ISPRS Journal of Photogrammetry and Remote Sensing 147 (2019) 42–55

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings

Qi Chen[a,c], Lei Wang[b,d,*], Yifan Wu[d], Guangming Wu[c], Zhiling Guo[c], Steven L. Waslander[b]

[a] Faculty of Information Engineering, China University of Geosciences (Wuhan), Wuhan 430074, China
[b] Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada
[c] Center for Spatial Information Science, University of Tokyo, Kashiwa 277-8568, Japan
[d] AtlasAI Inc., Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Keywords:
Roof segmentation
Building detection
Large-scale dataset
Automatic mapping
Deep learning

ABSTRACT

As an important branch of deep learning, convolutional neural network has largely improved the performance of building detection. For further accelerating the development of building detection toward automatic mapping, a benchmark dataset bears significance in fair comparisons. However, several problems still remain in the current public datasets that address this task. First, although building detection is generally considered equivalent to extracting roof outlines, most datasets directly provide building footprints as ground truths for testing and evaluation; the challenges of these benchmarks are more complicated than roof segmentation, as relief displacement leads to varying degrees of misalignment between roof outlines and footprints. On the other hand, an image dataset should feature a large quantity and high spatial resolution to effectively train a high-performance deep learning model for accurate mapping of buildings. Unfortunately, the remote sensing community still lacks proper benchmark datasets that can simultaneously satisfy these requirements. In this paper, we present a new large-scale benchmark dataset termed Aerial Imagery for Roof Segmentation (AIRS). This dataset provides a wide coverage of aerial imagery with 7.5 cm resolution and contains over 220,000 buildings. The task posed for AIRS is defined as roof segmentation. We implement several state-of-the-art deep learning methods of semantic segmentation for performance evaluation and analysis of the proposed dataset. The results can serve as the baseline for future work.

1. Introduction

Buildings, which serve as the most significant place for human livelihood, are a key element on digital mapping of urban areas. With the rapid urban development, tremendous efforts are continually allocated to creating and updating location information of buildings for various fields, such as urban planning, land investigation, change detection, and military reconnaissance. Aerial photogrammetry has been an effective technology for accurate mapping of buildings due to its capability for high-resolution imaging over large-scale areas. Unfortunately, automatic mapping of buildings is still limited by the insufficient detection/segmentation accuracy on aerial images. Most cases require considerable amounts of manual intervention.

Recent progress in computer vision (CV) field indicates that, with support from sufficient computing power and large training datasets (Cordts et al., 2016; Deng et al., 2009; Everingham et al., 2010; Lin et al., 2014), deep learning techniques such as convolutional neural network (CNN) (LeCun et al., 1989) can substantially improve the performance of object detection and semantic segmentation from first-person or ground-level imagery (Han et al., 2018; He et al., 2016; Krizhevsky et al., 2012). This condition strongly suggests that deep learning will play a critical role in promoting the accuracy of building detection toward practical applications of automatic mapping; thus, a publicly available dataset is significant for fair comparisons to accelerate the development of this research field. However, although several public datasets address building detection, and numerous existing studies use these datasets for experiments, critical problems remain because of the following reasons:

(1) As Fig. 1 shows, buildings designed for different purposes may present distinct patterns of roof surface and boundary, which requires a deep learning model to construct an extremely high-dimensional feature space to describe the large intra-class variability. Accordingly, an image dataset also requires a wide coverage and

* Corresponding author at: Mechanical and Mechatronics Engineering Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada.
E-mail address: lei.wang@uwaterloo.ca (L. Wang).

https://doi.org/10.1016/j.isprsjprs.2018.11.011
Received 27 July 2018; Received in revised form 10 October 2018; Accepted 9 November 2018

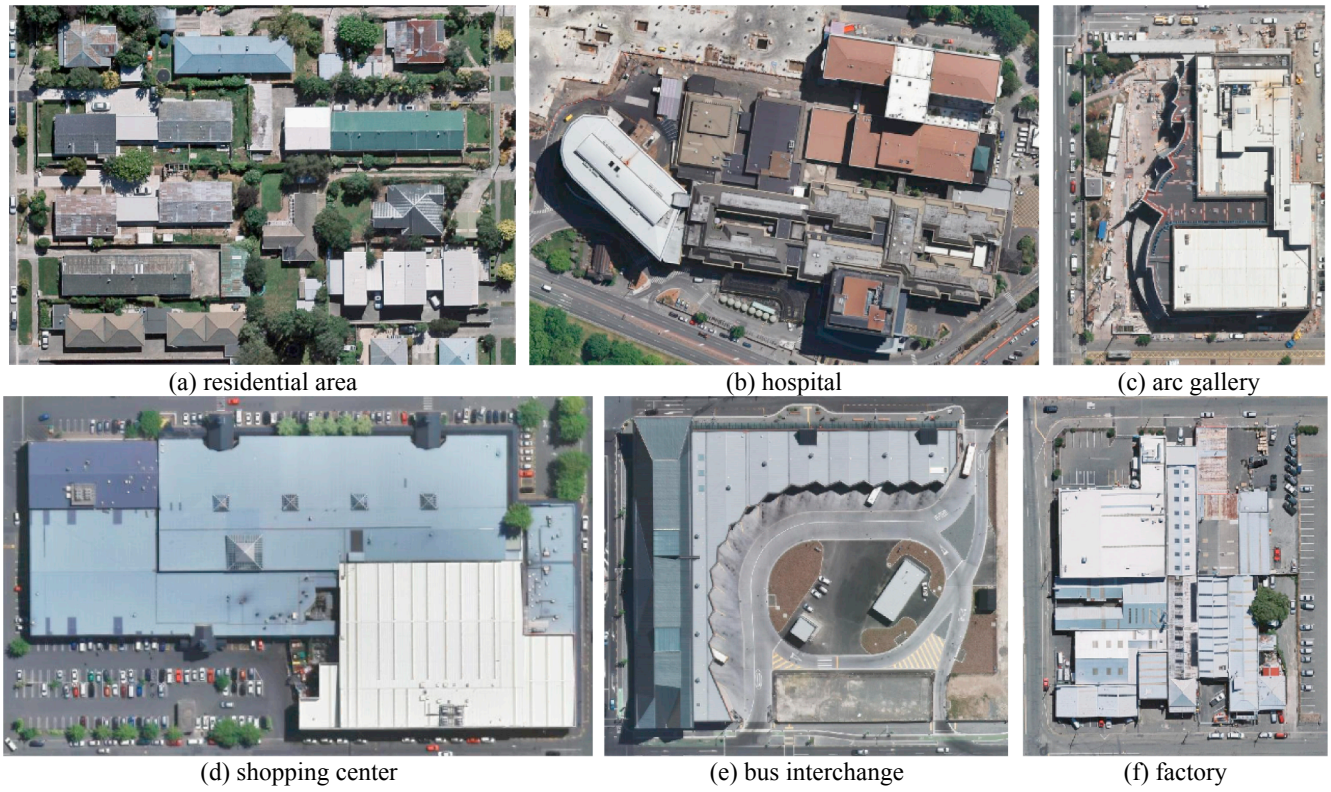

Fig. 1. The intra-class variability of buildings.

large quantity of data to increase the feature diversity for training an effective deep learning model. In addition, to achieve accurate building mapping, the boundaries or edges of buildings should be clearly visible on the images, which further requires the image dataset to have high spatial resolution. However, the remote sensing (RS) community still lacks proper benchmark datasets that can simultaneously satisfy the above requirements. Therefore, a specific dataset that focuses on advancing the development of automatic building mapping should be created.

(2) In digital orthophoto maps (DOMs), the relief displacement caused by imperfect rectification leads to varying misalignment between the *roof outlines* and *footprints* of buildings. As a result, training a model to directly learn the pattern of building footprints will increase the difficulty for effective feature representation. For low buildings, detecting footprints is roughly equivalent to extracting roof outlines, whereas for most high buildings, the task can be more complicated due to remarkable misalignment. Although adopting true DOMs (TDOMs) can theoretically eliminate the influence of relief displacement, other problems associated with TDOM generation may create barriers for accurate roof segmentation. Fig. 5 provides additional details on this issue.

To address the above-mentioned issues, we present a new large-scale benchmark dataset named Aerial Imagery for Roof Segmentation (AIRS). The proposed dataset provides a large quantity of aerial imagery with 7.5 cm resolution and covers almost all of Christchurch City in New Zealand with about 457 km^2 of land and over 220,000 buildings. The task posed for AIRS is defined as *roof* segmentation, and it is entirely unaffected by relief displacement.

