Building Extraction from Remotely Sensed Images by Integrating Saliency Cue

Er Li, *Member, IEEE*, Shibiao Xu, *Member, IEEE*, Weiliang Meng, *Member, IEEE*, and Xiaopeng Zhang, *Member, IEEE*

***Abstract*—In this paper, we propose a novel two-step building extraction method from remote sensing images by integrating saliency cue. We first utilize classical features such as shadow, color, and shape to find out initial building candidates. A fully connected conditional random field model is introduced in this step to ensure that most of the buildings are incorporated. While it is hard to further remove the mislabled rooftops from the building candidates by only using classical features, we adopt saliency cue as a new feature to determine whether there is a rooftop in each segmentation patch obtained from previous step. The basic idea behind the use of saliency information is that rooftops are more likely to attract visual attention than surrounding objects. Based on a specifically designed saliency estimation algorithm for building object, we extract saliency cue in the local region of each building candidate, which is integrated into a probabilistic model to get the final building extraction result. We show that the saliency cue can provide an efficient probabilistic indication of the presence of rooftops, which helps to reduce false positives while without increasing false negatives at the same time. Experimental results on two benchmark datasets highlight the advantages of the integration of saliency cue and demonstrate that the proposed method outperforms the state-of-the-art methods.**

***Index Terms*—Buildings, fully connected conditional random field (CRF), saliency.**

I. Introduction

AUTOMATIC extraction of buildings from remote sensing images is key to a wide range of applications, including landscape analysis, three-dimensional urbanscene reconstruction, map updating, etc. To pursuit efficient, generic and accurate extraction result, many approaches have been proposed in the past few years [1]–[6]. Although important advances have been achieved, it is still a challenging task to guarantee high-quality building extraction result over various imagery. The main difficulty comes from the significant diversity of the appearance, density, and structure complexity of buildings across different regions.

To distinguish building rooftops from nonrooftop objects in aerial or satellite images, a variety of cues were exploited in previous work, such as shape, color, strong edges, corners, and shadows. Notice that a single cue is insufficient to adaptively identify rooftops under different circumstances, recent works focus on combining multiple cues to design more reliable descriptors of buildings [3], [5], [6]. Such methods first select the most probable rooftop candidates by integrating several cues and then a stochastic optimization process is followed to further refine the candidates. While these methods have demonstrated impressive results, the problem is that the optimal balance between the number of missing rooftops and the number of mislabeled rooftops is still hard to find. To obtain a low number of missing rooftops, one has to keep as much as possible rooftops during the candidate selection step. However, this is fragile since more undesirable nonrooftop objects may be incorporated at the same time, which will lead to a noticeable increase in the number of mislabeled rooftops even after the following refinement procedure.

In this paper, we propose a novel method to address the above problems by integrating the saliency cue during building extraction. Saliency, which is highly related to human visual perception, measures the importance and informativeness of one object in the scene [7]. It is observed that human vision processing system has a remarkable ability to automatically focus attention onto several interested regions which have high visual saliency within the field of view. By mimicking the mechanism involved in human vision system when selecting candidate salient objects in a scene, many models have been proposed to computationally extract salient image regions, which broadly benefits applications including image segmentation, object recognition [8], [9], and detection task [10], [11]. Motivated by this fact, we consider exploring the saliency cue for building extraction from remote sensing images, based on the assumption that rooftops have higher saliency than the other objects nearby in the local context. However, the challenge for remote sensing images is that multiple buildings with complex appearance are involved in one scene, which limits the straightforward application of existing saliency detection model. To address this challenge, we develop a two-step scheme with a novel use of saliency feature to obtain a low number of both missing rooftops and mislabeled rooftops in the final extracted result. We first follow the pipeline in [6] to get an initial segmentation of rooftops. Given an image with RGB information, the pipeline first decomposes the image into

Manuscript received May 5, 2016; revised July 12, 2016; accepted August 11, 2016. Date of publication September 13, 2016; date of current version February 13, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61331018, Grant 61620106003, Grant 91338202, Grant 61502490, and Grant 61671451, and in part by the National High-Tech Research and Development Program of China (863 Program) with No.2015AA016402. E. Li and S. Xu contributed equally to this work and share the first authorship. *(Corresponding author: Xiaopeng Zhang.)*

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: er.li@ia.ac.cn; shibiao.xu@ia.ac.cn; weiliang.meng@ia.ac.cn; xiaopeng.zhang@ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2016.2603184

perceptually homogeneous regions by using Gaussian mixture model (GMM) clustering model. Those segments belonging to shadows and vegetations are identified according to the intensity and greenness. Next, the remaining segments are classified into probable rooftops and probable nonrooftops depending on their shape, size, compactness, and surrounding shadow. Based on the classification of these segments, a higher order multilabel conditional random field (CRF) model is used to obtain the final extraction result. However, in contrast to the GMM clustering model used in the original pipeline, we introduce a full connected CRF model for the segmentation, which allows us to decrease the number of missing rooftops. Then in the second step, we adopt the saliency cue to determine whether there is a rooftop in each segmentation patch obtained from previous step. The calculated saliency map of each patch gives a good indication of the presence of rooftops. The patches without obvious salient objects will be assigned a low probability of containing rooftops, which helps to discard the mislabeled rooftops efficiently. We evaluate our method on publicly available benchmark datasets and experimental results show that the proposed algorithm outperforms the state-of-the-art building extraction methods.

The proposed approach expands upon the recent work in [6] and improves the original method by making several key contributions as follows:

1) We develop an original two-step scheme for building extraction from remote sensing images based on a novel use of visual saliency feature. Compared with the method in [6], a new step is proposed which integrates visual saliency cue for a further refinement of the building extraction result. To better describe the visual saliency of rooftops in remote sensing images, a specifically designed saliency estimation algorithm is also proposed.
2) We also introduce a fully connected CRF model for the segmentation of rooftops in the first step of our method, which brings an additional improvement in the global accuracy.

II. Related Work

There has been extensive work in the area of building extraction from remote sensing imagery over the years. In this section, we will give an overview of the most relevant work.

Much of the work in the field identifies building rooftops by using single cue, such as shapes [12], strong edges [1], corners [2], and shadows [13]. For example, Liu *et al.* [12] proposed model matching techniques based on node graph search to find the correct building rooftop shape. However, the method is restricted to building rooftops with separated and rectilinear structure. Saeedi and Zwick [14] detected line segments at several levels on the original image, and then generated the true rooftop hypotheses based on the extracted straight lines and the initial image segments. The method is sensitive to the quality of the extracted edge map. Sirmacek and Ünsalan [2] present a scale invariant feature transform keypoints based method for urban area extraction and building detection. Their method needs a priorly given building templates for the subgraph matching, which makes it difficult for images containing variety of buildings with complex shapes. Ok *et al.* [13] proposed a shadow-based framework to extract building rooftops from single optical very high resolution (VHR) satellite images. They first obtained the potential building regions based on shadows and light direction by applying a fuzzy landscape generation approach, and then utilized CRF optimization at pixel level to detect the final building regions. The method tends to generate incomplete rooftop extraction result if the shadows are broken due to noise and occlusion. Both Femiani *et al.* [15] and Ok [5] improved the method by running the CRF optimization globally and removing mislabeled rooftops with incorrect shadow information.

Recently, using combination of multiple cues has gained much attention since it is more robust for more complex scenes. Sirmacek and Ünsalan [16] combined shadow, edge information, and roof color in a two-step process. In their method, coarse rooftop candidates were selected first using shadow and color, then Canny edge map was utilized to verify the proposals and refine the rooftop contour. Cote and Saeedi [4] based their extraction method on shapes and corners. They first segmented the input image into smaller blobs through k-means clustering algorithm and selected the candidate rooftop blobs according to their shapes. Corners were then detected to refine the outline of rooftops through level set curve evolution method.

