IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 19, 2019, accepted July 10, 2019, date of publication July 15, 2019, date of current version August 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2928646

Super-Resolution Integrated Building Semantic Segmentation for Multi-Source Remote Sensing Imagery

ZHILING GUO[1], GUANGMING WU[1], XIAOYA SONG[1,2], WEI YUAN[1], QI CHEN[3], HAORAN ZHANG[1], XIAODAN SHI[1], MINGZHOU XU[1], YONGWEI XU[1], RYOSUKE SHIBASAKI[1], AND XIAOWEI SHAO[1,4]

[1]Center for Spatial Information Science, The University of Tokyo, Kashiwa 277-8568, Japan
[2]Key Laboratory of Cold Region Urban and Rural Human Settlement Environment Science and Technology, School of Architecture, Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin 150006, China
[3]School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China
[4]Earth Observation Data Integration and Fusion Research Initiative, The University of Tokyo, Tokyo 153-8505, Japan

Corresponding author: Xiaowei Shao (shaoxw@iis.u-tokyo.ac.jp)

This work was supported in part by the Grant-in-Aid for Early-Career Scientists from the Japan Ministry of Education, Culture, Sports, Science, and Technology (MEXT), under Grant 19K15260, and in part by the Japan Society for the Promotion of Science (JSPS).

ABSTRACT Multi-source remote sensing imagery has become widely accessible owing to the development of data acquisition systems. In this paper, we address the challenging task of the semantic segmentation of buildings via multi-source remote sensing imagery with different spatial resolutions. Unlike previous works that mainly focused on optimizing the segmentation model, which did not enable the severe problems caused by the unaligned resolution between the training and testing data to be fundamentally solved, we propose to integrate SR techniques with the existing framework to enhance the segmentation performance. The feasibility of the proposed method was evaluated by utilizing representative multi-source study materials: high-resolution (HR) aerial and low-resolution (LR) panchromatic satellite imagery as the training and testing data, respectively. Instead of directly conducting building segmentation from the LR imagery by using the model trained using the HR imagery, the deep learning-based super-resolution (SR) model was first adopted to super-resolved LR imagery into SR space, which could mitigate the influence of the difference in resolution between the training and testing data. The experimental results obtained from the test area in Tokyo, Japan, demonstrate that the proposed SR-integrated method significantly outperforms that without SR, improving the Jaccard index and kappa by approximately 19.01% and 19.10%, respectively. The results confirmed that the proposed method is a viable tool for building semantic segmentation, especially when the resolution is unaligned.

INDEX TERMS Building segmentation, deep learning, remote sensing, super-resolution.

I. INTRODUCTION

Since the achievement of a wide variety of vital tasks such as urban monitoring, demographic modeling, and disaster surveillance strongly rely on the detection of important land features, the semantic segmentation of buildings via remote sensing imagery has become a significant research topic in recent years [1], [2]. To conduct the building segmentation task, the methods such as graph theory-based [3] and clustering-based [4] are usually inappropriate due to the complexity and variety of remote sensing imagery [5]. Furthermore, in terms of conventional classification-based segmentation methods [6]–[8], which mainly rely on handcrafted features, the concentration on merely a few of the particular and salient features, such as the structure, outline, and color, means that the models inevitably lack strong capability to represent the abstract characteristics of buildings [9]. Thus, the high-performance generalization of building segmentation remains a formidable challenge.

Lately, the rapid development of deep convolutional neural networks (DCNN) [10] has led to the construction of several models that have achieved great success with the task of

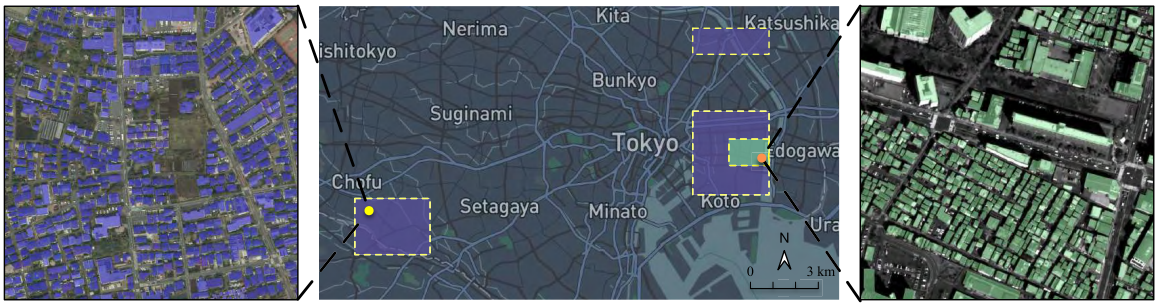
The associate editor coordinating the review of this manuscript and approving it for publication was Jan Chorowski.

building semantic segmentation in terms of both accuracy and computational efficiency. The pioneering work on the topic can be traced to 2015, when Paisitkriangkrai *et al.* [11] proposed effective semantic pixel labeling using CNN and conditional random fields (CRF) [12] to perform building segmentation with competitive classification accuracy. Subsequently, in 2016, inspired by fully convolutional networks (FCNs) [13], Kampffmeyer *et al.* [14] designed architecture that allows end-to-end learning of the pixel-to-pixel semantic segmentation for buildings, and small land features were proven to be detected accurately as well. In 2017, Guo *et al.* [15] utilized ensemble convolutional neural networks (ECNN) to identify village buildings by using Google's satellite map and Bing Maps with high accuracy. And the development of hourglass-shaped networks (HSNs) such as UNet [16] and SegNet [17] motivated Liu *et al.* [18] to propose an enhanced HSN. Their model included an inception module, which replaced the typically used convolutional layers, and which results in a network with multi-scale receptive areas with rich context. In contrast to studies that aimed to modify the structure of CNN, Bischke *et al.* (2017) [19] and Wu *et al.* (2018) [20] chose to optimize the loss function by applying multi-task loss and multi-constraint loss, respectively. The results demonstrated that optimization of the loss function could significantly improve the performance of classic FCNs in certain building segmentation tasks. In addition, to facilitate the development of parsing the earth through satellite imagery, a challenge named Deepglobe [21] was held during the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) in 2018, and the following is a brief overview of some representative studies. Zhao *et al.* [22] conducted extraction by using Mask R-CNN [23] with building boundary regularization. Delassus and Giot [24] proposed a fusion strategy based on a deep combiner using segmentation of both the results of different CNNs and the input data to segment. By using a new FCN variant named TernausNetV2, Iglovikov *et al.* [25] could extract buildings even at the instance level. Focusing on small buildings, Dickenson and Gueguen [26] utilized CNN to output rotated rectangles for symbolized building footprint extraction. Li *et al.* [27] used a building extraction method based on ensemble learning to perform the segmentation. Furthermore, a recent study in 2019, Wu *et al.* [28] utilized stacked fully convolutional networks and a feature alignment framework for multi-label land-cover segmentation with high accuracy.