Our contributions can be summarized as follows: (1) After refinement of the open-source data released by Land Information of New Zealand (LINZ),[1] we construct a large-scale aerial image dataset containing accurate roof outline annotations. Compared with other existing works, the presented dataset concentrates more on testing the capability of algorithms in precisely segmenting building roofs; such capability is significant for practical application of automatic mapping in the future. (2) Several state-of-the-art semantic segmentation methods based on deep learning models, with performance evaluation and comparison, are implemented and applied on our dataset; these methods can serve as the baseline results for future work. The dataset is now publicly available, and an online leaderboard is also presented to evaluate the results submitted by researchers.[2]

The remainder of the paper is organized as follows. Section 2 reviews the related previous works. Section 3 introduces the AIRS dataset. Section 4 describes the baseline methods. Section 5 presents and discusses the evaluation and comparison results for those methods. Section 6 draws the study conclusions.

2. Previous work

2.1. Related datasets

The ImageNet dataset (Deng et al., 2009), which provides over 14 million labeled images containing 22,000 categories, has enabled an early breakthrough in applying deep learning algorithms for object recognition in the CV field. This dataset was originally designed for image classification, which requires the algorithms to determine the presence of particular objects in an image. Subsequently, other large-scale datasets, such as SUN (Xiao et al., 2010), PASCAL VOC (Everingham et al., 2010), KITTI (Geiger et al., 2012), Microsoft COCO (Lin et al., 2014), and Cityscapes (Cordts et al., 2016), have been introduced and successfully used for recognition tasks, such as object detection and pixel-level and instance-level semantic segmentation. On the other hand, as many studies focused on detection of single category,

[1] https://data.linz.govt.nz.

[2] https://www.airs-dataset.com/.

other public datasets have been created for popular challenges, such as face detection (Huang et al., 2007) or pedestrian detection (Dollár et al., 2012).

The above CV datasets are mainly created for understanding first-person or ground-level images. In comparison, aerial or satellite imagery in the RS field offers a different perspective. Thus, the features learned from the ground-based imagery to RS data are difficult to generalize and apply. To create large-scale datasets specific for RS applications, various attempts have been first made for RS scene/image classification with satellite data. UC-Merced dataset (Yang and Newsam, 2010) is one of the earliest satellite datasets; it consists of 21 categories and 100 images per category. The images are organized as small patches of 256 × 256 pixels with roughly 30 cm resolution and are then used for scene classification. Afterward, other similar datasets, such as WHU-RS (Hu et al., 2015), RSSCN7 (Zou et al., 2015), NWPU VHR-10 (Cheng et al., 2016), AID (Xia et al., 2017), NWPU-RESISC45 (Cheng et al., 2017), and PatternNet (Zhou et al., 2018), which are mostly collected from Google Earth imagery, have been released in succession for this task. The Functional Map of the World is currently the largest public satellite dataset for RS scene classification (Christie et al., 2017); it not only offers over 1,000,000 images containing 63 categories but also provides metadata and revisit images for temporal reasoning of areas located around the world.

Creating a large-scale dataset for semantic labeling, which is another major topic in photogrammetry and RS, is more difficult than scene classification, as pixel-level annotation is required for ground-truth production. A previous significant dataset used on this task is the ISPRS urban classification and building reconstruction benchmark (Rottensteiner et al., 2012), which contains aerial imagery and airborne laser-scanning point clouds for detection of roads, buildings and trees. Later on, this benchmark has been updated and extended for purposes where a 2D semantic labeling contest is organized by providing TDOMs and the corresponding digital surface models (DSMs) derived from dense matching techniques.[3] On the other hand, considering the difficulty of making full annotation for large-scale imagery, many public datasets create benchmarks that only concern buildings or roads, in which the most similar work to ours includes the Massachusetts Buildings Dataset (Mnih, 2013), the Inria Aerial Image Labeling Dataset (Maggiori et al., 2017c), and the SpaceNet Dataset adopted in the recent DeepGlobe challenge (Demir et al., 2018). Section 3.3 provides a detailed comparison between AIRS and these closely related datasets.

2.2. CNN and building detection

Building detection from optical RS imagery has drawn considerable attention over the decades. Prior to deep learning, much of the work in this field is conducted by first extracting features, such as strong edge, shape design, roof color, shadow evidence, local context, or their combination, based on specific knowledge about building rooftops and then applying techniques, such as template matching (Sirmacek and Unsalan, 2009), mathematical morphology (Huang and Zhang, 2012; Zhang et al., 2016), active contours (Ahmadi et al., 2010; Liasis and Stavrou, 2016), graph-based methods (Li et al., 2017; Ok, 2013), random forest (Du et al., 2015), and support vector machine (Inglada, 2007; Turker and Koc-San, 2015) for building detection. In spite of achieving important advances, successful application of these methods is mostly at the cost of a careful manual design of the features and is largely based on prior knowledge and human experience. However, considering the complexity and variety of building shapes, roof surfaces, imaging conditions, and the spatial context, the performance of these methods can easily be limited to certain building shapes, specific regions, and high-quality initial segmentation results by applying low-level hand-crafted features.

As an important branch of deep learning, CNNs have achieved remarkable improvement in image classification and object detection (Han et al., 2018). Recently, they have been increasingly applied in building detection or semantic segmentation from RS imagery due to their capability of extracting high-dimensional and highly discriminative features without manual design. The early applications are usually performed under a patch-based CNN architecture (Alshehhi et al., 2017; Guo et al., 2017, 2016; Vakalopoulou et al., 2015). In this strategy, the CNN model labels every pixel by classifying the image patch centered on that pixel with a sliding window, which leads to extensive overlapping computations for an image in the training and test phases. When dealing with large-scale datasets, the efficiency of these approaches can be significantly affected by the heavy computation cost. Meanwhile, the patch-based CNN models are probably incapable to generate accurate contours and may produce irregular building outlines; thus, post-processing for the segmentation maps remains necessary in certain cases (Alshehhi et al., 2017).

Fully convolutional network (FCN) largely improves the efficiency by transforming the fully connected layers of classic CNN into convolutional layers (Long et al., 2015). FCN provides an end-to-end learning framework for image semantic segmentation. In this framework, instead of predicting a single label for an input patch, a label map is generated to achieve pixel-to-pixel classification. By far, FCNs have been successfully applied in building detection from aerial and satellite imagery (Maggiori et al., 2017a; Shrestha and Vanneschi, 2018). Previous comparative results on semantic segmentation verify the advantage of FCN over the standard patch-based strategy in terms of accuracy and efficiency (Volpi and Tuia, 2017).

Following the fully convolutional fashion, several CNN models under the bottom-up/top-down architecture, such as U-Net (Ronneberger et al., 2015), DeconvNet (Noh et al., 2015), SharpMask (Pinheiro et al., 2016), SegNet (Badrinarayanan et al., 2017), and feature pyramid network (FPN) (Lin et al., 2016), have been proposed to further improve the performance of image segmentation in the CV field. In general, the development of these models is motivated by the reduced spatial information of the input image after going through several convolution and pooling layers. This reduction can give rise to boundary blur or insufficient edge accuracy of segmentation results. Therefore, in the bottom-up/top-down architecture, the feature map with reduced spatial information is fed into a progressive upsampling process rather than directly used for prediction, and lateral connections are applied on feature maps at different scales to preserve the semantic and spatial information.

The above advances in CV inspire new attempts on building detection and image segmentation from RS imagery. The latest work includes explicit multi-scale feature fusion (Maggiori et al., 2017b) and multi-constraints on additional predictions (Wu et al., 2018) based on the bottom-up/top-down architecture for building detection as well as explicit encoding of equivariance in classic CNN (Marcos et al., 2018) or applying multi-task learning based on the hypercolumn architectures (Volpi and Tuia, 2018) for semantic segmentation. Under this background, we implement three deep learning models that utilizes multi-scale feature fusion in different strategies as the baseline methods for the AIRS benchmark.

3. AIRS dataset

3.1. Reason for focusing on roof segmentation

In practical terms, building footprint is the final product that can be used for various applications. However, for human visual systems, roof outline is the key pattern needed for building recognition. As shown in Fig. 2, building footprints can be produced from aerial imagery through human editing via two common methods. The first method is stereo mapping from overlapping image pairs; this method can extract relatively accurate footprints with elevation information. The other method

[3] http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html.

