experimental samples are sampled from different cities throughout the country, hence the data in Table 6 can be regarded as a rough estimation for the PCE. The calculated PCEs in Table 5 are fit with the rural population distributions of China.

Some unsatisfying examples of clustered buildings are showed in Fig. 10. The detected objects highlighted in yellow rectangles are overlapped or contained. Especially, the appearances of these ambiguity examples are mostly irregular, strip-like, large-size. This phenomenon can cause a problem that the detected results may not beneficial to the APs. It is because the APs are based on the metrics of precision and recall. In the field of object detection, the precision and recall are based on the accurate statistics of the detected objects. But the overlapped results of the clustered buildings may result in inaccurate statistics. However, these detected results can not be absolutely regarded as mistakes. On the contrary, the detected objects of these overlapped bounding boxes are clustered buildings, and they are true when taking no account of the manually annotated ground truths. There are two reasons for this phenomenon. On one hand, the clustered buildings have flexible structures which may cause ambiguity problem. On the other hand, the annotation work is subjective for persons. Take one strip-like object of image (b) in Fig. 10 as an example, one person annotates the object using single rectangle box, while the other person uses double. Accordingly, the occurrences of ambiguous results are unavoidable in testing stage. The ambiguity problem of the clustered objects could be a possible research topic in the future.

[7] http://www.mohurd.gov.cn/xytj/tjzljsxytjgb/index.html.

Table 6
The village population statistics of China (2019).

Name of regions	Average population of administrative villages	Name of regions	Average population of administrative villages	Name of regions	Average population of administrative villages
Beijing	925.24	Zhejiang	1226.43	Chongqing	2295.12
Tianjin	803.08	Anhui	2942.67	Sichuan	1334.90
Hebei	1061.72	Fujian	1474.64	Guizhou	1975.39
Shanxi	825.50	Jiangxi	1839.60	Yunnan	2548.04
Shandong	805.15	Inner Mongolia	1214.88	Tibet	451.24
Liaoning	1597.38	Henan	805.15	Shaanxi	1320.47
Jilin	1443.20	Hubei	1457.12	Gansu	1175.28
Heilongjiang	1889.93	Hunan	1457.12	Gansu	900.98
Shanghai	1967.68	Guangdong	2560.88	Ningxia	1668.60
Jiangsu	2495.42	Guangxi	2860.93	Xinjiang	1261.04
Hainan	2003.00	**National Total**	1506.24	–	–