Despite their success in several building semantic segmentation tasks, the discussions on building extraction via multi-source remote sensing imagery of which the spatial resolution differs are quite inadequate. With the dramatically increasing availability of new large-scale remote sensing data sources, the ever-expanding choices of datasets can be utilized in semantic segmentation tasks [29]–[31], and the case that training and testing datasets obtained from multiple sources with different resolution would be inevitable and ubiquitous in many practical applications [32].

In general, differences between the resolution of the training and testing datasets would greatly influence building semantic segmentation. Three factors are mainly responsible for the problems in this regard. First, the resolution defines the ability of a single pixel to cover the Earth's surface, which would cause the same building to appear to have a different size in multiple remote sensing images of different resolution. A recent study [33] indicated that the factor of building size strongly impacts upon the capability of the DCNN model, and a model trained by using a building of a specific size would find it difficult to detect buildings of a significantly different size. Second, the resolution indicates the ability of the image to represent small objects. Thus, a small-sized land feature would be deformed or ignored in a low-resolution (LR) image due to the limited resolution. Many studies regarding small object detection [34]–[36] demonstrated the difficulty of solving this problem. Furthermore, as an important indicator, resolution measures the richness of information contained in remote-sensing imagery [37], in which a different resolution represents a different frequency information distribution, which greatly affects the features of the building such as its color, outline, and texture [38]. For the aforementioned reasons, a DCNN model trained at a specific resolution would find it fundamentally difficult to correctly represent the features of the testing dataset at another resolution, and this would result in a poor generalization of semantic segmentation. Thus, overcoming the constraint of resolution differences among multi-source remote sensing imagery would facilitate the development of building semantic segmentation to a considerable extent.

To deal with the severe problems caused by resolution difference between multi-source remote sensing imagery, the solution can be mainly classified into image transform based [39], data augmentation based [40], and transfer learning based [41] methods. With regard to image transform, the usual approach would be to downscale the high-resolution (HR) imagery into LR space by using downsampling methods [42] or to upscale LR imagery to HR space using a single filter such as bicubic interpolation [43]. The relevant drawbacks are obvious since downsampling would lead to undesired side-effects such as the loss of spatial information whereas interpolation would generate insufficient large gradients along edges and high-frequency regions by simply weighted averaging neighboring LR pixel values [44], small buildings would not be the same as larger ones even if up-scaled. With regard to data augmentation, methods such as color transformation, affine transformation, rotation, and linear scaling could enrich the variety of the training dataset, but could not supplement important features such as high-frequency information effectively at LR. And about transfer learning, although it owns the capability to rebuild the model based on utilizing the knowledge acquired from the previous task, once the feature-space and information distribution changes caused by resolution, the preparation for adequate amount of new training dataset is still unavoidable, which limits the efficiency and scalability in practical applications.



(a) Training data examples (b) Study area (c) Testing data examples

FIGURE 1. Materials. (a) and (c) show examples of the training and testing data, including high-resolution aerial imagery in which the corresponding buildings are annotated in purple and low-resolution satellite imagery with the relevant buildings annotated in green. (b) The study area divided into training and testing areas colored purple and green, respectively.

Given the difficulties faced by the methods mentioned above, super-resolution (SR) [45] has emerged as a promising alternative strategy to solve the problem. Aimed at increasing the image resolution while providing finer spatial details than those captured by the original acquisition sensors, SR could balance the size and detail of land features between the training and testing datasets to a certain degree [46]. In addition, as a highly ill-posed problem, SR operation is considered to be a one-to-many mapping from LR to HR space, which can have multiple solutions. Recent studies on DCNN-based SR models have shown tremendous capability in super-resolving an LR image into HR space, showing that generating high-quality SR remote-sensing imagery is achievable. A detailed review of additional DCNN-based SR models and their corresponding applications was recently published [47].

In this study, contrary to previous work, we propose to integrate super-resolution (SR) techniques into the existing segmentation framework to address the problem of building semantic segmentation in multi-source remote sensing imagery with different spatial resolution. To validate the feasibility of the proposed method, two high-performance DCNN-based models, namely efficient sub-pixel convolutional neural network (ESPCN) [48] and UNet, are adopted to perform SR and the semantic segmentation operation, respectively. In addition, three-band RGB HR aerial imagery and single-band grayscale LR panchromatic satellite imagery are selected as representative multi-source remote sensing imagery to conduct training and testing, respectively. It is worth emphasizing that, to the best of our knowledge, there has not been any empirical study using SR techniques for the building semantic segmentation from multi-source imagery with different resolution.

The main contributions of this study are three fold:

- We discussed the challenge and limitation of recent deep learning based studies on building semantic segmentation of building while under multi-source imagery with different resolution circumstance.
- We innovatively presented a novel SR integrated building semantic segmentation framework to tackle the problem caused by the unaligned resolution between training and testing data, and investigated the feasibility of the proposed method based on comprehensive experiments.
- The experimental results demonstrate the proposed method could achieve state-of-the-art performance, and the IoU and Kappa is approximately 19.01% and 19.10% higher than that of the method without SR, respectively. It indicates the effects of SR on segmentation performance in remote sensing imagery, which would benefit the remote sensing community from literature review to future directions.

The remainder of this paper is organized as follows. Section II introduces the area we studied and the data source. Then, the workflow of the proposed method is explained in Section III, where details of the algorithms as well as the evaluation metrics are also presented. After that, the experimental results and discussion appear in Section IV and V. Finally, the conclusions are drawn in Sections VI.

II. DATA

A. STUDY AREA

As one of the world's highest density urban areas, Tokyo contains intensely dense buildings with a huge diversity and complexity. Such characteristics of urban landscape lead spatial resolution to play an important role in semantic segmentation task. In this study, we deliberately selected some representative study areas in downtown Tokyo to demonstrate the feasibility of SR in building semantic segmentation. Figure 1b shows the detailed study area. We divided the entire area into training and testing areas indicated in purple and green, respectively. The training area covered $33km^2$ and is mainly located in the Setagaya, Koto, and Sumida districts, which include a wide variety of land use categories such as residential, commercial, and industrial areas. In addition, an area of $3km^2$ in the Koto district with comprehensive land use was selected to perform testing.