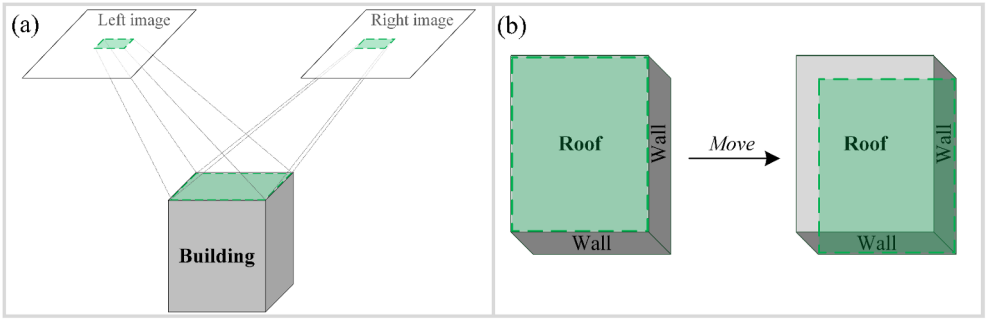

Fig. 2. Manual editing methods to produce building footprints from (a) original stereo-image pairs and (b) DOMs.

extracts 2D footprints from DOMs in an approximate manner: the annotator first draws the roof outline and then moves the outline to align with the visible edges of the footprint. Importantly, these methods both rely on accurate annotation of the roof shape. *Directly* extracting the footprint without drawing the roof outline in advance will cause difficulties for humans. Therefore, we believe that extracting roof outlines can be a significant step for automatic mapping of buildings, as roof segmentation is not only a more straightforward learning task with less difficulty but is also more consistent with human-like interpretation mechanism.

3.2. Data acquisition and refinement

Fig. 3 displays the study area of the AIRS dataset. The area of interest (AOI) covers about 457 km^2, including almost the full area of Christchurch, the largest city in the South Island of New Zealand. The aerial imagery and the corresponding building map are open-source data provided by LINZ Data Service. The photographs were captured during the flying seasons of 2015 and 2016, and the open-source images were ortho-rectified DOMs with RGB channels and 7.5 cm resolution in the projection of New Zealand Transverse Mercator. The building footprints, which are stored as 2D polygons in a shapefile, were produced in June 2016 by manual annotation.

Several annotators are hired to scan and refine all the building outlines within the AOI to provide reliable ground truths for roof segmentation experiment. A large part of the work involves moving the footprints back to the roofs when significant misalignment can be observed, as shown in Fig. 4. The original shapefile included numerous false annotations (probably generated from different data source). These samples are removed during refinement. We check the whole study area at least twice. We cooperate with the team of WHU building dataset (Ji et al., 2018) to finish the annotation work. After refinement, we create a relatively pure ground-truth dataset, which contains 226,342 buildings, for roof segmentation.

3.3. Comparison between AIRS and the current datasets

Table 1 presents the comparison of the statistics between the proposed AIRS dataset and the four closely related datasets. The comparison shows that ISPRS Dataset features a relatively small coverage, which limits its applicability for training high-performance deep learning models. The Massachusetts, Inria, and SpaceNet datasets cover large areas. However, the images provided by these datasets exhibit a relatively lower resolution than those of AIRS (⩾30 cm vs. 7.5 cm). Although lower resolution will cause no significant influence on the object-level recognition of building targets, it will significantly affect the performance of high-precision mapping, especially for small buildings. In addition, the Massachusetts dataset directly adopts the OpenStreetMap (Haklay and Weber, 2008) labels as ground truths, which may bring considerable noises (i.e., incorrect or missing

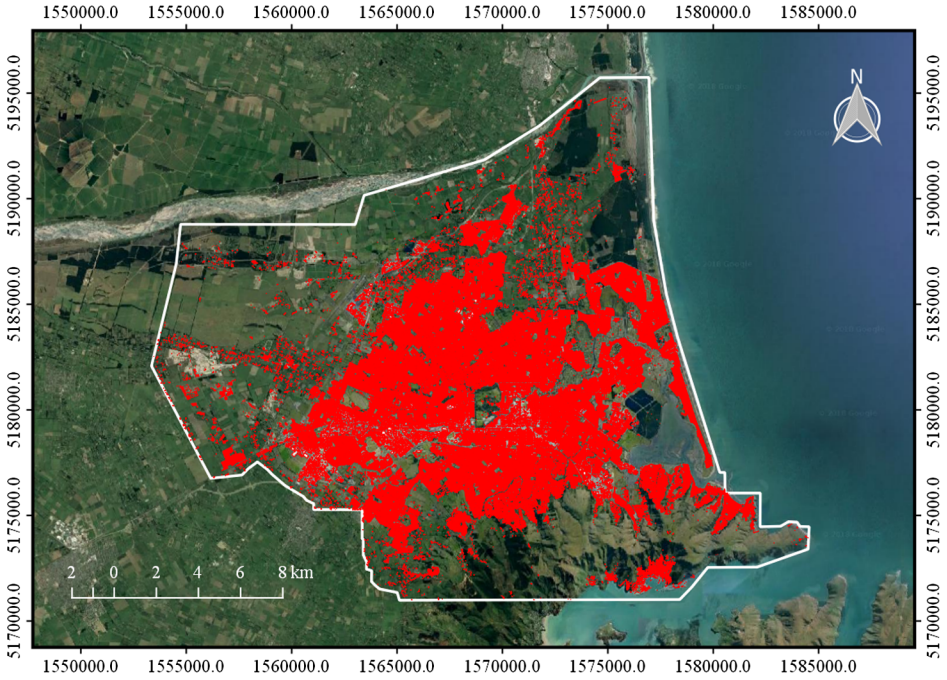

Fig. 3. Geolocation of the proposed AIRS dataset. The white polygon indicates the AOI. All the annotated buildings for benchmarking are highlighted in red.

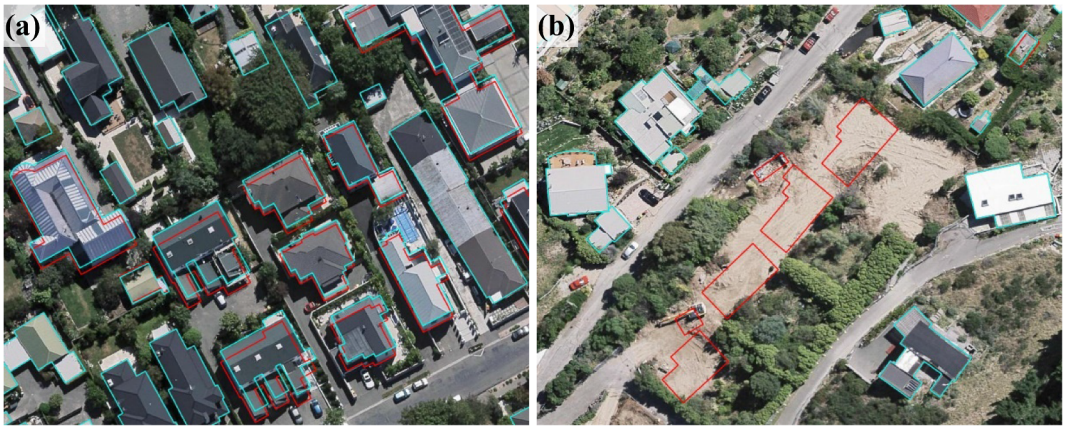

Fig. 4. Building outlines before and after refinement, denoted by red and cyan colors, respectively. The refinement work includes: (a) moving the footprint back to roof outline and (b) removal of false annotations.

annotations) caused by crowdsourcing.

Another noteworthy problem is that the four related datasets provide footprints rather than roof outlines as ground truths. In Fig. 5, some representative patches are cropped from those datasets to indicate this problem. Although many building roofs appear consistent with their footprints in aerial imagery, for high buildings or small view-elevation angle, significant misalignment can occur (see Fig. 4a, b), presenting complicated patterns for learning by a model. Therefore, providing footprints will increase the difficulty and complexity of training effective models for accurate building mapping. The ISPRS dataset provides TDOMs rectified with dense-matching DSM. Thus, misalignment can be partially resolved. However, the errors of dense matching may cause notable texture distortion at building boundaries (see Fig. 4c, d). Considering that edge information is useful for building outline extraction, this dataset may be unsuitable for developing algorithms toward automatic building mapping.

3.4. Splits of dataset

All the aerial images within the AOI are first merged into a single mosaic and then tiled into smaller images for better handling. The tiled images feature a size of 10000 × 10000 pixels, with a 10% overlap between adjacent tiles. As a result, the whole region is tiled into 1047 images. To facilitate research, as shown in Fig. 6, we split the dataset by randomly assigning the images to the following three subsets:

- *Training set.* This set contains 857 images and the corresponding roof labels for training.
- *Validation set.* This set contains 94 images and the corresponding roof labels for validation.
- *Test set.* This set contains 96 images for benchmark testing. Currently, for fair comparison, the corresponding roof labels are not disclosed. They will be used for evaluating the results submitted to the public challenge centered on the AIRS dataset.