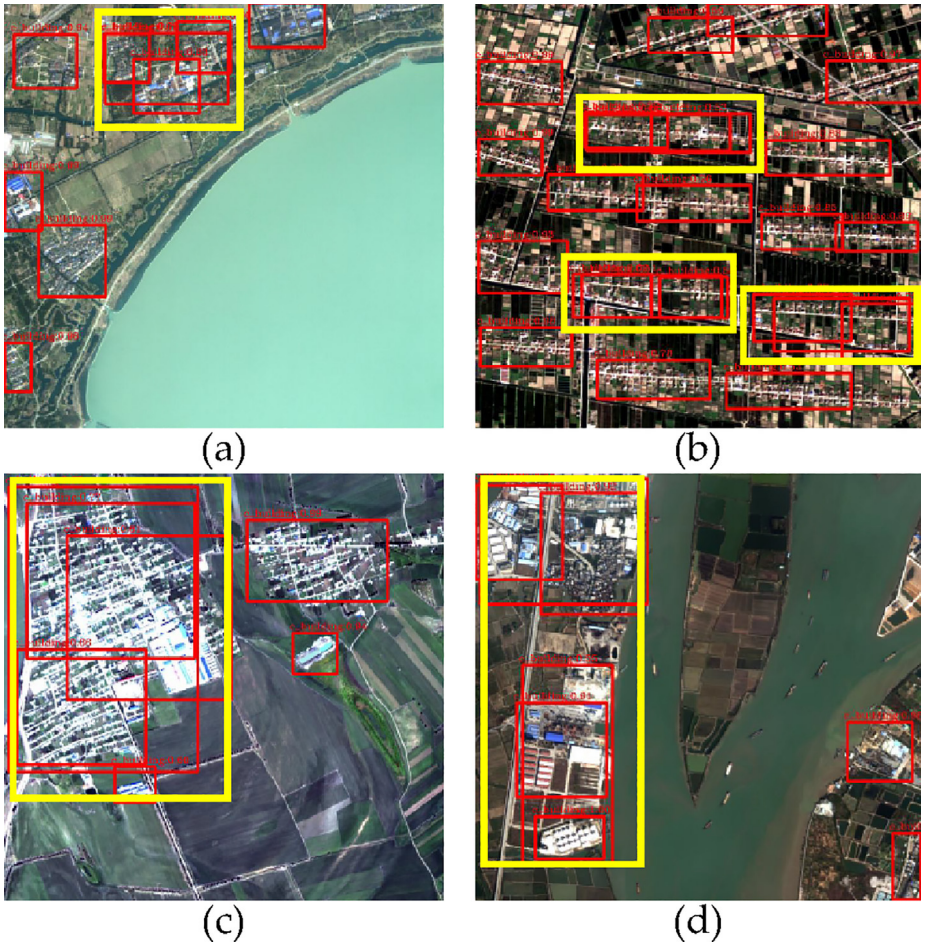

Fig. 10. Examples of unsatisfying detection results due to the flexible structures of the clustered buildings. Those objects highlighted in yellow rectangles are overlapped or contained.

5. Conclusion

In this paper, we proposed a concept of Clustered Building Detection (CBD) which mainly contributes to develop the detection techniques of cluster objects. Specifically, we proposed a building saliency estimation method and two strategies for the guide fusion of the saliency maps and original images. Typically, the proposed fusion strategy of Hard Guide Filter (HGF) for Faster R-CNN achieved the best AP scores. The AP score reached 0.410 and the AP_{50} score reached 0.797. Most notably, combining with the CBD and the density saliency map, a Population Capacity Estimation (PCE) algorithm is introduced. This algorithm can easily predict the potential population capacity of certain areas. Moreover, a Clustered Building Detection Dataset (CBDD) based on the Gaofen-2 satellite images is contributed. We hope that this newly proposed dataset CBDD will be one of the benchmark databases for the researchers to develop novel algorithms for CBD. The experimental results manifest the effects of the proposed method both

qualitatively and quantitatively. However, the ambiguity problem of the clustered objects could be a possible research topic in the future.

CRediT authorship contribution statement

Kang Liu: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Ju Huang:** Investigation, Resources, Visualization. **Mingliang Xu:** Writing - review & editing, Formal analysis. **Matjaž Perc:** Writing - review & editing, Formal analysis. **Xuelong Li:** Supervision, Project administration.

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

This work is supported by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences (No. QYZDY-SSW-JSC044), and by the National Natural Science Foundation of China (No. 61871470).