B. DATA SOURCE

The aerial and panchromatic satellite imagery were used as the training and testing datasets, respectively. The remote

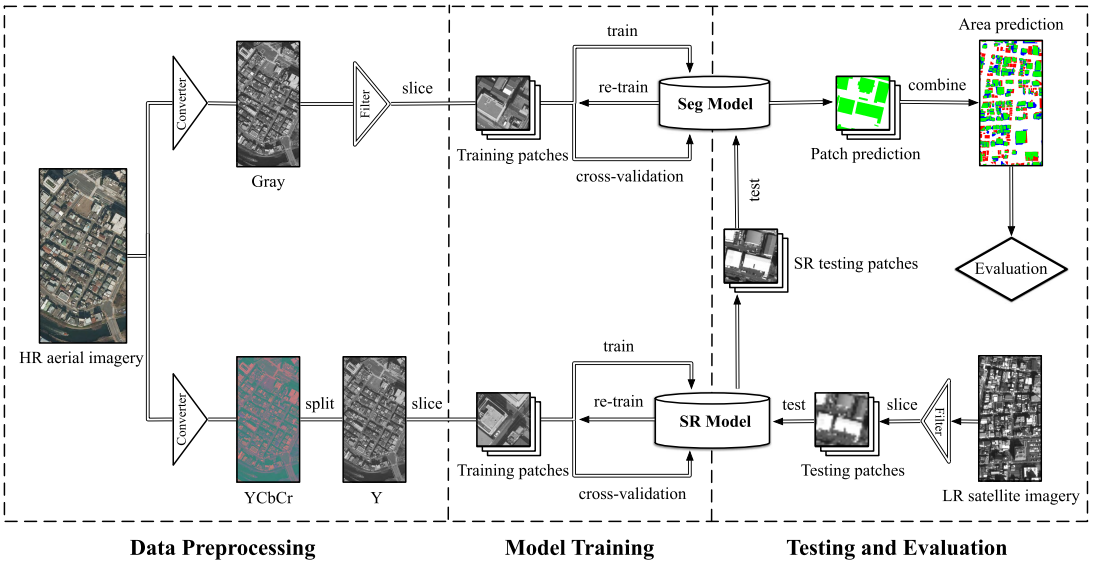


FIGURE 2. Framework of our building semantic segmentation method.

top-view three-band RGB aerial imagery in the training area was acquired in March 2016 with a resolution of 0.160*m*, and the source panchromatic imagery in the testing area was captured by the WorldView-2 sensor in May 2016 with spatial and radiometric resolution of 0.500*m* and 16-bit, respectively. In terms of the annotated dataset, a total of approximately 60,000 and 3,000 building footprints are contained within the training and testing areas, respectively. To best represent the building footprints, a polygon-based method via QGIS was used to conduct the annotation, in which, the polygon maximizes the shape of a building from an orthophoto, and any adjoining buildings are marked as a single building. Owing to the limitations of interpretation based on human-based vision, a few small errors are inevitable especially for high-density areas in LR satellite imagery. Some examples of training and testing imagery and their corresponding annotations are shown in Figure 1a in purple and in Figure 1c in green, respectively.

III. METHODS

In this section, we present our novel framework for an SR integrated building semantic segmentation method. As shown in Figure 2, the three main procedures in the framework are: data processing, model training, and testing with related evaluation. The two processes that precede the testing stage can be considered as running in parallel in terms of both segmentation and SR model generalization. First, aerial imagery of the study area obtained from the same source undergoes parallel data preprocessing to generate training data for semantic segmentation and SR integration. Subsequently, the obtained data is fed into the proposed upper UNet and lower ESPCN to train the segmentation and SR model, respectively. Here, in both models, 70% of the training data is used for training, and the remaining 30% is used for cross-validation. To evaluate the quality of the segmentation model, we apply six commonly used evaluation metrics that include precision, recall, overall accuracy [49], F1-score [50], the kappa coefficient [51], and the Jaccard index or intersection over union (IoU) [52]. The SR model is assessed by using the peak signal-to-noise ratio (PSNR) [53], which is usually taken as an approximation to human perception of reconstruction quality. It should be noted that both segmentation and SR models would be retained in case the bad results are generated when conducting cross validation. After that, in the testing and evaluation procedure, we first input the processed LR satellite data into the trained SR model to generate related upscaled SR data; then, the trained segmentation model with proper hyperparameters is adopted to enable the generated testing SR satellite data to be used to make predictions. Finally, the quality of the semantic segmentation results is evaluated by the segmentation assessment criteria mentioned above. To clearly reflect the capability of different models, here, the evaluation metrics are calculated without any post-processing for both the semantic segmentation and SR processes.

This section details first the data preprocessing step, followed by the training strategies of the SR and segmentation models. Lastly, the testing method and related assessment criteria are proposed and explained.

A. DATA PREPROCESSING

Data preprocessing is conducted in parallel to generate training data for both the segmentation and SR models. With respect to the segmentation, the three-band RGB HR aerial imagery is first converted into grayscale to align it with the single-band panchromatic testing LR satellite imagery; then, after applying basic color normalization methods such as adaptive histogram equalization [54], the aerial imagery

is sliced into patches sized 224 × 224 pixels by using a random sliding window to generate data for model training and cross-validation purposes. Simultaneously, corresponding ground truth patches with consistent size are generated via the annotation dataset as well. In terms of the SR process, considering humans are more sensitive to luminance changes [55], we convert the aerial imagery from RGB into YCbCr color space, and only take the luminance channel in the YCbCr color space into consideration. Similar to the process of segmentation, the converted aerial imagery in the luminance channel is sliced into small patches sized 224 × 224 pixels. In addition, to reduce the influence caused by data availability, data augmentation techniques [56] are also adopted to enrich the training data for both the segmentation and SR processes.

B. SEGMENTATION MODEL

Several effective segmentation models have been introduced in Section I, to demonstrate the feasibility of proposed framework, in this study, we propose to adopt UNet architecture as a representative segmentation model to conduct building semantic segmentation.

UNet is one of a state-of-the-art models for image semantic segmentation, and has been successfully applied to perform different tasks with high accuracy and efficiency. The network architecture can be divided into two parts: a contracting path and a symmetric expansive path. The contracting path, which is regarded as a variant of VGG [57], contains five consecutive blocks for feature extraction and downsampling. Each of these blocks consists of two 3×3 unpadded convolutions followed by 2×2 max pooling, which provides the abstracted form of the representation while enlarging the receptive field. The expansive path, which can be considered as the reverse operation of the contraction path, comprises four blocks and each contains an upsampling of the feature map followed by a 2×2 convolution. Importantly, before feeding the extracted feature map into the next block, the feature map generated in the contraction path with the same shape is integrated inside by concatenation. In addition, the number of feature channels are doubled and divided in half after each downsampling and upsampling, respectively. The non-saturated activator known as a rectified linear unit (ReLU) is adopted after each convolutional operation to perform nonlinear mapping. This architecture makes UNet suitable for mining very deep and abstract features.