A major challenge of combining multiple cues is dealing with the heterogeneity among different features. To address the challenge, probabilistic model is exploited for its fault tolerance against noise and uncertainty in the extracted features. Benedek *et al.* [3] integrated several low-level features, including local gradient orientation density, roof homogeneity, roof colors, and shadows through a hierarchical framework. A multitemporal Marked Point Process (MPP) model combined with a bilayer Multiple Birth and Death stochastic optimization process was performed to flexibly fuse the heterogeneous features. More recently, Li *et al.* [6] proposed a novel higher order CRF-based method to achieve accurate rooftops extraction, which incorporated pixel-level feature and segment-level feature for the identification of rooftops.

In addition to the classical features, new features are also proposed for more accurate building extraction. Kovacs and Sziranyi [17] introduced a novel aerial building detection method based on region orientation as a new feature. The orientation feature was estimated from local gradient analysis and applied in various steps to integrate with multiple classical features throughout the whole framework. Later, Kovacs and Ok [18] developed a building detection approach by integrated urban area knowledge, which was obtained based on the feature points produced by the modified Harris for edges and corners method.

Saliency cue, as a more general feature with adaptivity to different circumstances, has also been introduced in [19] for the target detection in satellite images. Their method aimed to detect and classify variable target objects in high-resolution broad-area satellite images by using saliency and gist features, but without discussion on the extraction of target objects. Zhao *et al.* [20] focused on airport target recognition of aerial images based on saliency-constraint feature. Several recent works attempt to exploit the benefits of using saliency cue for building extraction from remote sensing images. Yang *et al.* [21] took advantage of visual saliency and Bayesian model to rapidly locate rooftop areas. They directly calculated saliency map in the whole image, however, the globally generated saliency

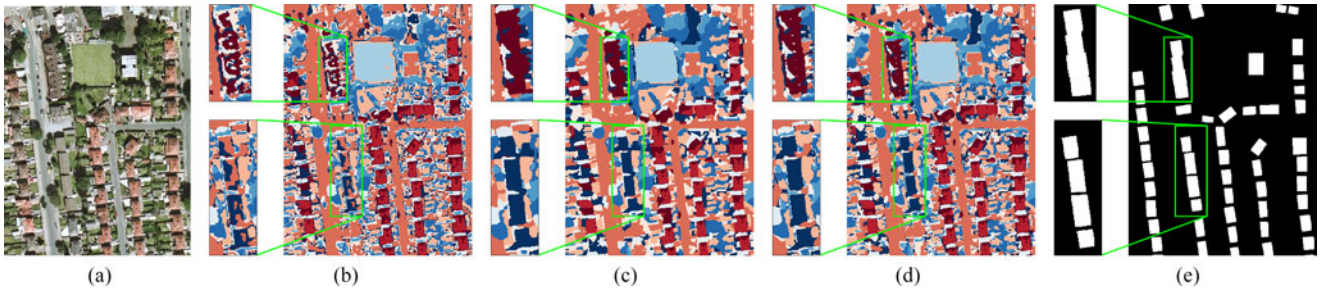


Fig. 1. Comparisons of the image presegmentation results using fully connected CRF model and GMM clustering method. (a) Original image. (b) Result of using GMM clustering method. (c) Result of using the four-neighbor grid CRF model. (d) Result of using the proposed fully connected CRF model. (e) The ground truth of rooftops. In (b)–(d), different colors represent different clusters. Notice the improvements of the rooftop segments contained in the green rectangles.

map is insufficient to reveal all of the potential rooftops with varying appearance. Cretu and Payeur [22] trained a support vector machine using descriptors derived from visual attention for the detection of buildings in aerial images. Their method had the limitation of requiring manually segmented masks for buildings and streets for the training. In our work, saliency cue is exploring in the local context of rooftops, which can provide more reliable measurement of the presence of rooftops. Furthermore, we integrate the saliency cue with classical features in an effective manner through probabilistic model and thus no user interaction is required.

III. Extraction Algorithm

The proposed method takes a remote sensing image with only RGB information as input. The whole process of our algorithm mainly consists of two steps: An initial segmentation of rooftops and a following refinement of the segmentation results by utilizing saliency cue. In the first step, we apply the same methodology as in [6] except that a new fully connected CRF model is adopt to overcome the loss of recall caused by inaccurate presegmentation in the original approach. In the second step, we compute the saliency map for each of the extracted rooftop and refine the segmentation result by integrating saliency cue into a probabilistic model. In the following, we will elaborate the details of each step.

A. Initial Segmentation

The goal of this step is to select potential rooftops from remote sensing image through an initial segmentation. We follow a similar pipeline as in [6] to obtain the initial building candidates. We observe that the method in [6] is able to maintain a relatively high precision on various images, however, the recall becomes unstable when it comes to images containing rooftops with large noise. The prime reason for the dramatic decreasing of recall in this case is the unreliable presegmentation generated by the GMM clustering. Since the GMM method merely clusters pixels in color space, the rooftops with corrupted color and low contrast are broken into small pieces, which were incorrectly identified as nonrooftops in the next step due to their irregular shapes. Such a failure case is shown in Fig. 1(b), even a smoothing operation is performed on the image before the GMM clustering, it is still hard to attain reliable segments of the rooftops contained in the green rectangles.

In order to reduce the presegmentation errors, we propose to employ the fully connected CRF model for the presegmentation. Different from traditional grid CRF model, a fully connected model establishes pairwise potential on every pair of pixels in the image. Thus, the Gibbs energy of a fully connected model takes the form:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \psi_i(x_i) + \sum_{\substack{i \in \mathcal{V}, j \in \mathcal{V} \\ i \neq j}} \psi_{ij}(x_i, x_j) \tag{1}$$

where $\mathcal{V}$ represents the set of all image pixels, ψ_i and ψ_{ij} denote the unary potential and pairwise potential, respectively, and x_i is the label taken by pixel i. Another difference is that the pairwise potential of our fully connected model is defined as the combination of spatial information and color contrast [23]. Let I_i, I_j and p_i, p_j denote the color and pixel coordinates of pixels i, j, respectively, then the pairwise potential is expressed with the following term:

$$\psi_{ij}(x_i, x_j) = \begin{cases} 0, & \text{if } x_i = x_j \\ \omega_1 \exp\left(-\frac{\|I_i - I_j\|^2}{2\theta_\beta} - \frac{\|p_i - p_j\|^2}{2\theta_\alpha}\right) & \\ \quad + \omega_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\lambda}\right), & \text{otherwise} \end{cases} \tag{2}$$

here ω_1 and ω_2 are the weight coefficients controlling the impacts of spatial term and color term, parameters $\theta_\alpha, \theta_\beta, \theta_\lambda$ denote the bandwidth of the Gaussian kernel function. $\theta_\alpha, \theta_\beta$ control the degrees of nearness and similarity, and θ_λ is used to characterize the effect of removing small isolated regions [23].

The proposed fully connected model improves the presegmentation quality in two aspects. First, the pairwise potential is defined over all pairs of pixels, which allows the model to capture long-range interactions; thus, the segmentation of objects associated with long-range context is augmented. Second, unlike the commonly used pairwise potential where only color contrast is considered, the proposed pairwise term incorporates both color contrast and spatial distance. Therefore, the proposed model is able to generate more accurate segmentation of objects with noise caused by sampling and proximity to other objects. Fig. 1(d) illustrates the benefits of using fully connected model.