References

[1] Y. Li, C. Lin, H. Li, W. Hu, H. Dong, Y. Liu, Unsupervised domain adaptation with self-attention for post-disaster building damage detection, Neurocomputing 415 (2020) 27–39.
[2] X. Lu, W. Ji, X. Li, X. Zheng, Bidirectional adaptive feature fusion for remote sensing scene classification, Neurocomputing 328 (2019) 135–146.
[3] X. Li, M. Chen, F. Nie, Q. Wang, A multiview-based parameter free framework for group detection, in: Proc. Conference on Artificial Intelligence, 2017, pp. 4147–4153. .
[4] G. Huang, Z. Wan, X. Liu, J. Hui, Z. Wang, Z. Zhang, Ship detection based on squeeze excitation skip-connection path networks for optical remote sensing images, Neurocomputing 332 (2019) 215–223.
[5] Q. Wang, S. Liu, J. Chanussot, X. Li, Scene classification with recurrent attention of VHR remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 57 (2) (2019) 1155–1167.
[6] B. Chen, Z. Chen, L. Deng, Y. Duan, J. Zhou, Building change detection with rgb-d map generated from uav images, Neurocomputing 208 (2016) 350–364.
[7] X. Huang, W. Yuan, J. Li, L. Zhang, A new building extraction postprocessing framework for high-spatial-resolution remote-sensing imagery, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10 (2) (2017) 654–668.
[8] P.J. Wang, X. Sun, W.H. Diao, K. Fu, Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery, IEEE Transactions on Geoscience and Remote Sensing 58 (5) (2020) 3377–3390.
[9] L. Agarwal, K.S. Rajan, Integrating MSER into a Fast ICA approach for improving building detection accuracy, in: Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2018, pp. 4831–4834.
[10] M. Norman, H. Shafri, M. Idrees, S. Mansor, B. Yusuf, Spatio-statistical optimization of image segmentation process for building footprint extraction using very high-resolution Worldview 3 satellite data, Geocarto International 35 (10) (2020) 1124–1147.
[11] M. Shahzad, M. Maurer, F. Fraundorfer, Y.Y. Wang, X.X. Zhu, Buildings detection in vhr sar images using fully convolution neural networks, IEEE Transactions on Geoscience and Remote Sensing 57 (2) (2019) 1100–1116.
[12] X. Li, M. Chen, F. Nie, Q. Wang, Locality adaptive discriminant analysis, in: Proc. International Joint Conference on Artificial Intelligence, IJCAI, 2017, pp. 2201–2207. .
[13] J. Li, J. Cao, M. Feyissa, X. Yang, Automatic building detection from very high-resolution images using multiscale morphological attribute profiles, Remote Sensing Letters 11 (7) (2020) 640–649.
[14] J. Huang, G. Xia, F. Hu, L. Zhang, Accurate building detection in VHR remote sensing images using geometric saliency, in: Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2018, pp. 3991–3994.
[15] Y. Liu, Z. Zhang, R. Zhong, D. Chen, Y. Ke, J. Peethambaran, C. Chen, L. Sun, Multilevel building detection framework in remote sensing images based on convolutional neural networks, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (10) (2018) 3688–3700.
[16] M. Bhimra, U. Nazir, M. Taj, Using 3d residual network for spatio-temporal analysis of remote sensing data, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, 2019, pp. 1403–1407.
[17] K. Reda, M. Kedzierski, Detection, classification and boundary regularization of buildings in satellite imagery using faster edge region convolutional neural networks, Remote Sensing 12 (14). .
[18] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, X. Zhu, Buildings detection in VHR SAR images using fully convolution neural networks, IEEE Transactions on Geoscience and Remote Sensing 57 (2) (2019) 1100–1116.
[19] Q. Li, Y. Wang, Q. Liu, W. Wang, Hough transform guided deep feature extraction for dense building detection in remote sensing images, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, 2018, pp. 1872–1876.
[20] L. Sun, Y. Tang, L. Zhang, Rural building detection in high-resolution imagery based on a two-stage CNN model, IEEE Geoscience and Remote Sensing Letters 14 (11) (2017) 1998–2002.
[21] C.Y. Xu, C.Z. Li, Z. Cui, T. Zhang, J. Yang, Hierarchical semantic propagation for object detection in remote sensing imagery, IEEE Transactions on Geoscience and Remote Sensing 58 (6) (2020) 4353–4364.
[22] C. Li, R. Cong, C. Guo, H. Li, C. Zhang, F. Zheng, Y. Zhao, A parallel down-up fusion network for salient object detection in optical remote sensing images, Neurocomputing 415 (2020) 411–420.
[23] Z.P. Dong, M. Wang, Y.L. Wang, Y. Zhu, Z.Q. Zhang, Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features, IEEE Transactions on Geoscience and Remote Sensing 58 (3) (2020) 2104–2114.
[24] R. Hamaguchi, K. Nemoto, T. Imaizumi, S. Hikosaka, Detecting buildings of any size using integration of CNN models, in: Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2018, pp. 1280–1283.
[25] D. Zhu, S. Xia, J. Zhao, Y. Zhou, M. Jian, Q. Niu, R. Yao, Y. Chen, Diverse sample generation with multi-branch conditional generative adversarial network for remote sensing objects detection, Neurocomputing 381 (2020) 40–51.
[26] F. Alidoost, H. Arefi, A CNN-based approach for automatic building detection and recognition of roof types using a single aerial image, Journal of Photogrammetry Remote Sensing and Geoinformation Science 86 (5–6) (2018) 235–248.
[27] Z. Shu, X. Hu, J. Sun, Center-point-guided proposal generation for detection of small and dense buildings in aerial imagery, IEEE Geoscience and Remote Sensing Letters 15 (7) (2018) 1100–1104.
[28] S. Ji, S. Wei, M. Lu, Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set, IEEE Transactions on Geoscience and Remote Sensing 57 (1) (2019) 574–586.
[29] J. Yang, L. Ji, X. Geng, X. Yang, Y. Zhao, Building detection in high spatial resolution remote sensing imagery with the U-rotation detection network, International Journal of Remote Sensing 40 (15) (2019) 6036–6058.
[30] T. Bai, Y. Pang, J. Wang, K. Han, J. Luo, H. Wang, J. Lin, J. Wu, H. Zhang, An optimized Faster R-CNN method based on DRNet and RoI align for building detection in remote sensing images, Remote Sensing 12 (5). .
[31] F. Yang, H. Fan, P. Chu, E. Blasch, H. Ling, Clustered object detection in aerial images, in: Proc. IEEE/CVF International Conference on Computer Vision ICCV, 2019, pp. 8310–8319.
[32] Y. Xie, J. Cai, R. Bhojwani, S. Shekhar, J. Knight, A locally-constrained YOLO framework for detecting small and densely-distributed building footprints, International Journal of Geographical Information Science 34 (4) (2020) 777–801.
[33] J. Redmon, S.K. Divvala, R.B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2016, pp. 779–788.
[34] L. Du, L. Li, D. Wei, J.S. Mao, Saliency-guided single shot multibox detector for target detection in sar images, IEEE Transactions on Geoscience and Remote Sensing 58 (5) (2020) 3366–3376.
[35] X. Yao, J. Han, L. Guo, S. Bu, Z. Liu, A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and crf, Neurocomputing 164 (2015) 162–172.
[36] H. Chen, L. Zhang, J. Ma, J. Zhang, Target heat-map network: An end-to-end deep network for target detection in remote sensing images, Neurocomputing 331 (2019) 375–387.
[37] G. Gao, Q. Liu, Y. Wang, Counting dense objects in remote sensing images, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP, 2020, pp. 4137–4141.
[38] C. Shiyong, O. Meynberg, P. Reinartz, Bayesian linear regression for crowd density estimation in aerial images, in: Proc. Joint Urban Remote Sensing Event JURSE, 2017, pp. 1–4. .
[39] Q. Wang, J. Gao, W. Lin, Y. Yuan, Learning from synthetic data for crowd counting in the wild, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2019, pp. 8198–8207.
[40] B. Sirmacek, P. Reinartz, Automatic crowd analysis from airborne images, in: Proc. 5th International Conference on Recent Advances in Space Technologies, 2011, pp. 116–120.
[41] H. Song, X. Liu, X. Zhang, J. Hu, Real-time monitoring for crowd counting using video surveillance and gis, in: Proc. 2nd International Conference on Remote Sensing, Environment and Transportation Engineering, 2012, pp. 1–4. .
[42] B. Sirmacek, P. Reinartz, Automatic crowd density and motion analysis in airborne image sequences based on a probabilistic framework, in: Proc. IEEE International Conference on Computer Vision Workshops ICCV Workshops, 2011, pp. 898–905. .
[43] Z. Ma, X. Wei, X. Hong, Y. Gong, Bayesian loss for crowd count estimation with point supervision, in: Proc. IEEE/CVF International Conference on Computer Vision ICCV, 2019, pp. 6141–6150.