4. Baselines and methods

Several state-of-the-art deep learning models are implemented as the baseline methods for benchmarking. ResNet is currently a commonly-used architecture in object recognition, as it allows effective training of very deep neural networks. Moreover, the architecture of ResNet has been proven powerful in feature extraction for RS images (Kaiser et al., 2017; Kemker et al., 2018; Lin et al., 2017; Zhong et al., 2018); thus, in AIRS challenge, the ResNet-101 (He et al., 2016) is adopted as the base structure for the three baseline methods applied on roof segmentation.

4.1. FPN

We first select FPN as a typical model of the bottom-up/top-down architecture and the first baseline method in our work. As shown in Fig. 7, FPN extends the classic ResNet architecture by progressively upsampling the feature map at different spatial scales, creating an in-network feature pyramid. Hence, the model can be considered the combination of a bottom-up structure (the ResNet-101) and a top-down structure, in which several lateral connections are applied to merge the feature maps of the same spatial size from the two structures. The lateral connection aims to combine high-resolution, semantically weak features from the bottom-up structure with low-resolution, semantically strong features from the top-down structure to create semantically strong feature maps with high resolution. What follows the final feature map is an inference structure, which is used to generate prediction results for each pixel of the input image.

4.2. FPN with multi-scale feature fusion (MSFF)

The plain FPN architecture creates feature maps at different spatial scales. However, only a single feature map is used for prediction. Other studies have demonstrated that explicitly fusing multi-scale features

Table 1
Comparison between AIRS and current state-of-the-art benchmarks.

Dataset	ISPRS	Massachusetts	Inria	SpaceNet	AIRS (ours)
Location	Vaihingen/Potsdam	Massachusetts	10 regions in USA and Austria	5 cities around the world	Christchurch
Data Type	CIR/IRRGB + DSM	RGB	RGB	8-band imagery	RGB
Target of Segmentation	6 land-cover classes	Building	Building	Building, road	Building
Coverage (km^2)	1.4/3.4	340	810	5555	457
Resolution (cm)	9/5	100	30	30–50	7.5
OpenStreetMap Labels	–	Yes	–	–	–
TDOM	yes	–	–	–	–
Refined to Roof Outline	–	–	–	–	Yes

Fig. 5. Samples of current state-of-the-art public datasets for building detection. (a) and (b) show the misalignment between roof outlines and footprints in Inria Aerial Image Labeling and SpaceNet datasets, respectively; (c) and (d) show the texture distortion of building boundaries in Vaihingen and Potsdam in ISPRS dataset, respectively.

can improve the performance of image segmentation (Chen et al., 2016; Hariharan et al., 2015; Marmanis et al., 2018); thus, in the second baseline method, we extend the plain FPN by fusing the feature maps on different scales for prediction. This extended version is denoted as FPN with MSFF (FPN + MSFF). As shown in Fig. 8, the feature maps on four different scales from the top-down structure of FPN are rescaled to the same spatial dimension and summed up to generate a new feature map for prediction.

4.3. Pyramid scene parsing network (PSPNet)

The third baseline method is PSPNet (Zhao et al., 2017), which is one of the state-of-the-art deep learning models for semantic image segmentation. In comparison with FPN, which mainly exploits local multi-scale information, PSPNet focuses more on exploring global information in different scales. To enhance the representation of context information among different sub-regions, PSPNet applies a pyramid pooling module on the feature map generated by ResNet to create pooled feature maps at different levels. These features are then merged and used for prediction.

As shown in Fig. 9, the PSPNet applied on AIRS challenge follows its original design. The pyramid pooling module features four levels with different bin sizes and appended to the feature map of ResNet-101. A 1×1 convolutional layer is applied on each pyramid level to reduce its dimension to a fixed depth. The sum of the channels of the four levels is set equal to the dimension of the original feature map. Afterward, all the pooled feature maps are upsampled to the same size of the original feature map. The five feature maps are then concatenated as the final pooling feature, which is followed by an inference structure for prediction.

4.4. Implementation details

Table 2 shows the detailed architectures of the three baseline methods for AIRS challenge. The outputs of Conv1, Conv2, Conv3, Conv4, and Conv5 are denoted as C_1, C_2, C_3, C_4, and C_5, respectively. The feature maps in the feature pyramid corresponding to C_2, C_3, C_4, and C_5 are P_2, P_3, P_4, and P_5, respectively. P_2 is the final feature map of FPN for prediction. For FPN + MSFF, the final feature map is the fusion result of P_2, P_3, P_4, and P_5. For PSPNet, the pooled feature maps PO_1, PO_2, PO_3, and PO_4 are first generated from C_5 by pyramid pooling. Then, the final feature map is created by concatenating C_5 and the four pooled feature maps. The three models share the same inference structure that generates the probability map in a fully convolutional fashion. The final roof segmentation result in the form of a binary map is obtained from the probability map using a threshold of 0.5, which means that each pixel will be classified to the label (roof or non-roof) of the highest probability.

For each convolutional layer of the models, a batch normalization (BN) layer (Ioffe and Szegedy, 2015) is applied after the convolution operation; then, the BN output is further handled by the nonlinear activation function of the rectified linear unit (Nair and Hinton, 2010). All the upsampling operations are performed via bilinear interpolation.

We implement the three baseline methods using the public platform TensorFlow (Abadi et al., 2016) on a 64-bit Ubuntu system equipped with an NVIDIA GeForce GTX 1080 Ti GPU. We initialize the ResNet-101 structure with the publicly available pretrained model (He et al., 2016). The weights of additional parameters are initialized using the standard Gaussian normal distribution with a mean of 0 and a standard deviation of 0.01. We use a weight decay of 0.0001 and a momentum of 0.95 (Krizhevsky et al., 2012).

The tiled images of AIRS are further divided into small crops to feed into the models. A large batch size is helpful for efficient model training; however, since the image spatial resolution in AIRS is high, it is important to have a relatively large crop size to provide sufficient context information. Due to the limited GPU memory, larger crop size means smaller batch size. With careful tuning, we set crop size to 401×401 pixels and use batch size 2 in all our experiments. A sliding

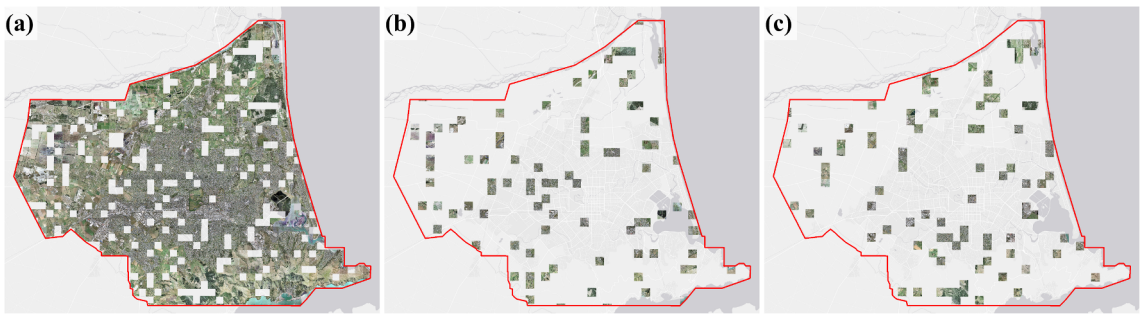

Fig. 6. Geographical distribution of images in: (a) training set, (b) validation set, and (c) test set of AIRS dataset.

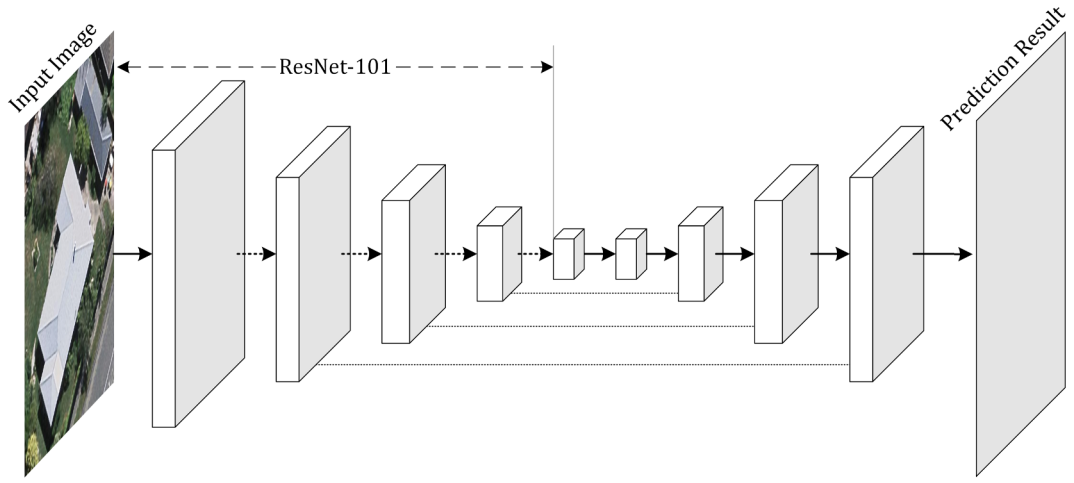

Fig. 7. Architecture of FPN.