Specifically, once the pixels of a given image are classified into ten classes using GMM clustering as did in [6], we initialize the unary potentials of the proposed fully connected CRF

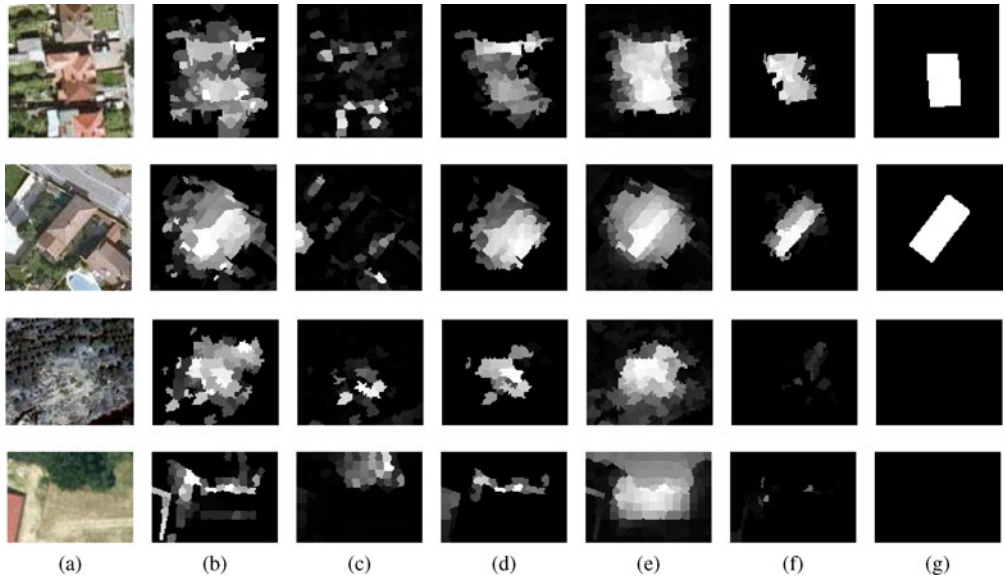


Fig. 2. Saliency map computed by different methods on the obtained image patch $\mathcal{R}_I$. (a) Image patches. (b) Results of GS method [26]. (c) Results of SF method [25]. (d) Results of wCtr method [31]. (e) Results of MR method [24]. (f) Results of our saliency estimation algorithm. (g) Ground truth. Brighter color indicates higher value.

model from the classification and then obtain the refined presegmentation through optimizing the CRF model. Based on the new presegmentation result, segments belonging to vegetation and shadows are first extracted using color features. Then, the remaining segments are classified into probable rooftops and probable nonrooftops depending on geometric shape feature and shadow information, i.e., we use area size, eccentricity, and compactness to identify irregular segments that are not likely to belong rooftops and further prune out the probable rooftops by checking the shadow information among its neighboring segments. Finally, we perform a higher order multilabel CRF segmentation based on the initial classification of the pixels through above steps. The proposed new presegmentation result is able to produce more accurate segmentation for the rooftops, which effectively reduces the number of rooftops that are misclassified as probable nonrooftops, thus achieving high recall in the initial segmentation of rooftops.

B. Saliency-Based Refinement

1) Saliency Detection for Rooftop: The goal of this step is to further refine the segmented rooftops generated from above step by removing the mislabeled rooftops. This is not easy since we have already integrated multiple cues to identify rooftops in the first step, such as geometric shape, shadows, and color. Thus, the remaining mislabeled rooftops often share several features with actual rooftops. In order to single out the mislabeled rooftops, different cues should be taken into account. Inspired by recent work on saliency detection, we propose to exploit saliency information for further verification of rooftops. The basic idea behind the use of saliency information is that rooftops are more likely to attract visual attention than surrounding objects. Therefore, high saliency value should be detected in the local area where the rooftop is located. If no salient part is detected, then the probability of the area containing a rooftop is assumed to be low. Since remote sensing images are not taken to frame individual object, we calculate the saliency map in the vicinity of each rooftop segment resulting from the initial segmentation rather than working on the whole image.

For each rooftop candidate $\mathcal{R}$ obtained from the initial segmentation, we calculate its ROI $\mathcal{R}_I$ as the region within the bounding box of $\mathcal{R}$. To compensate for the blurry of the boundary of the buildings, we enlarge the bounding box by 100% so that the whole rooftop is ensured to be contained in the image patch $\mathcal{R}_I$. We borrow the framework from [24] to estimate the saliency value of each pixel in $\mathcal{R}_I$. We have investigated several modern saliency detection methods [24]–[28], while all of these methods are able to give reasonable result in some image patches, their results tend to be affected by intensity artifacts caused by shadows or illumination ambiguity. Moreover, the methods are hard to handle image patches without noticeable foreground objects (see Fig. 2). We build our saliency estimation algorithm based on the two-stage framework in [24], since it provides us a flexible and efficient way to involve additional prior knowledge about the characteristic of rooftops during the saliency estimation. In our implementation, we use superpixels instead of pixels for saliency detection. This not only accelerates the calculation but also achieves a smoother saliency detection result, as demonstrated in many studies [29]. Specifically, we adopt the SLIC algorithm [30] to generate superpixels for $\mathcal{R}_I$. Then, we represent image $\mathcal{R}_I$ as a graph $G = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ denotes N generated superpixels and E is the set of undirected edges connecting adjacent superpixels in V.

We first generate an initial saliency map through analyzing the commonly used prior knowledge for saliency detection: contrast prior and center prior, which are supported by psychological evidence [7] [29]. As we expect the extracted salient region to coincide with probable rooftops contained in $\mathcal{R}_I$, three saliency

Fig. 3. Pipeline of estimating saliency cue. (a) Original image. (b) Boundary connectivity measure in (4). (c) Region contrast measure in (3). (d) Background constraints measure in (6). (e) Initial saliency map from (7). (f) Final saliency map from (10). Brighter color indicates higher value.

measures based on the prior knowledge are proposed for saliency estimation:

1) *Region Contrast:* Image region with high contrast surroundings generally attracts more visual attention [29]. We measure the region contrast of v_i by summing its weighted appearance distance to all other superpixels, with the following term:

$$RC(v_i) = \sum_{j=1}^{N} \exp(-D_s^2(v_i, v_j)/\sigma_c^2) D_c(v_i, v_j) \quad (3)$$

here $D_s(v_i, v_j)$ is the Euclidean distance between the centers of superpixels v_i and v_j, $D_c(v_i, v_j)$ is the Euclidean distance between their average colors in Lab color space, and σ_c controls the strength of spatial weighting.

2) *Boundary Connectivity:* As suggested by attention theory about visual saliency [7], the location of salient object is more likely to be close to the center of image, which means image boundary are mostly background [31]. This is not true for remote sensing images, since multiple buildings can be found in different locations over a whole region. However, in our case the image is divided into several regions based on the initial segmentation result of buildings; therefore, we can assume that the expected salient part is located in the center area of each region. We quantify this cue by calculating the connectivity between superpixel v_i and image boundary:

$$BC(v_i) = \frac{Len_b(v_i)}{\sqrt{Len(v_i)}} \quad (4)$$

where the length of region v_i's perimeter $Len(v_i)$ is defined by summing the weighted geodesic distance from v_i to each superpixel in the image

$$Len(v_i) = \sum_{j=1}^{N} \exp(-G^2(v_i, v_j)/\sigma_s^2) \quad (5)$$

where geodesic distance $G(v_i, v_j)$ is defined as the accumulated edge weights $D_c(v_k, v_{k+1})$ along the shortest path $(v_i, \ldots, v_k, v_{k+1}, \ldots, v_j)$ from v_i to v_j, and σ_s controls the distance sensitivity of the boundary connectivity. Similarly, $Len_b(v_i)$ represents the length of region v_i's perimeter on boundary, which only sums the distances from v_i to the superpixels on the image boundary.

3) *Background Constraints:* Isolated shadows and vegetations in the image may cause misleading saliency detection result. Therefore, we introduce additional background constraints for shadows and vegetations extracted in the first step:

$$BS(v_i) = \delta(v_i) \quad (6)$$

here $\delta(v_i)$ is 1 if v_i belongs to shadows or vegetations and 0 otherwise.