[44] X. Li, B. Zhao, Video distillation, Science China Information Sciences doi:10.1360/SSI-2020-0165. .
[45] P.A. Viola, M.J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[46] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2005, pp. 886–893.
[47] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.
[48] P.A. Viola, M.J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. Computer Society Conference on Computer Vision and Pattern Recognition CVPR, 2001, pp. 511–518.
[49] P.F. Felzenszwalb, D.A. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2008. .
[50] P.F. Felzenszwalb, R.B. Girshick, D.A. McAllester, Cascade object detection with deformable part models, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2010, pp. 2241–2248. .
[51] X. Li, D. Song, Y. Dong, Hierarchical feature fusion network for salient object detection, IEEE Transactions on Image Processing 29 (2020) 9165–9175.
[52] R.B. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2014, pp. 580–587.
[53] R.B. Girshick, Fast r-cnn, in: Proc. IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1440–1448. .
[54] S. Ren, K. He, R. Girshick, J. Sun, R.-C.N.N. Faster, Towards real-time object detection with region proposal networks, in: Proc. Advances in Neural Information Processing Systems NIPS, 2015, pp. 91–99.
[55] K. He, G. Gkioxari, P. Dollr, R.B. Girshick, Mask r-cnn, CoRR abs/1703.06870. .
[56] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Proc. European Conference Computer Vision, ECCV, vol. 9905, 2016, pp. 21–37. .
[57] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollr, Focal loss for dense object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence PP (99) (2017) 2999–3007. .
[58] H. Law, J. Deng, Cornernet: Detecting objects as paired keypoints, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Proc. European Conference Computer Vision, ECCV, vol. 11218, 2018, pp. 765–781. .
[59] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, Centernet: Keypoint triplets for object detection, in: Proc. IEEE/CVF International Conference on Computer Vision, 2019, pp. 6568–6577.
[60] Z. Dong, G. Li, Y. Liao, F. Wang, P. Ren, C. Qian, Centripetalnet: Pursuing high-quality keypoint pairs for object detection, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition CVPR, 2020, pp. 10516–10525.
[61] N. Jean, M. Burke, M. Xie, W.M. Davis, S. Ermon, Combining satellite imagery and machine learning to predict poverty, Science 353 (6301) (2016) 790–794.
[62] X. Tong, Q. Lu, G. Xia, L. Zhang, Large-scale land cover classification in gaofen-2 satellite imagery, in: Proc. IEEE International Geoscience and Remote Sensing Symposium IGARSS, 2018, pp. 3599–3602.
[63] W. Liu, Z. Wang, X. Liu, N. Zeng, D. Bell, A novel particle swarm optimization approach for patient clustering from emergency departments, IEEE Transactions on Evolutionary Computation 23 (4) (2019) 632–644.
[64] W. Liu, Z. Wang, Y. Yuan, N. Zeng, K. Hone, X. Liu, A novel sigmoid-function-based adaptive weighted particle swarm optimizer, IEEE Transactions on Cybernetics 51 (2) (2021) 1085–1093.
[65] N. Zeng, D. Song, H. Li, Y. You, F.E. Alsaadi, A competitive mechanism integrated multi-objective whale optimization algorithm with differential evolution, Neurocomputing 432 (12). .
[66] N. Zeng, Z. Wang, W. Liu, H. Zhang, K. Hone, X. Liu, A dynamic neighborhood-based switching particle swarm optimization algorithm, IEEE Transactions on Cybernetics (2020) 1–12.