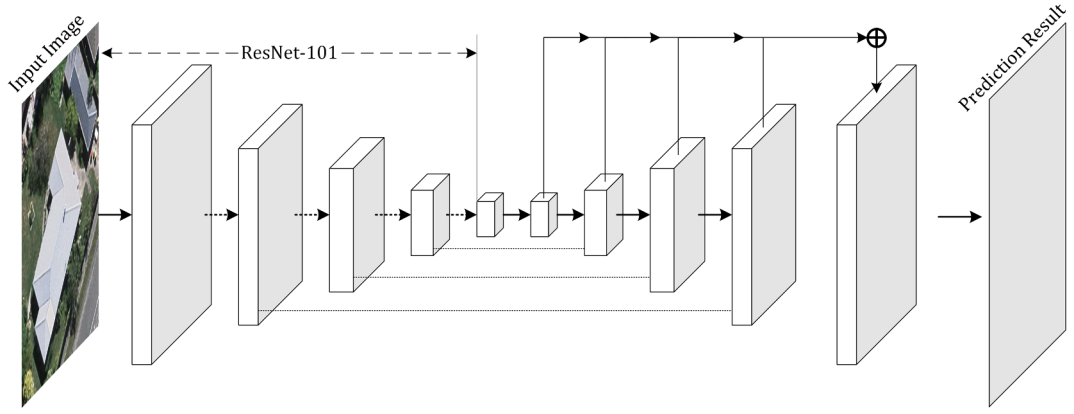

Fig. 8. Architecture of the FPN + MSFF.

window with a step size of 134 (one third of 401) is used to generate the crops. Thus, every two consecutive crops have a 66.6% overlap horizontally or vertically.

We use several standard data augmentation methods to resist overfitting. We first apply random left-right flipping (with 50% probability) for each training crop. Besides that, random scaling with ratio of 0.8, 0.9, 1.0, 1.1 or 1.2 is applied, the scaled crops are then cropped or padded to the fixed size of 401 × 401 pixels. Afterward, random rotation of 0, 90, 180 or 270 degrees is performed with equal probability. Furthermore, dropout with probability of 0.05 is used to randomly mask out pixel values. Meanwhile, with 50% probability, zero-mean Gaussian noise with a variance of 5 is added for each channel of the image independently.

We use the union of training set and validation set of AIRS (951

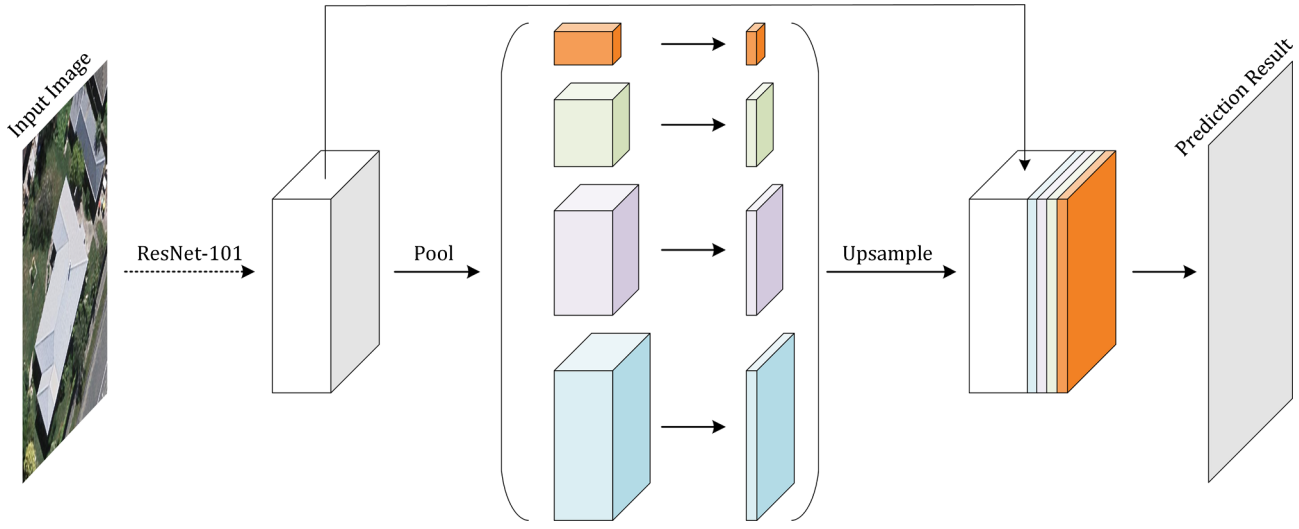

Fig. 9. Architecture of PSPNet.

Table 2
Configurations of the three baseline deep learning models for AIRS challenge.

Layer/Module	Output name	Output size	FPN	FPN + MSFF	PSPNet
Conv1	C_1	201 × 201	7 × 7, 64, stride 2		
Conv2_x	C_2	101 × 101	3 × 3 max pool, stride 2		
			$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix} \times 3$		
Conv3_x[a]	C_3	51 × 51	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix} \times 4$		
Conv4_x[a]	C_4	26 × 26 (51 × 51)	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix} \times 23$		
Conv5_x[a]	C_5	13 × 13 (51 × 51)	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix} \times 3$		
Conv6	P_5	13 × 13	1 × 1, 256		–
Lateral4	P_4	26 × 26	C_4 ~ (1 × 1, 256); P_5 ~ upsample sum		
Lateral3	P_3	51 × 51	C_3 ~ (1 × 1, 256); P_4 ~ upsample sum		
Lateral2	P_2 (Final Feature of FPN)	101 × 101	C_2 ~ (1 × 1, 256); P_3 ~ upsample sum		
Fusion	Final Feature of FPN + MSFF	201 × 201	–	P_2 ~ (3 × 3, 256), upsample; P_3 ~ (3 × 3, 256), upsample; P_4 ~ (3 × 3, 256), upsample; P_5 ~ (3 × 3, 256), upsample sum	–
Pyramid Pool	PO_1, PO_2, PO_3, PO_4	1 × 1, 2 × 2, 3 × 3, 6 × 6	–		max pool 1 × 1, 256
Concat	Final feature of PSPNet	51 × 51			C_5; (PO_1 ~ PO_4) ~ upsample concatenate
Inference	Probability	401 × 401	3 × 3, 512 dropout 1 × 1, 2; upsample; softmax		

[a] For FPN and FPN + MSFF, the feature map is downsampled with a stride of 2 by Conv3_1, Conv4_1, and Conv5_1; for PSPNet, following the original implementation, no downsampling is performed in these layers.

images, more than 5 million crops in total) to train and report results on the 96 images of the test set. The average building/non-building ratio of all crops is 0.738, which indicates that there is no obvious bias between positive and negative training samples. Each model is trained end-to-end by mini-batch stochastic gradient descent and backpropagation algorithms (LeCun et al., 1989) for approximately 2 weeks. The numbers of iterations for PSPNet, FPN, and FPN + MSFF are 696 K, 532 K, and 516 K, respectively, which are inversely proportional to the complexity of the models. During training, the initial learning rate is set as 0.001, which decreases by 10 times at 300 K. To reduce random variations for each baseline method, the median error of the last five models with 10 K iteration interval (e.g., PSPNet models at 656 K, 666 K, 676 K, 686 K, and 696 K) is reported for comparison (Goyal et al., 2017). In prediction phase, each test image is divided into crops and fed into the models. The generated probability maps of all the crops are merged to one probability map of 10,000 × 10,000 pixels. For the overlapping areas between crops, the average probability is calculated as output.