Fusing these saliency measures together, we initialize the saliency map for $\mathcal{R}_I$ as

$$S_1(v_i) = (1 - BC(v_i)) \cdot RC(v_i) \cdot (1 - BS(v_i)). \quad (7)$$

We then compute the final saliency map through manifold ranking on a graph [24]. The initial saliency map is binary segmented with an adaptive threshold set as the mean saliency over the entire saliency map, and the superpixels with saliency value above the threshold are viewed as the labeled queries in a ranking problem. Then, the saliency value of each element in V is expressed as its ranking score according to its relevance to the queries. Let the saliency values of superpixels in set V be $\mathbf{f} = \{f_i\}_{i=1}^N$, we thus find the optimal saliency value by solving the following optimization problem:

$$\bar{\mathbf{f}} = \arg\min_f \left(\sum_{(i,j)\in E} \omega_{ij} \left\| \frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}} \right\|^2 + \sum_{i=1}^{N} \lambda_i \|f_i - y_i\|^2 \right) \quad (8)$$

here $\mathbf{y} = \{y_i\}_{i=1}^N$ is an indication vector, where $y_i = 1$ if $S_1(v_i)$ is greater than the threshold and $y_i = 0$ otherwise, λ_i controls the impact of labeled query v_i, ω_{ij} is the weight defined on edge connecting superpixel v_i and v_j:

$$\omega_{ij} = \exp\left(-\frac{\|I_i - I_j\|^2}{2\sigma^2} \right) \quad (9)$$

the strength of edge weight is controlled by constant σ, and $d_{ii} = \sum_j \omega_{ij}$. Considering that the ranking function is quadratic, we can optimize it efficiently by least-square method. Consequently, resulted optimal saliency value $\bar{\mathbf{f}}$ can be written as

$$\bar{\mathbf{f}} = (\mathbf{D} - \Lambda\mathbf{W})^{-1}\mathbf{y} \quad (10)$$

where $\Lambda = \text{diag}\{\frac{1}{1+\lambda_1}, \ldots, \frac{1}{1+\lambda_n}\}$, $\mathbf{D} = \text{diag}\{d_{11}, \ldots, d_{nn}\}$ and $\mathbf{W}$ represents the weight matrix $[\omega_{ij}]_{n\times n}$. Notice that here we deduce $\bar{\mathbf{f}}$ using unnormalized Laplacian matrix to achieve better performance as revealed in [24]. Once getting the final saliency map for superpixels, we assign the saliency value of each pixel as that of the superpixel it belongs to. Fig. 3 illustrates the results of each individual step during the saliency map estimation.

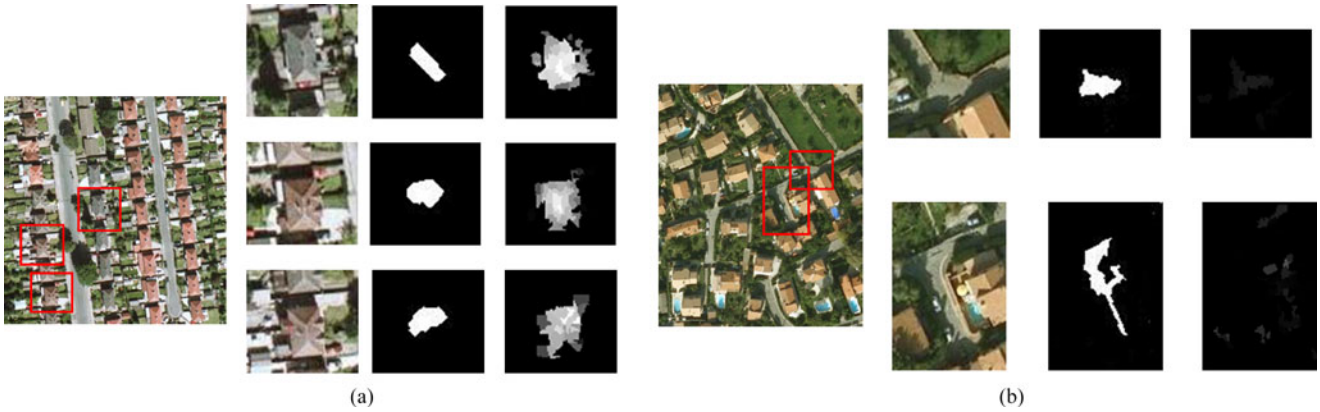


Fig. 4. Benefits of introducing saliency cue. The first four columns (a) shows how saliency cue helps to recover the missing parts of rooftops. The last four columns (b) shows how saliency cue helps to suppress mislabeled rooftops. For each example, from left to right we show the original image, close-up view of red rectangle area, initial rooftop extraction result from the first step in the red rectangle area, estimated saliency maps of rectangle area.

The estimated saliency map provides an insightful description of the existence of rooftops. If there is rooftop within ROI $\mathcal{R}_I$, it has the ability to recover the missing parts of rooftops that the initial segmentation fails to detect. As illustrated in Fig. 4(a), the initial segmentation produces incomplete extraction results for the gabled rooftops, however, the saliency map creates much better representation of the whole rooftop. More importantly, if there is no rooftop within ROI $\mathcal{R}_I$ actually, saliency map indicates a low probability of the existence of rooftops in this patch. As shown in Fig. 4(b), part of road is mislabeled as rooftops due to the adjacent shadows. While this error is hard to avoid in previous method, the estimated saliency map clearly suggests that the patch is unlikely to contain rooftops.

2) Refinement: The obtained saliency map allows us to further refine the segmented rooftops from the first step. We optimize the final segmentation results by integrating saliency cue into a fully connected CRF model as discussed in Section III-A. Thus, the unary potential in (1) is divided into two components, appearance potential and saliency potential:

$$\sum_{i\in\mathcal{V}}\psi_i(x_i)=\sum_{i\in\mathcal{V}}\psi_i^A(x_i)+\sum_{i\in\mathcal{V}}\psi_i^S(x_i) \tag{11}$$

here $\psi_i^A(x_i)$ is the appearance potential defined as the negative log of the likelihood of label x_i being assigned to pixel i, and can be deduced from the initial segmentation results of Section III-A as did in [6]. For the saliency potential $\psi_i^S(x_i)$, we accumulate the saliency map of each candidate rooftop to form the final saliency map $\bar{\mathbf{f}}$ of the whole image, then define $\psi_i^S(x_i)$ as

$$\psi_i^S(x_i)=-\exp(\theta_s\cdot\bar{\mathbf{f}}_i) \tag{12}$$

here we use an exponential function to emphasize the saliency cue and θ_s denotes the scaling factor for the exponential, which we empirically set to 1.5. We finally optimize the fully connected CRF model over the whole image through the mean field approximation algorithm described in [32].

IV. Evaluation

In this section, we evaluate our approach on various challenging real world datasets. We first perform quantitative evaluation on the publicly available benchmark datasets provided in [3]. We also illustrate the improvement of the proposed method compared to several state-of-the-art methods on benchmark datasets provided in [18]. Parameter settings and limitations of the proposed methods will be discussed last.

A. Evaluation Metrics

We evaluate our results on pixel level by computing the widely used precision (P), recall (R), and F-score (F_1) measures [33], which are defined as

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

$$R=\frac{TP}{TP+FN}, \text{and} \tag{14}$$

$$F_1=\frac{2PR}{P+R}. \tag{15}$$

Here, TP represents true positives and corresponds to the number of pixels correctly labeled as rooftop in both ground truth and segmentation result. FP represents false positives and corresponds to the number of pixels mislabeled as rooftop. FN represents false negatives and corresponds to the number of pixels mislabeled as nonrooftop. F_1 measures the overall performance through the weighted harmonic of precision and recall. Additionally, we also evaluate the object-level performance by counting the missing and falsely labeled rooftops (MO and FO, respectively), and then calculate the F-score at object level using the same formula. We utilize an overlapping threshold of 60% to determine the number of MO and FO as described in [13].

B. Evaluation on the SZTAKI-INRIA Benchmark

The SZTAKI-INRIA dataset [3] consists of nine aerial and satellite images from different geographical regions and contains 665 buildings with significantly different building appearance. The manually annotated ground truth data are also provided for validation. Detailed properties of the dataset can be found in [17].