Considering the characteristics of UNet, some advantages of adopting UNet architecture as segmentation model to conduct building semantic segmentation can be listed as follows. First, the architecture of UNet performs pixel-to-pixel and end-to-end mapping from input to output, which enables precise localization for the building segmentation result. Second, UNet can generate results in HR space by recovering HR representations. Instead of using pooling operators after successive convolutional layers, the architecture adopts upsampling with a large number of feature channels to increase the output resolution. In addition, the model has the capability to augment feature space by fusing the context from imagery acquired at different resolutions. Because HR features extracted from a contracting path and LR features upsampled by using an expansive path are combined through the process of concatenation, the feature space can be augmented to a certain degree.

In this study, we modified the original UNet architecture in some important ways. To avoid dead neurons in the back-propagation step as well as to benefit from initialization, we use leaky ReLU [58] instead of ReLU after each convolution. Concretely, the convolution operation which performs element-wise multiplication via kernels, can be formulated as follows:

$$z = \sum_{i=1}^{h_f} \sum_{j=1}^{w_f} \sum_{d=1}^{c_l} \Theta_{i,j,d,d'} \times x_{i,j,d} + b_{d'} \tag{1}$$

where h_f, w_f represent the height and width of the kernel Θ, c_l is the number of channels for input x in layer l, and b in shape $1 \times 1 \times 1 \times d'$ donates the bias.

Then, leaky ReLU ϕ is utlized to generate the hypothesis from z:

$$\phi(z) = \begin{cases} z & if\ z > 0 \\ 0.01z & otherwise \end{cases} \tag{2}$$

Subsequently, batch normalization [59] is also added and extensively applied after each non-linearity to accelerate the training and reduce internal covariate shift. The two parameters in batch normalization, scale γ and shift β, can be learned by:

$$Y_B = \gamma \frac{X_B - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta \tag{3}$$

where X_B and Y_B denote all input and output in mini-batch B. μ_B and σ_B^2 refer to mean and variance of corresponding mini-batch.

Furthermore, to avoid over-fitting, we eliminate the redundant features by adopting dropout [60], and the final binary classification of either building or non-building is predicted by using the sigmoid function. Here, the cross entropy expressed by Equation 4 is used to penalize the inconsistency between prediction $\hat{Y}$ and ground truth Y. Further, H and W are the height and width of both the prediction and ground truth, respectively.

$$L(Y, \hat{Y}) = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \Big(Y_{i,j} \times \log(\hat{Y}_{i,j}) + (1 - Y_{i,j}) \times \log(1 - \hat{Y}_{i,j}) \Big) \tag{4}$$

C. SR MODEL

Aimed at recovering HR imagery from its LR information, SR is an important category of techniques for image processing and offers an excellent opportunity to facilitate the development of different remote sensing applications including building semantic segmentation. In recent years,

deep learning based SR methods have been investigated quite intensively and have achieved state-of-the-art performance among various benchmarks of SR, some breakthrough studies such as SRCNN [61], VDSR [62], LapSRN [63], SRGAN [64], etc. To explore the feasibility of integrating SR into the building semantic segmentation task while simultaneously considering the characteristics of the available data source, we propose to adopt a typical deep learning based single-image super-resolution (SISR) method named ESPCN to increase the resolution of LR panchromatic satellite imagery to match that of the HR aerial imagery at some point. Ideally, the reconstructed SR imagery could be augmented with high-frequency information on condition that the spatial resolution is similar to that of the HR aerial imagery.

Instead of upscaling the LR input imagery X^{LR} into HR space before reconstruction, ESPCN directly extracts feature maps from LR space with the help of successive hidden convolutional layers. To generate SR imagery X^{SR} from X^{LR} with an upscaling factor r, X^{LR} with a shape of $h \times w \times c$ would undergo L layers of convolution operations. The first $L-1$ layers can be described as:

$$f^1(X^{LR}, \Theta_1, b_1) = g(\Theta_1 \times X^{LR} \times b_1) \quad (5)$$

$$f^l(X^{LR}, \Theta_{1:l}, b_{1:l}) = g(\Theta_l \times f^{l-1}(X^{LR}) \times b_l) \quad (6)$$

where Θ_l and b_l with $l \in (1, L-1)$ represent the learnable hyperparameters weights and biases, respectively. Function g is the activator ReLU used to perform nonlinear mapping.

The final layer f^L applies an efficient sub-pixel convolution operation, which learns an array of complex upscaling filters to upscale the LR feature maps into the HR output X^{SR}. The formula as follows:

$$X^{SR} = f^L(X^{LR}, \Theta_L, b_L) = PS(\Theta_L \times f^{L-1}(X^{LR}) + b_L) \quad (7)$$

where PS is a periodic shuffling operator that reshapes the feature maps of layer $L-1$ from shape $h \times w \times c \cdot r^2$ into a tensor of shape $rh \times rw \times c$. Weights Θ_L are in the shape $h_f \times w_f \times c_{L-1} \times c \cdot r^2$.

During training, the input LR imagery X^{LR} can be synthesized efficiently by sub-sampling HR aerial imagery X^{HR} from shape $rh \times rw \times c$ to $h \times w \times c$ using a Gaussian filter. After generating the result in each epoch, the loss function pixel-wise mean squared error (MSE) (Equation 8) is used to measure the discrepancy between reconstructed X^{SR} and original X^{HR}, both in shape $rh \times rw \times c$. In addition, early stopping is adopted to end the training process once the model performance no longer improves after 100 epochs on cross-validation data.

$$L(X^{HR}, X^{SR}) = \frac{1}{r^2 hw} \sum_{i=1}^{rh} \sum_{j=1}^{rw} (X_{i,j}^{HR} - f_{i,j}^{L}(X^{LR}))^2 \quad (8)$$

In terms of the spatial resolution of the multi-source remote sensing imagery used in this study, the HR aerial imagery is approximately three times higher than LR panchromatic imagery; therefore, three SR models are trained by ESPCN by assigning the values 1, 2, and 3 to the upscaling factor r, respectively.

D. ASSESSMENT CRITERIA

The quality of the results obtained after semantic segmentation and the use of the SR model is evaluated by applying criteria based on a confusion matrix and image quality assessment (IQA), respectively.

We assess the properties of the resulting segmentation $\hat{Y}$ with regard to the ground truth Y via six criteria: precision, recall, overall accuracy, F1-score, the kappa coefficient, and the Jaccard index. For the sake of simplicity, tp, fn, fp, and tn, represent the basic terms in the confusion matrix: true positive, false negative, false positive, and true negative, respectively.