5. Results and discussion

5.1. Overall evaluation

The evaluation metrics of intersection over union (IoU) (Jaccard, 1912) and F1-score (Powers, 2011) are used to reflect the overall performance of the baseline methods. On the other hand, precision and recall, which indicate the correctness and completeness of the roof segmentation results, respectively, are also adopted for evaluation. As shown in Table 3, we first evaluate the segmentation results of all images in the test set as a whole for each method.

Table 3
Evaluation of the three baseline methods for all the test images in AIRS. For every metric, the highest value is highlighted in bold.

Method	IoU	F1-score	Precision	Recall
FPN	0.882	0.937	**0.963**	0.913
FPN + MSFF	0.888	0.941	0.958	0.924
PSPNet	**0.899**	**0.947**	0.961	**0.933**

In general, the three baseline methods achieve good performance with IoU and F1-score higher than 0.88 and 0.93, respectively. Considering the large quantity and random distribution of the test data (96 images, each with 10,000 × 10,000 pixels), the complexity and variety of the dataset can be well managed by very deep neural networks on roof segmentation.

Among the three methods, PSPNet yields the highest IoU and F1-score. Results show similar performances of the three models in suppressing false positives. The advantage of PSPNet over the two versions of FPN mainly lies in the completeness of segmentation results. On the other hand, FPN + MSFF slightly outperforms FPN by achieving high recall, which shows that multi-scale feature fusion can improve the performance of models in completeness of the detected roofs.

Also note that the final feature map of PSPNet for prediction has the lowest spatial resolution (51 × 51 pixels, one-eighth of the input size), which indicates that the architecture design of the network plays a more important role in promoting segmentation performance rather than simply improving the resolution of the feature map through lateral connection and upsampling.

Fig. 10. Evaluation of the baseline methods in different areas. The yellow, red, blue, and white pixels of the evaluation maps represent the prediction results of true positives, false positives, false negatives, and true negatives, respectively. The values of I, F, P, and R refer to the evaluation results of IoU, F1-score, precision, and recall metrics for the whole image, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5.2. Detailed analysis

As shown in Fig. 10, representative images are selected from the test set to further analyze the performance of the baseline methods in different types of regions. Specifically, (a) is a residential area with densely distributed houses; (b) is the central business area of the city, with many irregular buildings around a large square; (c) is a typical industrial area, including many factory buildings at both sides of the railways; (d) is a complex area, which contains factories, government buildings, schools, and residential houses; and (e) is a village area with large vegetation coverage and few houses.

The baseline methods perform differently in different areas. In residential and village areas (Fig. 9a, e), building roofs can be mostly well segmented by the three methods, and most of the vegetation area can be excluded correctly. In these areas, high accuracy can be achieved by the three methods probably due to the consistency and regularity of roof size and shape. However, no improvement is reported by the MSFF strategy over the plain FPN. In the central business area (Fig. 10b), although PSPNet slightly outperforms the two FPNs, all the methods perform poorly mainly because of low recall. A possible explanation is that the roof surfaces in this area exhibit more complex patterns, especially for buildings with irregular shapes or small parts stacking on the major structure. In the industrial area (Fig. 10c), most factory buildings are well detected, and few railroad cars are mistaken. The performance difference among the methods is mainly caused by their recall capability (0.889 vs. 0.901 vs. 0.930), which is evidently reflected in the segmentation of the two large buildings in the image. The results in the complex area (Fig. 10d) validate our findings in other areas: for most residential houses with regular shapes and smooth roof surfaces, all the methods can achieve high segmentation accuracy; for industrial areas, especially large factory buildings, PSPNet outperforms the FPNs by detecting more complete structures, whereas FPN + MSFF presents a slightly improved performance compared with the plain FPN.

Fig. 11 shows typical scenarios where remarkable differences in the three methods can be observed. Specifically, for the building centered in (a), the two FPNs fail to segment the composite structure as a single object, whereas PSPNet achieves better results by “linking” the three parts together. The building pointed by the arrow in (b) possesses a two-layer structure. The shadow cast by the high parts prevents FPN from detecting the roof completely, whereas the other two methods are not affected by the shadow. The facilities and shadow on top of the building in (c) provide the roof with a complicated pattern, leading to significant miss detection in the results of the two FPNs. By contrast, PSPNet can still roughly segment the whole roof. The building centered in (d) is a large factory (with a bounding box of about 1700 × 2500 pixels); therefore, it must be divided into several crops (each with 401 × 401 pixels) and processed separately for prediction. In this case, the two FPNs suffer from serious miss detection of the central part of the roof in varying degrees, whereas PSPNet outperforms them with less false negatives. (e) shows several residential buildings with multiple layers or irregular shapes, which create certain difficulties for the three methods. However, an advantage of PSPNet can still be observed: it shows better performance in suppressing false detection when the FPNs mix up the ground or grass with the roof (as pointed by the black arrows). PSPNet also shows better capability in correctly segmenting the unsmooth area on the roof pointed by the green arrow.

As stated above, although the three methods exhibit close performances over the whole test area, for specific scenarios, PSPNet not only shows a certain advantage over the two FPNs in preserving the completeness of the roofs but also achieves better robustness to the irregularity and roughness of the roof surfaces. A possible explanation is that the pyramid pooling module in PSPNet aids the model in learning the context information of the input image, which reasonably leads to better global recognition of the building targets.

5.3. Areas for improvement

The current segmentation performance of the baseline methods still requires considerable improvement. Fig. 12 shows typical problems that have not been resolved. Occlusion of trees and shadows is one of the most representative problems, as shown in Fig. 12a. The baseline models cannot conduct human-like reasoning in these cases, thus preventing them from generating more practical segmentation results. Moreover, the models are easily confused by buildings under construction (Fig. 12b), although the surfaces of such objects usually present different patterns. In the central business area, when encountering buildings with complex roof surfaces or uncommon facilities on the top, the models can almost ignore the whole target (Fig. 12c) or omit the unsmooth area (Fig. 12d). On the other hand, surfaces presenting similar patterns with roofs, such as the square in Fig. 12d, can also be falsely detected. The last problem often occurs when processing large buildings. As shown in Fig. 12e, difficulty arises for extraction of global contextual information of the building due to the limited crop size. Hence, the smooth surface of the roof can be easily detected as roads or other negative samples. Generally, although the baseline methods perform well on detecting regular or typical roofs, they are still challenged by the intra-class variabilities of buildings and inter-class similarities between buildings and other background objects in certain cases.

The above problems provide at least two important directions for improvement. The first direction involves applying a generative adversarial network (Goodfellow et al., 2014) on the segmentation task to improve the reasoning capability of the models; thus, the complete structure can possibly be segmented even if the roof is not completely visible. The second involves conducting instance-aware segmentation (Dai et al., 2015) in different spatial scales; thus, the deep learning models can extract and understand object-level contextual information efficiently, which will aid in detecting large buildings. In addition, expanding training samples for specific types of areas or buildings also bears significance in enhancing the generalization capability of the models.

6. Conclusion

In this work, we present a publicly available large-scale aerial dataset for developing and evaluating methods to advance the development of automatic building mapping. We conduct refinement on the open-source data to accurately align the polygons with roof outlines and remove false annotations and to provide reliable ground truths for roof segmentation experiment. Furthermore, three state-of-the-art deep learning methods, including two versions of FPNs and PSPNet, are implemented as the baseline methods for benchmarking. The evaluation verifies the high performance of the deep learning models on roof segmentation and also demonstrates areas for future improvement.

Moreover, we want to point out a few limitations of this work. First and foremost, the proposed dataset focuses on evaluating the roof segmentation performance on aerial images, which means that the deep learning model trained by this dataset may not work well on images captured from satellite or other platforms. In this case, the dataset can be used to pre-train deep learning models for other types of images. Compared to training from scratch, fine-tuning the pre-trained model on target satellite images requires less samples. It also should be noted that when large relief displacement exists, roof segmentation results may not represent the actual building geolocations. To obtain more accurate geolocation results of buildings, providing more information (i.e., footprint along with roof outline) for the model to learn is a future direction worth exploring. These problems will be further studied in our follow-up work. Meanwhile, we will collect more high-resolution aerial imagery data from different regions around the world to enhance the variety of the current dataset.