TABLE I
NUMERICAL OBJECT-LEVEL AND PIXEL-LEVEL COMPARISON BETWEEN STATE-OF-THE-ART BUILDING DETECTION METHODS AND THE PROPOSED METHOD (SCRF) WITH BEST RESULTS IN BOLD, ON THE SZTAKI-INRIA BENCHMARK DATASET

Dataset		Object-level performance												Pixel-level performance									
		EV		SM		MPP		OSBD		HCRF		Prop. SCRF		EV		SM		MPP		HCRF		Prop. SCRF	
Name	#obj	MO	FO	MO	FO	MO	FO	MO	FO	MO	FO	MO	FO	P	R	P	R	P	R	P	R	P	R
BUDAPEST	41	11	5	9	1	2	4	3	1	1	0	**0**	**0**	0.73	0.46	0.84	0.61	0.82	0.71	**0.90**	0.75	0.84	**0.81**
SZADA	57	10	18	11	5	4	1	4	**0**	2	3	**2**	1	0.61	0.62	0.79	0.71	0.93	0.75	**0.85**	0.86	0.84	**0.90**
COTE D'AZUR	123	14	20	20	25	5	4	4	6	3	4	**3**	**2**	0.73	0.51	0.75	0.61	0.83	0.69	**0.75**	0.81	0.74	**0.84**
BODENSEE	80	11	13	18	15	7	6	8	7	4	3	**3**	**3**	0.56	0.30	0.59	0.41	0.73	0.51	**0.84**	0.76	0.80	**0.79**
NORMANDY	152	18	32	30	58	18	1	**4**	10	16	7	7	**6**	0.60	0.32	0.62	0.55	0.78	0.60	**0.79**	0.67	0.76	**0.76**
MANCHESTER	171	46	17	53	42	19	6	NA	NA	20	**3**	**10**	4	0.64	0.38	0.60	0.56	0.86	0.63	**0.82**	0.67	0.79	**0.80**
Overall F-score		0.827		0.771		0.936		0.948		0.948		**0.967**		0.517		0.631		0.726		0.786		**0.805**	

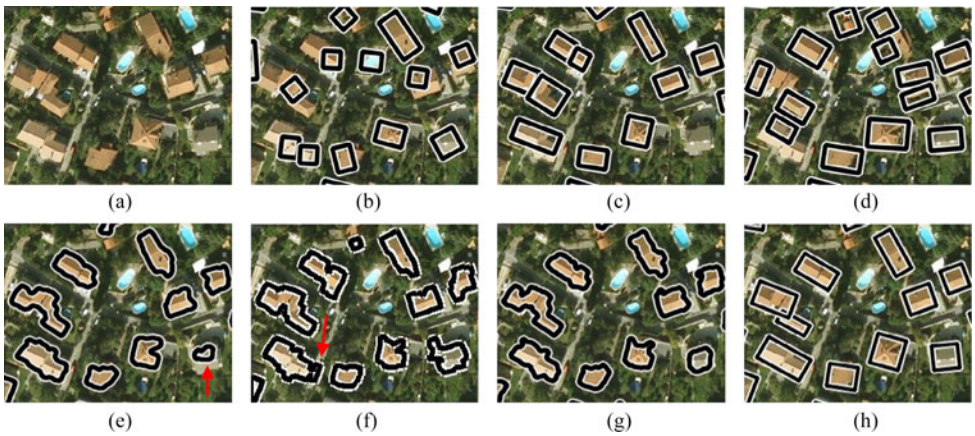


Fig. 5. Comparisons between the proposed method and the state-of-art methods. The result of EV, SM, MPP, and OSBD methods are from [17]. The red arrows in (e) and (f) indicate the incorrect segmentation results of the HCRF method and the OSBD method, respectively. (a) Original image. (b) EV [16]. (c) SM [34]. (d) MPP [3]. (e) HCRF [6]. (f) OSBD [17]. (g) Proposed SCRF. (h) Ground Truth.

Table I lists the numerical object-level and pixel-level comparisons of the proposed saliency-based CRF method (SCRF) against five state-of-the-art approaches: EV [16], SM [34], MPP [3], OSBD [17], and HCRF [6]. It is shown that the proposed method achieves overall improvement over the best of state-of-art method (HCRF) by 2% at pixel level and 2% at object level.

Fig. 5 offers qualitative comparison results on a selected site from COTE D'AZUR dataset. EV [16] method is able to localize most of the rooftops through extracted shadow information and estimated illumination direction. However, the method gives poor rooftop outlines as it strongly relies on detecting accurate and complete edge feature of rooftops, which is also an error prone task. Another limitation of EV method is that it can only handle buildings with rectangular shape. SM [34] method fails to capture the dark gray buildings in the right bottom corner of Fig. 5(a), since it is sensitive to low contrast with surroundings. The method also lacks the ability to detect inhomogeneous objects. MPP [3] method adopts probabilistic model to fuse multiple cues, which improves both MO and FO. However, the method is also restricted to rectangular buildings; thus, erroneously cutting across the buildings with complex shapes, as shown in the left part of Fig. 5(d).

While both OSBD [17] method and HCRF [6] method are able to produce high quality result at object level, notable artifacts still persist at the boundaries of several extracted buildings, which reduce the performance at pixel level. The OSBD method tends to merge two blobs close to each other during the shape refinement step and thus wrongly labels part of road as buildings, as indicated by the red arrow in Fig. 5(f). The HCRF method avoids introducing false positives in this area by discarding small blobs when selecting the candidate rooftops, but has the downside that some parts of rooftop are also discarded if the presegmentation fails to represent the shape of the rooftop. Such a failure case is shown in Fig. 5(g), the HCRF method only captures part of the building as indicated by the red arrow, due to the incorrect presegmentation result on this building.

Compared with state-of-the-art methods, the proposed method achieves better pixel-level accuracy and maintains comparable performance at object level on the given sample image. Compared with the OSBD method, our method inherently reduces false positives at pixel level since it is based on HCRF method. And compared with the original HCRF method, our SCRF method improves the recall significantly at the cost of a slightly decrease on precision. Our method improves the recall from two aspects: first, the improved presegmentation approach

TABLE II
NUMERICAL OBJECT LEVEL AND PIXEL LEVEL OF HCRF METHOD, APPLYING OUR APPROACH WITHOUT USING SALIENCY CUE (PROP.SCRF W/O SALIENCY), AND THE PROPOSED METHOD (SCRF) WITH BEST RESULTS IN BOLD, ON THE SZTAKI-INRIA BENCHMARK DATASET

Dataset		Object-level performance						Pixel-level performance								
		HCRF		Prop.SCRF w/o Saliency		Prop.SCRF		HCRF			Prop.SCRF w/o Saliency			Prop.SCRF		
Name	#obj	MO	FO	MO	FO	MO	FO	P	R	F_1	P	R	F_1	P	R	F_1
BUDAPEST	41	1	0	0	1	0	0	0.90	0.75	0.81	0.82	0.81	0.81	0.84	0.81	**0.82**
SZADA	57	2	3	2	3	2	1	0.85	0.86	0.85	0.83	0.90	0.86	0.84	0.90	**0.87**
COTE D'AZUR	123	3	4	3	5	3	2	0.75	0.81	0.77	0.72	0.85	0.78	0.74	0.84	**0.79**
BODENSEE	80	4	3	3	4	3	3	0.84	0.76	0.79	0.79	0.80	0.79	0.80	0.79	**0.79**
NORMANDY	152	16	7	6	10	7	6	0.79	0.67	0.72	0.70	0.78	0.74	0.76	0.76	**0.76**
MANCHESTER	171	20	3	9	5	10	4	0.82	0.67	0.73	0.76	0.81	0.78	0.79	0.80	**0.79**
Overall F-score		0.948		0.959		**0.967**			0.786			0.796			**0.805**	

Fig. 6. Comparison with the state-of-the-art method (HCRF) on MANCHESTER dataset. (a) Original image. (b) Result of HCRF. (c) Result of the proposed SCRF method. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

helps us to reduce the number of false negative rooftops; second, the integration of saliency cue further recovers the missing part of a whole rooftop. More importantly, the saliency cue successfully prevents the increase of false positive rooftops at the same time. We run our method without using saliency cue and give the numerical result in Table II, showing how the two steps affect the overall performance.