Precision and recall are both measures of relevance. Here, Precision (Equation 9) measures the proportion of relevant results in the list of all returned search results, and refers to the percentage of correctly predicted buildings to the total number of predicted buildings.

$$Precision = \frac{tp}{tp + fp} \quad (9)$$

Contrary to this, recall (Equation 10) measures the proportion of the relevant results returned by the segmentation model to the total number of relevant results that could have been returned, and refers to the correctly predicted buildings as a percentage of the exact total number of buildings.

$$Recall = \frac{tp}{tp + fn} \quad (10)$$

A trade-off between precision and recall is important. Thus, the F1-score, which takes both precision and recall into account and finds an optimal blend for them, is applied. The formula is as follows:

$$F_1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

where the relative contribution of precision and recall to the F1 score are the same.

Overall accuracy, as shown in Equation 12, is also an essential metric in semantic segmentation. It refers to the proportion of correctly predicted building and non-building areas of the total number of areas to predict.

$$Overall\ Acc = \frac{tp + tn}{tp + tn + fp + fn} \quad (12)$$

To measure the level of agreement between two objective annotators, kappa coefficient is also applied as follows:

$$Po = Overall\ Acc \quad (13)$$

$$Pe = \frac{(tp + fp) \times (tp + fn) + (fn + tn) \times (fp + tn)}{(tp + tn + fp + fn)^2} \quad (14)$$

$$Kappa = \frac{Po - Pe}{1 - Pe} \quad (15)$$

where Po is identical to the overall accuracy and refers to the observed agreement ratio, and Pe is the probability of

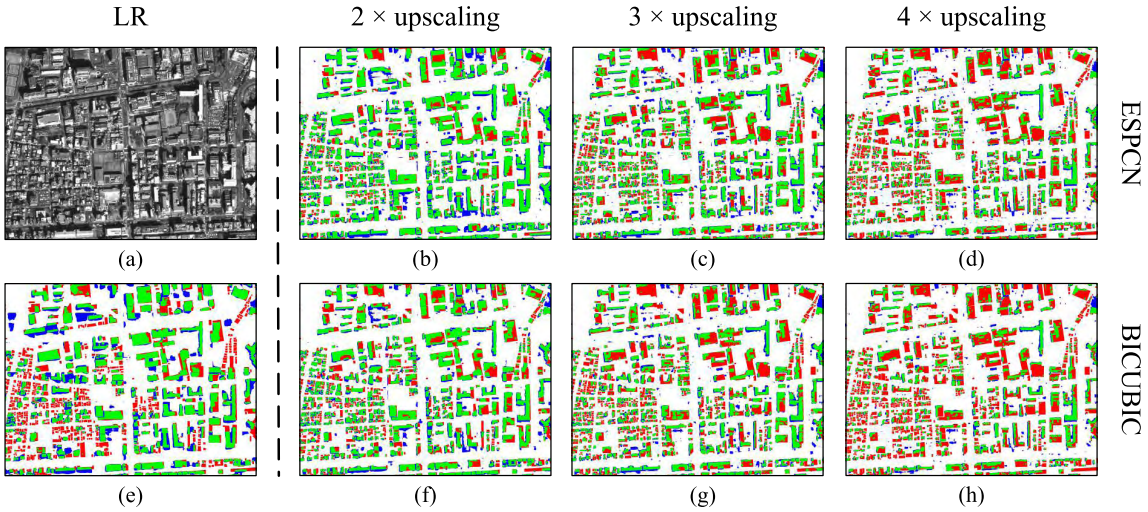


FIGURE 3. Qualitative results for test region 1.

the expected agreement when both annotators assign building areas randomly.

Moreover, as the most prevalent criterion for segmentation problems, the Jaccard index in Equation 16 is used to measure the dissimilarity between the predicted and extracted building areas.

$$Jaccard = \frac{tp}{tp + fp + fn} \tag{16}$$

All six of the segmentation assessment criteria mentioned above reach their best value at 1 and worst score at 0.

Regarding the qualitative performance of the SR model, the most widely used evaluation criterion, PSNR, is adopted to measure the reconstruction quality of transformation. The PSNR, which is an objective IQA method, is calculated based on the maximum possible pixel value (denoted as MAX) and the pixel-level MSE between HR imagery X^{HR} and super-resolved SR imagery X^{SR}. The corresponding formulas are as follows:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X^{HR} - X^{SR})^2 \tag{17}$$

$$PSNR = 10 \times \log_{10}(\frac{MAX^2}{MSE}) \tag{18}$$

We normalize the maximum possible pixel value between multi-source imagery by converting the value of 8-bit aerial imagery and 16-bit panchromatic imagery, and rescale both of them from 0 to 1. Thus, instead of relying on human visual perception, the quality of SR is computationally only related to the MSE.

IV. RESULTS

To demonstrate the feasibility of proposed SR integrated approach, we employ the same modified UNet model trained by HR aerial imagery as the backbone to test imagery in three main categories: LR, ESPCN based SR, and bicubic based interpolated imagery. Considering the exact resolution of HR aerial and LR panchromatic imagery as well as exploring the influence caused by their resolution difference, we upscale the testing LR panchromatic imagery with resolution $0.500m$ into 2, 3, 4 times by both ESPCN and bicubic interpolation methods. Thus, SR- and bicubic-based interpolated panchromatic imagery with resolution $0.250m$, $0.167m$, and $0.125m$ are generated. Moreover, to evaluate the robustness of methods, we deliberately divide the entire testing area into four regions based on land use, where buildings and other important land features present in different characteristics in terms of grayscale value, texture, structure, density, size, etc.

This section presents the qualitative and quantitative results of the building semantic segmentation of the four regions via different methods. More specifically, with respect to the qualitative results, the assessment criteria introduced in Section III-D are applied. The quantitative results are shown in Figure 3, 5, 7, and 9. In these figures, (**a**) and (**e**) are LR panchromatic imagery and the corresponding segmentation result, (**b**), (**c**), and (**d**) are the segmentation results generated by the ESPCN-based methods with upscale factors of 2, 3, and 4, respectively. Further, (**f**), (**g**), and (**h**) are the segmentation results generated by the bicubic-based methods with upscale factors corresponding to the ESPCN-based methods. The different colors: green, red, blue, and white, are used to indicate the tp, fn, fp, and tn pixels in the segmentation results, respectively. Moreover, for improved visualization, as shown in Figure 4, 6, 8, and 10, enlargements of selected representative subregions in each region are displayed in a yellow window to reveal the details, which reflect the effect of applying different methods.