Kang Liu received the bachelor degree from Xi'an Jiaotong University, Xi'an, China, in 2013, and the master degree from the University of Chinese Academy of Sciences, Beijing, China, in 2016. He is currently pursuing the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, and also as an Assistant Research Fellow with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. His research interests include Machine Learning, Computer Vision and Remote Sensing.

Ju Huang received the B.E. and M.E. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2013 and 2018, respectively. He is currently pursuing the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, and also as an assistant research fellow with the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences. His research interests include remote sensing, hyperspectral data processing, and machine learning.

Mingliang Xu received the B.E. and M.E. degrees from Zhengzhou University, China, in 2005 and 2008, respectively, and the Ph.D. degree from the State Key Laboratory of CAD & CG, Zhejiang University, China, in 2012, all in computer science. He is currently an Associate Professor with the School of Information Engineering, Zhengzhou University. His research interests include computer graphics and computer vision.

Matjaž Perc received his Ph.D. in 2007 from the University of Maribor, where he is now Professor of Physics and director of the Complexity Science Laboratory. He is a member of Academia Europaea and the European Academy of Sciences and Arts, and among top 1% most cited physicists according to Clarivate Analytics. He is also the 2015 recipient of the Young Scientist Award for Socio and Econophysics from the German Physical Society, and the 2017 USERN Laureate. In 2018 he received the Zois Award, which is the highest national research award in Slovenia. In 2019 he became Fellow of the American Physical Society. Matjaž is currently Editor of Physics Letters A and Chaos, Solitons & Fractals, and he is on the Editorial Board of New Journal of Physics, Proceedings of the Royal Society A, Journal of Complex Networks, EPL, European Physical Journal B, Scientific Reports, Royal Society Open Science, Applied Mathematics and Computation, and Frontiers in Physics.

Xuelong Li (Fellow, IEEE) is a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China.