To better demonstrate the advantages of our method, a more challenging example is given in Fig. 6. The image is taken from MANCHESTER dataset and contains a lot of rooftops with low contrast, image noise, and self-shadows, which makes rooftop extraction more difficult for the state-of-the-art methods. Table I shows weak performances of state-of-the-art methods on this dataset. Our method outperforms all of the compared methods and improves the object-level accuracy and pixel-level accuracy by 3% and 6% over the state-of-the-art performance. Qualitative results in Fig. 6 confirm the benefits of the improved presegmentation and the integration of saliency cue. More results of our method are given in Fig. 8.

C. Evaluation on the VHR Benchmark

We also evaluate the pixel-level performance of our method on the VHR Benchmark provided in [18]. The VHR benchmark dataset contains 14 image patches acquired from two different satellites, IKONOS-2 (1m), and QuickBird (60 cm). All imagery includes four multispectral bands (B, G, R, and NIR) with a radiometric resolution of 11 bits (16 bits images) per band. Ground truth data and the result of four state-of-the-art approaches including Grabcut [13], Multi-label Partitioning (MLP) [5], SSDF [35], and Urban Area Knowledge Integration-based method (UAKI) [18] are also provided along with the dataset.

The quantitative comparison of our method with state-of-the-art approaches is given in Table III. As shown in the table, the

Fig. 7. Comparison with state-of-the-art methods on VHR benchmark [18]. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red. (a) Original image. (b) Grabcut [13]. (c) MLP [5]. (d) SSDF [35]. (e) UAKI [18]. (f) Prop. SCRF. (g) Original image. (h) Grabcut [13]. (i) MLP [5]. (j) SSDF [35]. (k) UAKI. [18]. (l) Prop. SCRF.

proposed method is able to generate much better results than all of the other state-of-the-art approaches on 10 of the 14 image patches, and shows overall improvement over the UAKI method on all performance measures, with significant improvements for several images. While all of the methods mentioned in Table III adopt shadows as an important cue for the identification of rooftops, consequently, they also suffer from the noisy, incomplete, and ambiguous observation of shadows in the images. Redundant shadow information, which usually comes from mislabeled shadows caused by dark regions and shadows cast by walls and fences, tends to introduce large false positive areas. On the other hand, lost shadow information due to noise and occlusion, often results in large false negative areas. To overcome these limitations, MLP [5] utilizes a global multilabel graph partitioning strategy to recover the missing rooftops as result of lost shadow information; thus, the method can obtain much higher recall. SSDF [35] and UAKI [18] further improve the performance, however, they are still susceptible to unreliable shadows. Two examples are given in Fig. 7, notice the missed rooftops without distinct shadows in the first example and the mislabeled rooftops due to misleading shadows in the second example. Unlike the above methods, our method can detect

Fig. 8. More example results of the proposed method on the two benchmark datasets. The first three images are taken from Budapest, Normandy and Cot dAzur datasets in SZTAKI-INRIA benchmark, respectively. The last three ones are taken from image patch 3, 6, and 14 in VHR benchmark, respectively. Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

rooftops even when there is no visible shadow information with the help of our improved presegmentation. Furthermore, integration of saliency cue allows us to effectively alleviate the precision loss incurred by the misleading shadows.

Finally, as pointed in [18], the VHR benchmark dataset is purposefully selected to uncover the potential of the UAKI method. This explains the worse performance of our method on four images than the UAKI method. However, even in this case we still achieve an overall performance improvement, which further justifies the efficiency and robustness of the proposed saliency cue-based method.

D. Parameter Settings

The proposed SCRF method mainly involves the following parameters: Parameters for the fully connected CRF model and parameters for saliency estimation.

1) *Fully Connected CRF Model Parameters:* Fully connected CRF model is used twice in the proposed method, for the presegmentation in Section III-A and the final rooftop refinement in Section III-B2. We employ the same parameter settings for both of the two steps, since there are no significant differences between the two energy functionals. For the parameters used to calculate the pairwise potential as

TABLE III
PIXEL-LEVEL QUANTITATIVE RESULTS FOR GRABCUT [13], MLP [5], SSDF [35], UAKI [18] AND THE PROPOSED METHOD WITH BEST RESULTS IN BOLD, ON THE VHR BENCHMARK DATASET

Dataset	Pixel-level performance														
	GrabCut			MLP			SSDF			UAKI			Prop.SCRF		
Image Patch	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
#1	59.1	58.6	58.8	36.5	56.8	44.4	60.9	69.5	64.9	81.2	75.0	78.1	**83.8**	**84.9**	**84.3**
#2	70.8	49.8	58.5	76.8	78.9	77.8	79.8	76.1	77.9	74.3	86.4	79.9	**82.2**	**89.6**	**85.7**
#3	60.4	76.3	67.4	60.1	**90.2**	72.1	59.5	69.4	64.1	69.2	89.0	77.9	**84.3**	81.6	**82.9**
#4	54.6	64.8	59.3	52.4	76.7	62.3	63.5	54.4	58.6	**86.6**	78.8	**82.5**	79.5	**83.2**	81.3
#5	71.5	61.7	66.2	70.2	**89.5**	78.7	88.8	78.3	83.2	91.0	88.1	89.6	**94.2**	88.8	**91.4**
#6	46.3	80.0	58.7	23.8	74.4	36.1	67.1	83.8	74.5	**87.4**	68.2	76.7	82.0	**85.9**	**83.9**
#7	77.5	83.2	80.3	77.2	87.3	81.9	80.1	82.3	81.2	81.7	**88.8**	**85.1**	**83.9**	88.2	**86.0**
#8	72.2	69.4	70.8	68.1	**86.9**	76.4	83.5	70.0	76.2	**86.4**	83.6	85.0	83.9	86.4	**85.1**
#9	47.4	62.3	53.9	40.6	72.6	52.6	78.5	88.5	83.2	89.9	90.2	90.0	**94.7**	**90.6**	**92.6**
#10	30.6	71.5	42.8	20.0	71.4	31.3	72.2	75.0	73.5	61.0	73.0	66.4	**74.3**	**87.6**	**80.4**
#11	70.1	92.2	79.6	77.9	**95.9**	86.0	75.8	94.6	84.2	**83.7**	87.0	85.3	83.6	93.0	**88.0**
#12	46.5	17.3	25.2	41.1	32.2	36.1	37.0	7.74	12.8	**84.4**	**81.0**	**82.7**	81.8	79.9	80.8
#13	62.6	52.3	57.0	67.6	86.0	75.7	77.1	69.6	73.1	**86.2**	**85.1**	**85.6**	80.7	83.6	82.1
#14	61.1	43.1	50.5	66.6	71.3	68.8	73.2	67.4	70.2	84.3	**85.9**	**85.1**	**85.1**	79.5	82.2
Average	57.5	61.9	59.6	53.1	78.1	63.2	75.5	71.7	73.5	83.5	84.4	83.9	**84.0**	**86.4**	**85.2**