Figure 3 shows the qualitative results for test region 1, which mainly contains commercial and residential areas, in which the types of buildings are particularly diverse, whereas the non-building areas include several open sided

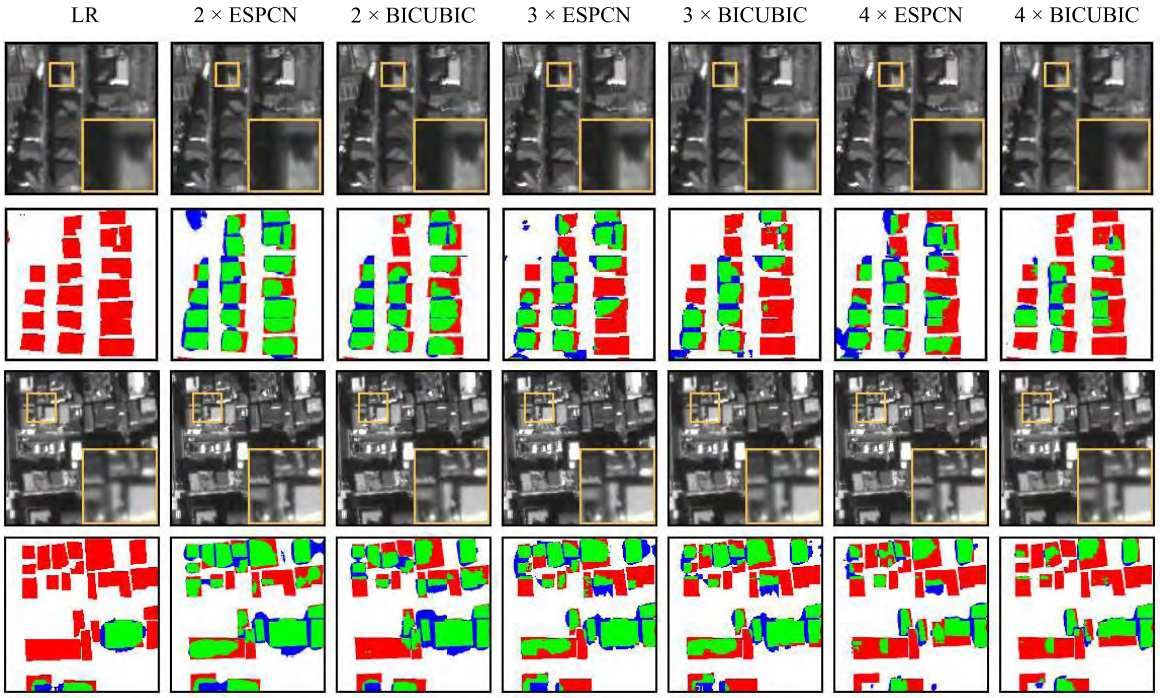


FIGURE 4. Qualitative results for representative subregions in test region 1.

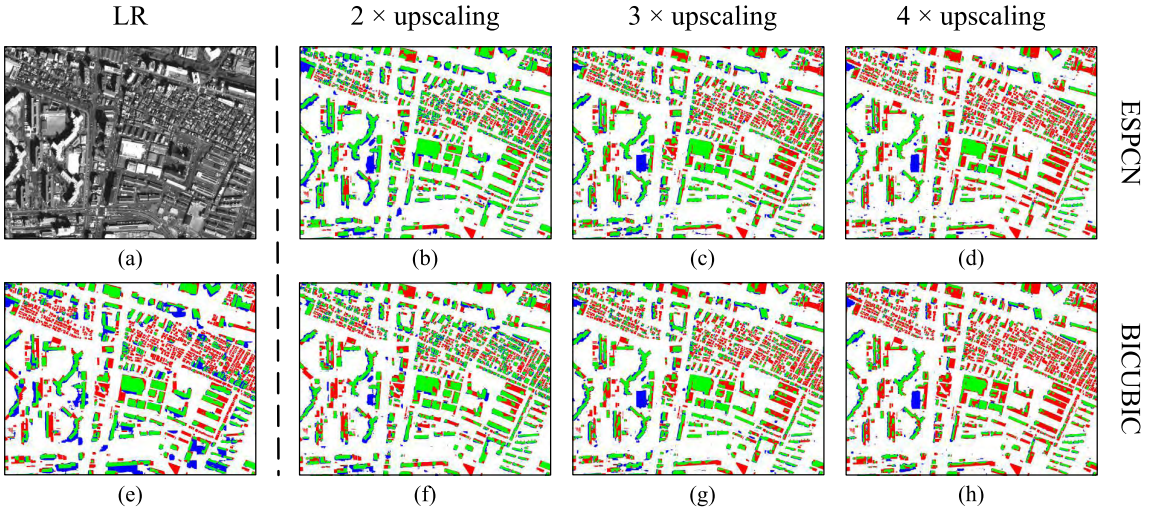


FIGURE 5. Qualitative results for test region 2.

car parks and sports grounds. The corresponding quantitative results generated by the different methods are provided in Table 1, and indicate that the proposed ESPCN method with an upscale factor of 2 outperformed other models in terms of the recall, overall accuracy, F1-score, kappa, and Jaccard index. With respect to precision, the results are worse than those of other upscaled imagery but are still more accurate than those obtained for the original LR imagery.

Notably, as shown in the first row of Figure 4, in some residential areas, the size of building as well as the separation distance between adjacent buildings is quite small in LR imagery, which considerably increases the challenge of segmentation, and makes it difficult to identify buildings at all. The use of ESPCN not only enlarges the size of building and distance between adjacent buildings like bicubic interpolation does but also enrich the texture information. The effect can easily be seen in the enlarged views and related segmentation results, where all buildings are well segmented by adopting ESPCN with an upscale factor of 2, and, ESPCN outperforms the simple interpolation methods for every respective upscale factor. Similar to the residential areas, as shown in the third row of Figure 4, the external outlines of