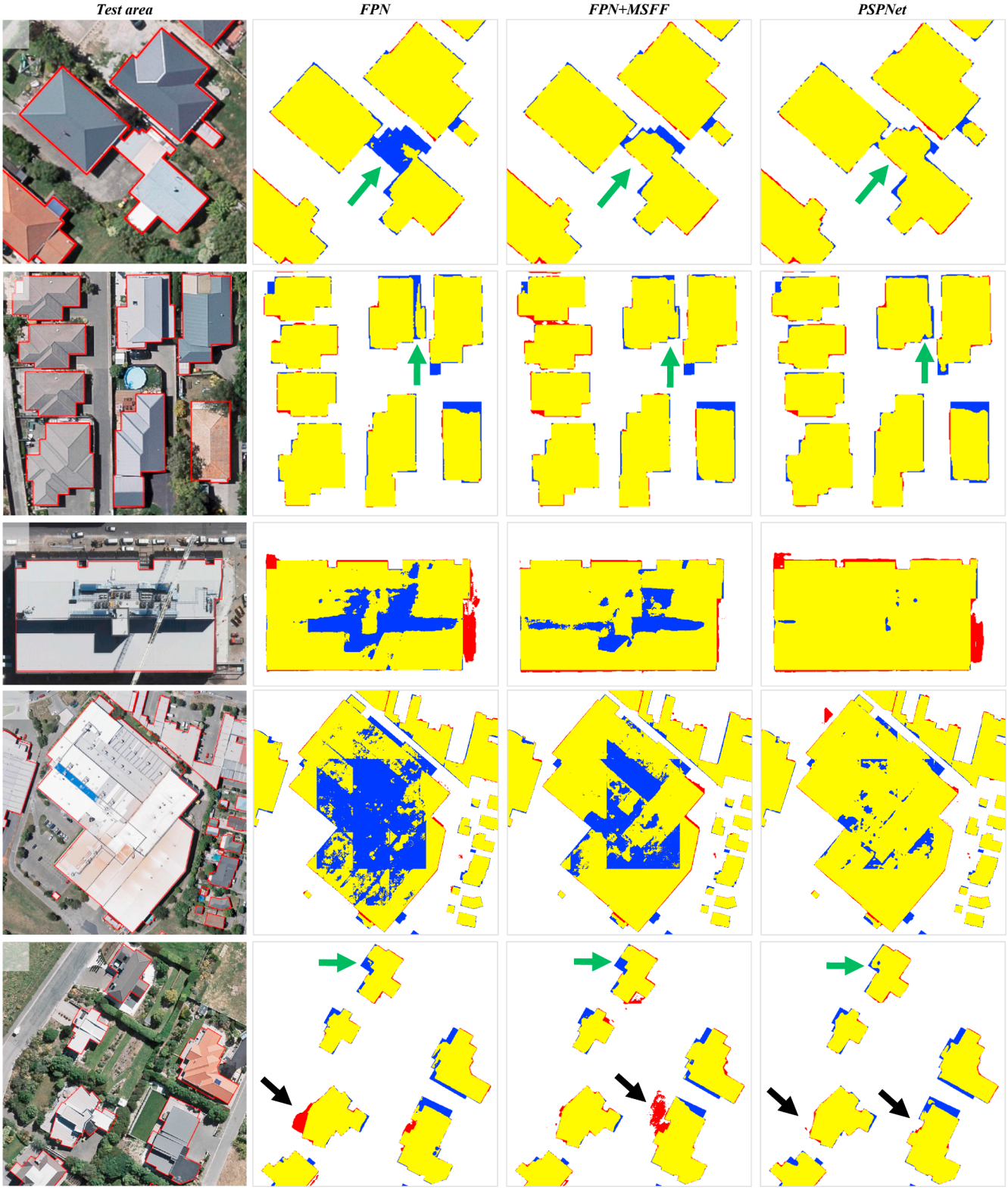

Fig. 11. Typical scenarios that demonstrate the performance differences among the baseline methods. The yellow, red, blue, and white pixels of the evaluation maps represent the prediction results of true positives, false positives, false negatives, and true negatives, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

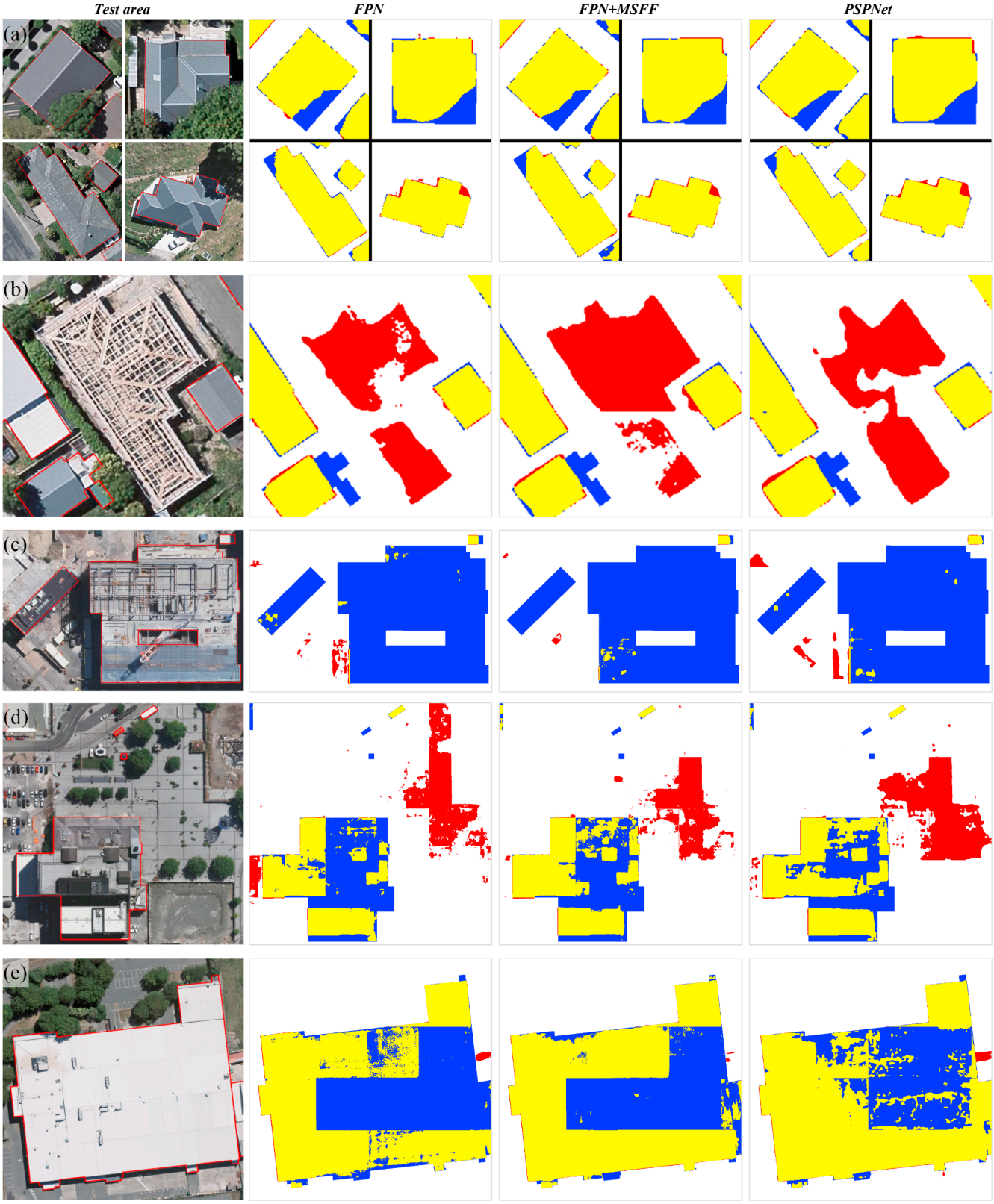

Fig. 12. Main problems of the baseline methods for future improvement. The yellow, red, blue, and white pixels of the evaluation maps represent the prediction results of true positives, false positives, false negatives, and true negatives, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgements

This work was partially supported by National Natural Science Foundation of China (No. 41601506); and China Postdoctoral Science Foundation (No. 2016M590730); and Japan Society for the Promotion of Science (JSPS) Grant (No. 16K18162). We want to thank National Topographic Office of New Zealand for their kind open-souring of the data, and also thank Prof. Shunping Ji from Wuhan University for helping with the annotation of the proposed dataset.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16. USENIX Association, Berkeley, CA, USA, pp. 265–283.

Ahmadi, S., Zoej, M.J.V., Ebadi, H., Moghaddam, H.A., Mohammadzadeh, A., 2010. Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours. Int. J. Appl. Earth Obs. Geoinf. https://doi.org/10.1016/j.jag.2010.02.001.

Alshehhi, R., Marpu, P.R., Woon, W.L., Mura, M.D., 2017. Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 130, 139–149. https://doi.org/10.1016/j.isprsjprs.2017.05.002.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615.

Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L., 2016. Attention to scale: scale-aware semantic image segmentation. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3640–3649. https://doi.org/10.1109/CVPR.2016.396.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE. https://doi.org/10.1109/JPROC.2017.2675998.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7405–7415. https://doi.org/10.1109/TGRS.2016.2601622.

Christie, G., Fendley, N., Wilson, J., Mukherjee, R., 2017. Functional Map of the World 2.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding. http://doi.org/10.1109/CVPR.2016.350.

Dai, J., He, K., Sun, J., 2015. Instance-aware semantic segmentation via multi-task network cascades, pp. 3150–3158. http://doi.org/10.1109/CVPR.2016.343.

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. DeepGlobe 2018: a challenge to parse the earth through satellite images, pp. 172–181. http://doi.org/10.1109/CVPRW.2018.00031.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 248–255. https://doi.org/10.1109/CVPRW.2009.5206848.

Dollár, P., Wojek, C., Schiele, B., Perona, P., 2012. Pedestrian detection: an evaluation of the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 34, 743–761. https://doi.org/10.1109/TPAMI.2011.155.

Du, S., Zhang, F., Zhang, X., 2015. Semantic classification of urban buildings combining VHR image and GIS data: An improved random forest approach. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2015.03.011.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88, 303–338. https://doi.org/10.1007/s11263-009-0275-4.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27. https://doi.org/10.1017/CBO9781139058452.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K., 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. http://doi.org/10.1561/2400000003.

Guo, Z., Chen, Q., Wu, G., Xu, Y., Shibasaki, R., Shao, X., 2017. Village building identification based on Ensemble Convolutional Neural Networks. Sensors (Switzerland).

Guo, Z., Shao, X., Xu, Y., Miyazaki, H., Ohira, W., Shibasaki, R., 2016. Identification of village building via Google Earth images and supervised machine learning methods. Remote Sens. https://doi.org/10.3390/rs8040271.

Haklay, M., Weber, P., 2008. OpenStreetMap: user-generated street maps. IEEE Pervasive Comput. 7, 12–18. https://doi.org/10.1109/MPRV.2008.80.

Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D., 2018. Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process. Mag. https://doi.org/10.1109/MSP.2017.2749125.

Hariharan, B., Arbeláez, P., Girshick, R., Malik, J., 2015. Hypercolumns for object segmentation and fine-grained localization. In: Proc. IEEE Comput. Soc. Conf. Comput Vis. Pattern Recognit. 07–12–June, pp. 447–456. https://doi.org/10.1109/CVPR.2015.7298642.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit. pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

Hu, J., Jiang, T., Tong, X., Xia, G.-S., Zhang, L., 2015. A benchmark for scene classification of high spatial resolution remote sensing imagery. In: Geosci. Remote Sens. Symp. (IGARSS), 2015 IEEE Int. pp. 5003–5006. https://doi.org/10.1109/IGARSS.2015.7326956.

Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled faces in the wild: a database for studying face recognition in unconstrained environments, vol. 1, 07–49 doi: 10.1.1.122.8268.

Huang, X., Zhang, L., 2012. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. https://doi.org/10.1109/JSTARS.2011.2168195.

Inglada, J., 2007. Automatic recognition of man-made objects in high resolution optical remote sensing images by SVM classification of geometric image features. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2007.05.011.

Ioffe, S., Szegedy, C., 2015. Batch Normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning – Volume 37, ICML'15. JMLR.org, pp. 448–456.

Jaccard, P., 1912. The distribution of the flora in the alphine zone. New Phytol. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multi-source building extraction from an open aerial and satellite imagery dataset. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2018.2858817.

Kaiser, P., Wegner, J.D., Lucchi, A., Jaggi, M., Hofmann, T., Schindler, K., 2017. Learning aerial image segmentation from online maps. IEEE Trans. Geosci. Remote Sens. 55, 6054–6068. https://doi.org/10.1109/TGRS.2017.2719738.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. 0–1. https://doi.org/10.1016/j.isprsjprs.2018.04.014.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. https://doi.org/10.1016/j.protcy.2014.09.007.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. https://doi.org/10.1162/neco.1989.1.4.541.

Li, E., Xu, S., Meng, W., Zhang, X., 2017. Building extraction from remotely sensed images by integrating saliency cue. IEEE J Sel. Top. Appl. Earth Obs. Remote Sens. 10, 906–919. https://doi.org/10.1109/JSTARS.2016.2603184.

Liasis, G., Stavrou, S., 2016. Building extraction in satellite images using active contours and colour features. Int. J. Remote Sens. https://doi.org/10.1080/01431161.2016.1148283.

Lin, H., Shi, Z., Zou, Z., 2017. Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images. IEEE Geosci. Remote Sens. Lett. 1–5.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2016. Feature Pyramid Networks for Object Detection. http://doi.org/10.1109/CVPR.2017.106.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: common objects in context. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 8693 LNCS, pp. 740–755. http://doi.org/10.1007/978-3-319-10602-1_48.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation ppt. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. https://doi.org/10.1109/CVPR.2015.7298965.

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

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017b. High-resolution aerial image labeling with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2017.2740362.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., Semantic, C., 2017. Can Semantic Labeling Methods Generalize to Any City ? The Inria Aerial Image Labeling Benchmark to cite this version : HAL Id : hal-01468452.

Marcos, D., Volpi, M., Kellenberger, B., Tuia, D., 2018. Land cover mapping at very high resolution with rotation equivariant CNNs: towards small yet accurate models. ISPRS J. Photogramm. Remote Sens. 1–12. https://doi.org/10.1016/j.isprsjprs.2018.01.021.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.

Mnih, V., 2013. Machine Learning for Aerial Image Labeling. (PhD Thesis).

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10. Omnipress, USA, pp. 807–814.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, https://doi.org/10.1109/ICCV.2015.178.

Ok, A.O., 2013. ISPRS Journal of Photogrammetry and Remote Sensing Automated detection of buildings from single VHR multispectral images using shadow information and graph cuts. ISPRS J. Photogramm. Remote Sens. 86, 21–40. https://doi.org/10.1016/j.isprsjprs.2013.09.004.

Pinheiro, P.O., Lin, T.-Y., Collobert, R., Dollár, P., 2016. Learning to refine object

segments. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016. Springer International Publishing, Cham, pp. 75–91.
Powers, D.M.W., 2011. Evaluation: From Precision, Recall and F-Measure To Roc, Informedness Markedness & Correlation. J. Mach. Learn. Technol doi: 0.1.1.214.9232.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, pp. 1–8. http://doi.org/10.1007/978-3-319-24574-4_28.
Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. I-3 293–298. https://doi.org/10.5194/isprsannals-I-3-293-2012.
Shrestha, S., Vanneschi, L., 2018. Improved fully convolutional network with conditional random fields for building extraction. Remote Sens. 10.
Sirmacek, B., Unsalan, C., 2009. Urban-area and building detection using SIFT keypoints and graph theory. IEEE Trans. Geosci. Remote Sens. 47, 1156–1167. https://doi.org/10.1109/TGRS.2008.2008440.
Turker, M., Koc-San, D., 2015. Building extraction from high-resolution optical spaceborne images using the integration of support vector machine (SVM) classification, Hough transformation and perceptual grouping. Int. J. Appl. Earth Obs. Geoinf. https://doi.org/10.1016/j.jag.2014.06.016.
Vakalopoulou, M., Karantzalos, K., Komodakis, N., Paragios, N., 2015. Building detection in very high resolution multispectral data with deep learning features. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), https://doi.org/10.1109/IGARSS.2015.7326158.
Volpi, M., Tuia, D., 2018. Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. ISPRS J. Photogramm. Remote Sens. 144, 48–60. https://doi.org/10.1016/j.isprsjprs.2018.06.007.
Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2016.2616585.
Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R., 2018. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. Remote Sens. https://doi.org/10.3390/rs10030407.
Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2017.2685945.
Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A., 2010. SUN database: large-scale scene recognition from abbey to zoo. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492. https://doi.org/10.1109/CVPR.2010.5539970.
Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems – GIS '10, pp. 270. https://doi.org/10.1145/1869790.1869829.
Zhang, Q., Huang, X., Zhang, G., 2016. A morphological building detection framework for high-resolution optical imagery over urban areas. IEEE Geosci. Remote Sens. Lett. 13, 1388–1392. https://doi.org/10.1109/LGRS.2016.2590481.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, https://doi.org/10.1109/CVPR.2017.660.
Zhong, Y., Han, X., Zhang, L., 2018. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 138, 281–294. https://doi.org/10.1016/j.isprsjprs.2018.02.014.
Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. . PatternNet: a benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2018.01.004.
Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12, 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299.