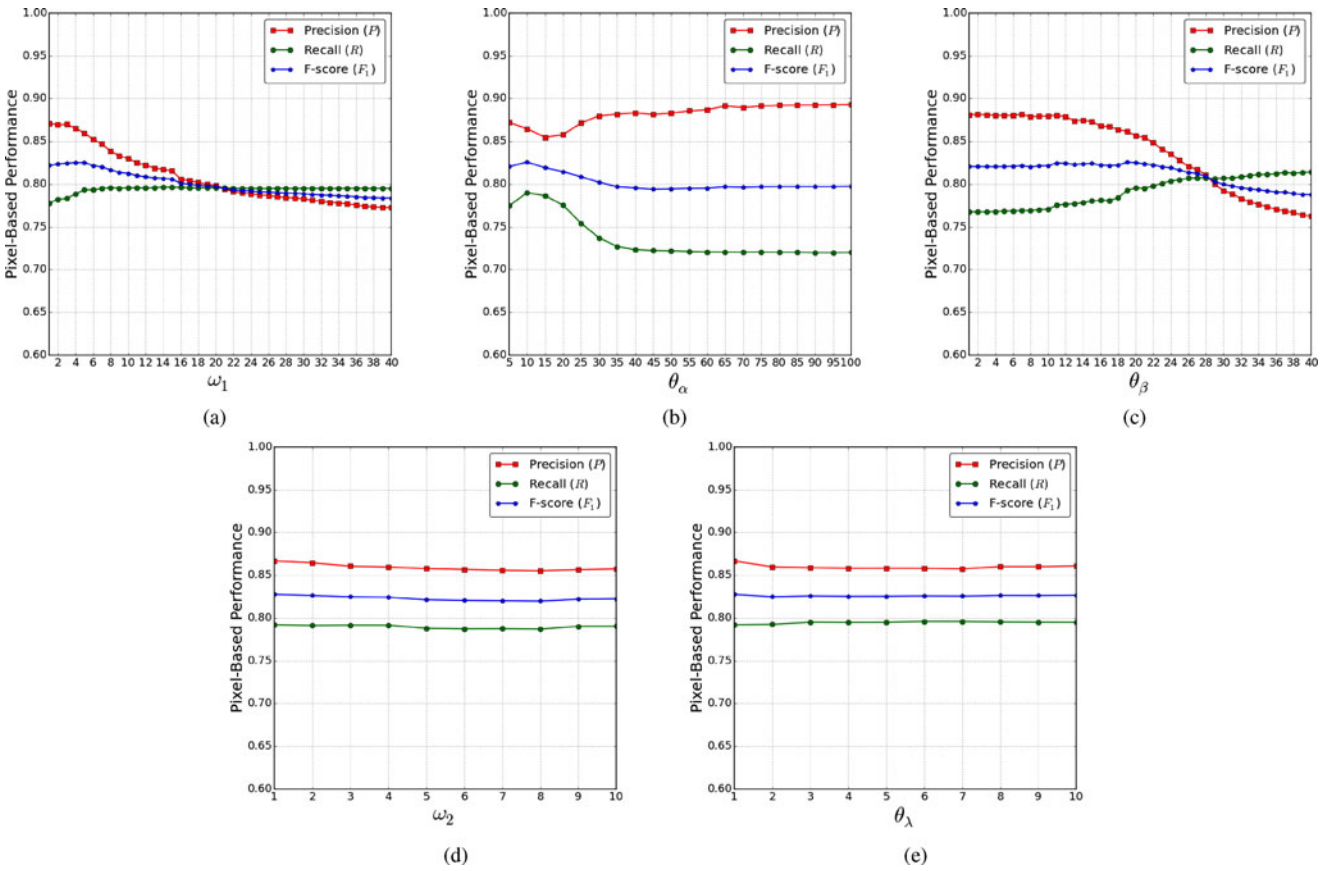


Fig. 9. Variation of accuracy with different parameter settings in (2). In each plot, nonvarying parameters are fixed as their optimal settings.

defined in (2), we initialize the parameters following the guidelines in [23] and then vary the parameters to search the optimal settings on our own dataset. The experimental results reveal that parameters ω_2 and θ_λ have relatively little impact on the accuracy of the final result, as shown in Fig. 9(d) and (e), which was also indicated in [23] . Thus, we set $\omega_2 = \theta_\lambda = 1.0$, the same as suggested in [23]. For parameters θ_α , θ_β, and ω_1, we test different parameter settings on the two benchmark dataset, and the effects of parameter variation are shown in Fig. 9(a)–(c). ω_1 weighs the

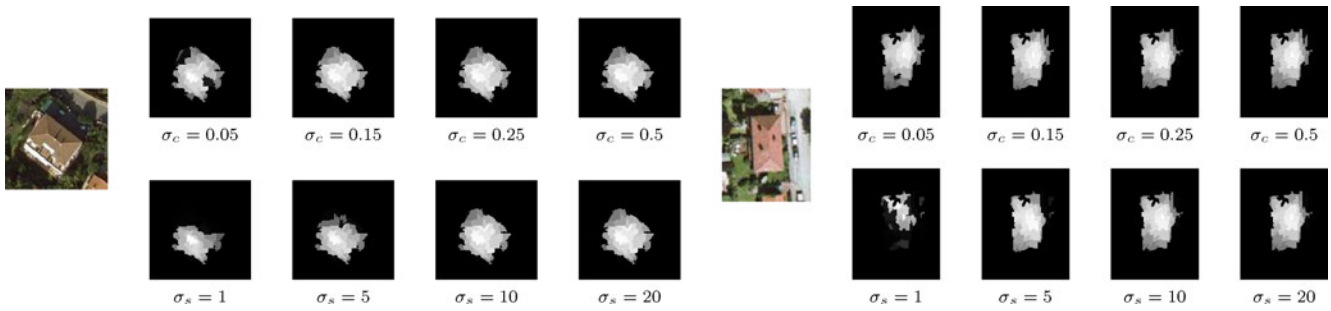


Fig. 10. Effects of choosing different σ_c and σ_s for saliency estimation. Nonvarying parameters are fixed as their optimal settings $\sigma_c = 0.25$, $\sigma_s = 10$. (a) $\sigma_c = 0.05$. (b) $\sigma_c = 0.15$. (c) $\sigma_c = 0.25$. (d) $\sigma_c = 0.5$. (e) $\sigma_c = 1$. (f) $\sigma_c = 5$. (g) $\sigma_c = 10$. (h) $\sigma_c = 20$.

Fig. 11. Failure cases. (a) Rooftops with nonuniform color appearance and similar color with background; (b) The proposed method fails to capture the whole rooftops. Ships in (c) are mislabeled as rooftops by the proposed method (d). Correct results (TP) are shown in green, false positives are shown in blue, and false negatives are shown in red.

impact of spatial term and color term. Large values of ω_1 leads to oversmooth segmentation result, which slightly improves the recall, but at the cost of significant drop in precision, as shown in Fig. 9(a). θ_α controls the spatial range of pairwise interaction. Accuracy increases as θ_α grows from 1 to 5, since the spatial smoothness helps to remove pixel-level noise in the local range of rooftop. However, as θ_α keeps growing, longer range interaction is considered. Therefore, rooftops that have similar color with background would be smoothed into the background, since they are now viewed as the noise in the background. As a result, recall decreases substantially, as shown in Fig. 9(b). θ_β modulates the effects of color contrast. There is little change in accuracy when θ_β gets low values. But when θ_β is too high, neighboring rooftops with similar colors would be merged and therefore causes more false positives and decreases the overall accuracy, as shown in Fig. 9(c). Through grid search in a reasonable range, we find the optimal setting as $\theta_\alpha = 10.0, \theta_\beta = 19.0$, and $\omega_1 = 5.0$ based on the experimental result.

2) *Saliency Estimation Parameters:* For the edge weight σ in (9), we apply the same value as θ_β since they share similar characteristic. For the other two weighting parameters σ_c and σ_s in (3) and (5), we experimentally set to 0.25 and 10 in our implementation. Results of testing a few different parameters of σ_c and σ_s on two sample images are shown in Fig. 10 and for the balance weight λ_i, different with the constant setting for all pixels as done in [24], we value the weight based on the background constraints $BS(v_i)$ defined in (6) to suppress the influence of shadows and vegetations. By experimentation, we set $\lambda_i = 10$ if $BS(v_i) > 0$ and $\lambda_i = 0.1$ otherwise. Regarding the parameter for superpixel generation, we specify the pixel number within each superpixel as 40 instead of fixing the number of superpixels, which gives better performance.

Other parameters, which mainly inherit from [6], are kept the same except for the eccentricity threshold and compactness threshold. We slightly decrease these two value to 0.10 so as to incorporate the rooftops with extremely irregular shape in the VHR benchmark dataset.

E. Limitations

First, our method assumes center prior and contrast prior for the estimation of saliency map, which are fairly common assumptions also made in other works. However, in dense area full of buildings, these two assumptions may be violated and thus the proposed region contrast and boundary connectivity saliency measures are insufficient to reflect the visual appearance of the rooftops, which leads to performance degradation. One typical failure case is shown in Fig. 11(a) and (b). In this case, the center area of rooftop is background, violating our center assumption. Furthermore, the region contrast measure fails to capture some rooftops due to their nonuniform color appearance and similar color with background area. To remedy this problem, more intelligent saliency estimation method will be explored in the future.