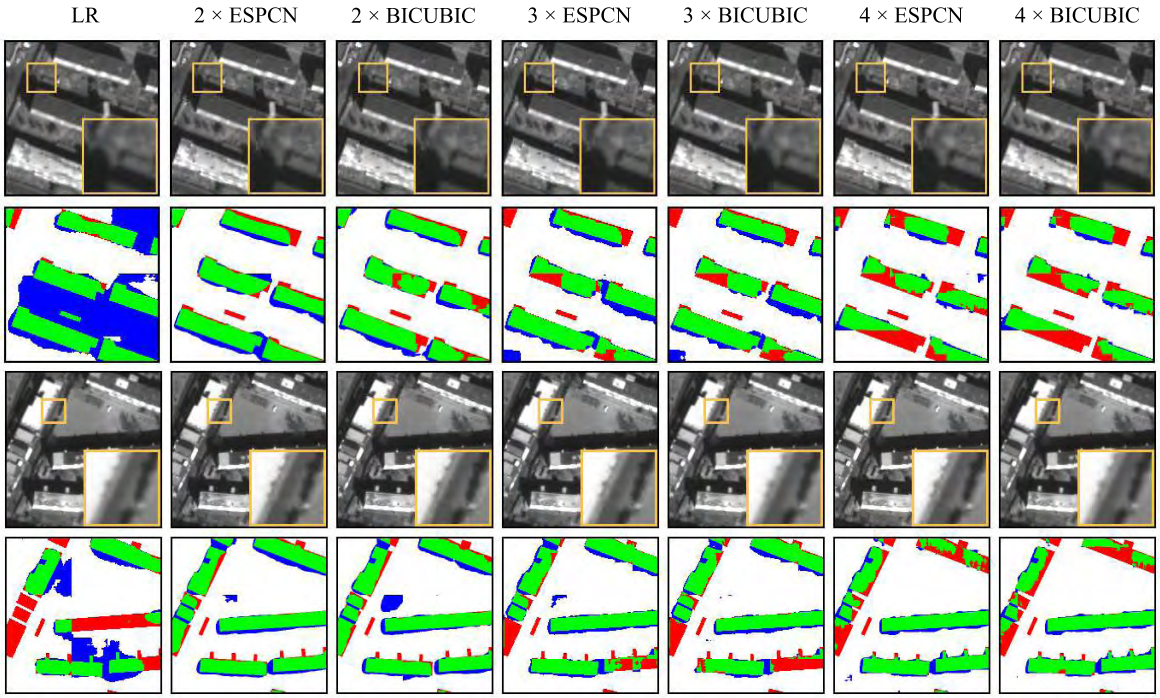


FIGURE 6. Qualitative results for representative subregions in test region 2.

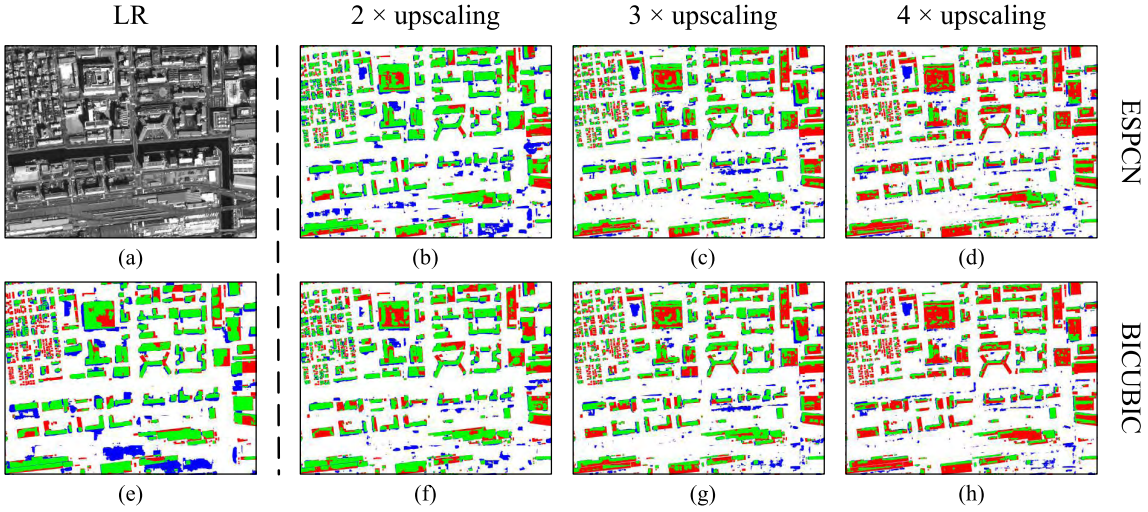


FIGURE 7. Qualitative results for test region 3.

buildings in the commercial area are particularly clear when using ESPCN, which produces more accurate segmentation results.

Region 2 is a mainly residential area, and the related qualitative results are shown in Figure 5. As with the results of region 1, because of the misalignment in resolution between the training and testing data, the majority of small and detached houses in LR imagery are misclassified as non-building areas. Apart from the prevalence of detached houses, the residential buildings in region 2 also include medium-rise mansions and apartments with a comparatively larger distance separating them, and the region also contains several small parks. Both the qualitative and quantitative results shown in Figure 5 and Table 2 demonstrate the effect of ESPCN on the semantic segmentation of residential buildings in the different categories.

Figure 6 shows some representative mansions and apartments as well as related segmentation details. Except for a few tiny accessory buildings and protruding architectural contours, buildings are correctly segmented with a low *fp* value by adopting ESPCN with an upscale factor of 2. In contrast, the use of LR imagery leads to the misclassification of many

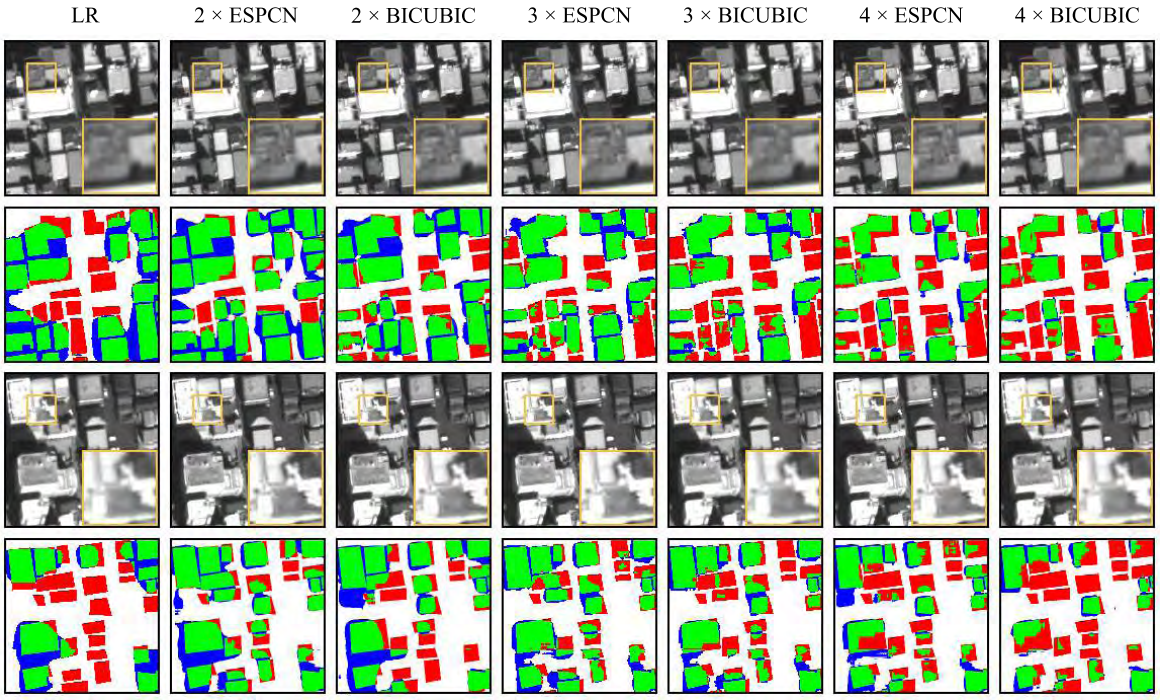


FIGURE 8. Qualitative results for representative subregions in test region 3.