Second, even if the proposed saliency cue helps us to exclude most of the mislabeled rooftops, a few false positives still persist. Such as the ships in Fig. 11(c), which have regular shape and highly salient appearance similar to rooftops. Therefore, the saliency cue is insufficient to remove the erroneously labeled

rooftops in this case, as shown in Fig. 11(d). Combination using of additional features [17], [18] might address these issues.

V. Conclusion and Future Work

We propose a novel two-step method for improving building extraction from remote sensing images. An improved presegmentation method based on fully connected CRF model is first used to reduce the false negatives, and then the method leverages saliency cue in the second step to further reduce the false positives. A specially designed saliency estimation algorithm is also introduced to make it suitable for detecting rooftops. Compared with several state-of-the-art methods on two benchmark datasets, the results show that our method achieves improved performance and that it generalizes well across varying image conditions. In future work, we plan to investigate more intelligent saliency estimation method to further improve the accuracy. In addition, using combination of other available cues is also a promising avenue for future research.

Acknowledgment

The authors would like to thank Dr. A. Manno-Kovacs for providing us with numerical results on SZTAKI-INRIA benchmark dataset to compare against the proposed method. We thank Dr. C. Benedek for providing the SZTAKI-INRIA[1] benchmark dataset. We thank Dr. A. Ö. Ok for providing the VHR[2] benchmark dataset.

References

[1] A. Katartzis and H. Sahli, "A stochastic framework for the identification of building rooftops using a single remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 259–271, Jan. 2008.

[2] B. Sirmacek and C. Ünsalan, "Urban-area and building detection using sift keypoints and graph theory," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 4, pp. 1156–1167, Apr. 2009.

[3] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, Jan. 2012.

[4] M. Cote and P. Saeedi, "Automatic rooftop extraction in nadir aerial imagery of suburban regions using corners and variational level set evolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 1, pp. 313–328, Jan. 2013.

[5] A. O. Ok, "Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts," *ISPRS J. Photogrammetry Remote Sens.*, vol. 86, pp. 21–40, Dec. 2013.

[6] E. Li, J. Femiani, S. Xu, X. Zhang, and P. Wonka, "Robust rooftop extraction from visible band images using higher order CRF," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4483–4495, Aug. 2015.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[8] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 37–44.

[9] Z. Ren, S. Gao, L.-T. Chia, and I.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[10] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Ciruits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Sep. 2006.

[11] W.-T. Li, H.-S. Chang, K.-C. Lien, H.-T. Chang, and Y. Wang, "Exploring visual and motion saliency for automatic video object extraction," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2600–2610, Jul. 2013.

[12] Z. Liu, S. Cui, and Q. Yan, "Building extraction from high resolution satellite imagery based on multi-scale image segmentation and model matching," in *Proc. Earth Observ. Remote Sens. Appl. Int. Workshop*, Jun. 2008, pp. 1–7.

[13] A. Ok, C. Senaras, and B. Yuksel, "Automated detection of arbitrarily shaped buildings in complex environments from monocular VHR optical satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 3, pp. 1701–1717, Mar. 2013.

[14] P. Saeedi and H. Zwick, "Automatic building detection in aerial and satellite images," in *Proc. 10th Int. Conf. Control Autom. Robot. Vis.*, 2008, pp. 623–629.

[15] J. Femiani, E. Li, A. Razdan, and P. Wonka, "Shadow-based rooftop segmentation in visible band images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 5, pp. 2063–2077, May 2015.

[16] B. Sirmacek and C. Ünsalan, "Building detection from aerial images using invariant color features and shadow information," in *Proc. 23rd Int. Symp. Comput. Inf. Sci.*, 2008, pp. 1–5.

[17] A. Manno-Kovacs and T. Sziranyi, "Orientation-selective building detection in aerial images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 108, pp. 94–112, Oct. 2015.

[18] A. Manno-Kovacs and A. Ok, "Building detection from monocular VHR images by integrated urban area knowledge," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2140–2144, Oct. 2015.

[19] Z. Li and L. Itti, "Saliency and gist features for target detection in satellite images," *IEEE Trans. Image Process.*, vol. 20, no. 7, pp. 2017–2029, Jul. 2011.

[20] D. Zhao, J. Shi, J. Wang, and Z. Jiang, "Saliency-constrained semantic learning for airport target recognition of aerial images," *J. Appl. Remote Sens.*, vol. 9, no. 1, 2015, Art. no. 096058.

[21] P. Yang, Z. Jiang, H. Feng, and Y. Ma, "Building detection based on saliency for high resolution satellite images," in *Proc. SPIE*, vol. 8918, pp. 89 180D-1–89 180D-6, 2013.

[22] A. M. Cretu and P. Payeur, "Building detection in aerial images based on watershed and visual attention feature descriptors," in *Proc. 2013 Int. Conf. Comput. Robot Vis.*, May 2013, pp. 265–272.

[23] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," *Adv. Neural Inf. Process. Syst.*, vol. 24, pp. 109–117, 2011.

[24] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. 2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[25] P. Krahenbuhl, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.

[26] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. 12th Eur. Conf. Comput. Vis.—Volume Part III*, 2012, pp. 29–42.

[27] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. 2013 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1155–1162.

[28] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. 2013 IEEE Int. Conf. Comput. Vis.,* Dec. 2013, pp. 153–160.

[29] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. 2011 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.

[30] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[31] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. 2014 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2814–2821.

[32] P. Kraehenbuehl and V. Koltun, "Parameter learning and convergent inference for dense random fields," in *Proc. 30th Int. Conf. Mach. Learn.*, May 2013, vol. 28, no. 3, pp. 513–521.

[33] B. zdemir, S. Aksoy, S. Eckert, M. Pesaresi, and D. Ehrlich, "Performance measures for object detection evaluation," *Pattern Recognit. Lett.*, vol. 31, no. 10, pp. 1128–1137, 2010.

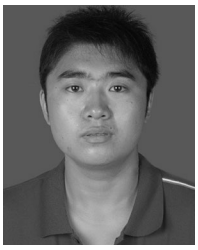
[34] S. Müller and D. W. Zaum, "Robust building detection in aerial images," *Int. Arch. Photogrammetry Remote Sens.*, vol. 36, no. B2/W24, pp. 143–148, 2005.

[35] C. Senaras and F. Yarman Vural, "A self-supervised decision fusion framework for building detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1780–1791, 2016.

[1] http://web.eee.sztaki.hu/remotesensing/building_benchmark.html
[2] http://biz.nevsehir.edu.tr/ozgunok/en/408

Er Li (M'13) received the B.S. degree in automation from Wuhan University, Wuhan, China, in 2007, and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2012.

From 2012 to 2013, he was a Postdoctoral Researcher with the Institute of Software, Chinese Academy of Sciences. From 2013 to 2014, he was a Postdoctoral Researcher with the Department of Engineering and Computing Systems, Arizona State University, Phoenix, AZ, USA. He is currently working as an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include image analysis, computer vision, and compute graphics.



Shibiao Xu (M'15) received the B.S. degree in information engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2009, and the Ph.D. degree in computer science from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His current research interests include vision understanding, three-dimensional reconstruction, and image-based modeling.

Weiliang Meng (M'16) received the B.E. degree in computer science from Civil Aviation University of China, Tianjin, China, in 2003, the M.Sc. degree in computer application from Tianjin University, Tianjin, China, in 2006, and the Ph.D. degree in computer application from the State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Chengdu, China, in 2010.

He is currently an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include geometry modeling, image-based modeling and rendering, and augmented reality.

Xiaopeng Zhang (M'11) received the B.S. and M.S. degrees in mathematics from Northwest University, Xian, China, in 1984 and 1987, respectively, and the Ph.D. degree in computer science from the Institute of Software, Chinese Academy of Sciences, Beijing, China, in 1999.

He is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His main research interests include computer graphics and computer vision.