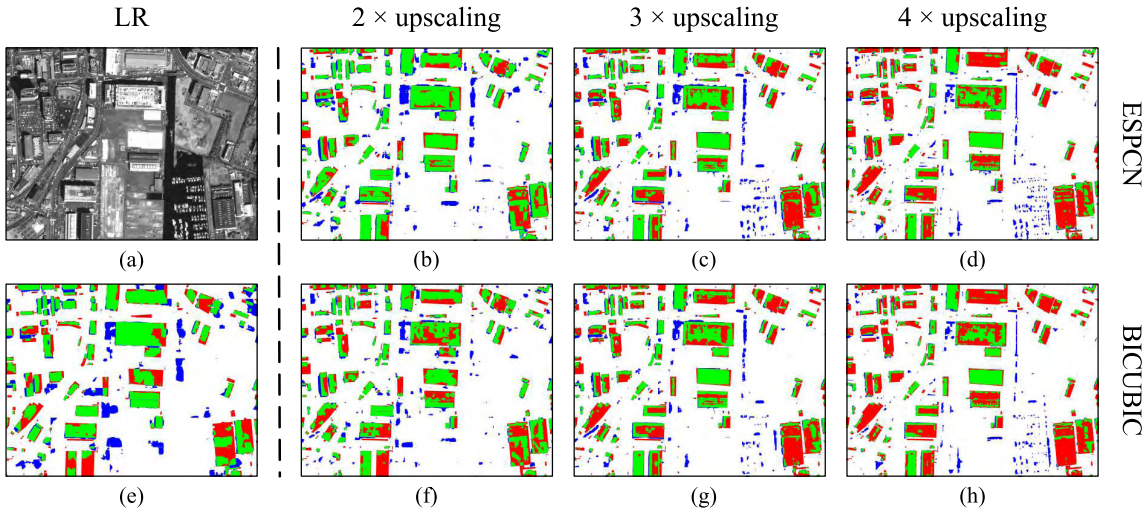


FIGURE 9. Qualitative results for test region 4.

roads and areas containing vegetation as buildings, whereas buildings are incorrectly detected.

As shown in Figure 7a, region 3 mainly consists of quasi-industrial zones occupied by light industrial and service facilities, with non-building land features such as a river and large-scale transport system also included. Intuitively, the qualitative results seem to suggest that ESCPN with an upscale factor of 2 outperformed LR and the other methods with fewer *fp* and *fn* results, especially in areas bordering the railway line and high-density building areas. Some representative results are presented in Figure 8.

The quantitative results in Table 3 also infer that SR imagery obtained with an appropriate upscale factor can achieve performance superior to that attainable with LR imagery in regions with comprehensive land features.

Region 4 shown in Figure 9a is situated in the vicinity of the Tokyo Bay estuary. This highly particular location consists of industrial areas with large factories and storage buildings as well as docks spread over the entire region. Land features particular to this location, such as containers, are widely distributed in the port, while barges are moored in the harbor. The quantitative results shown in Figure 9b to h and

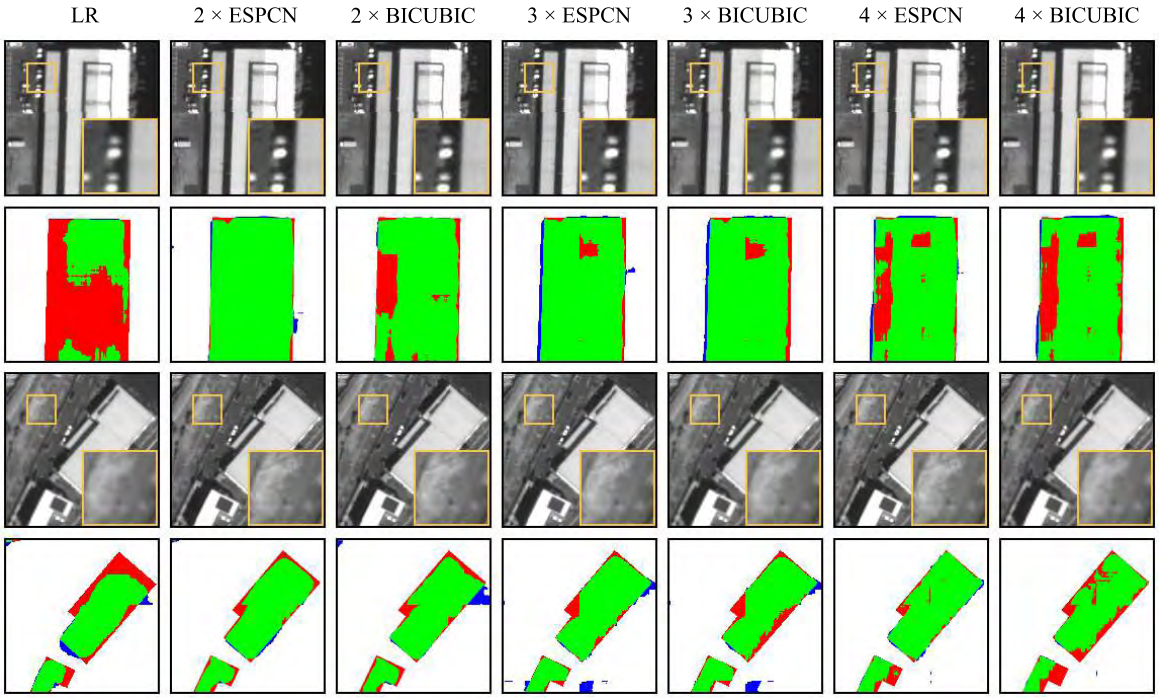


FIGURE 10. Qualitative results for representative subregions in test region 4.

TABLE 1. Quantitative results for test region 1.

Model	Scale	Resolution	Precision	Recall	Overall Acc	F1-score	Kappa	Jaccard
LR	1	0.500	0.693	0.621	0.803	0.655	0.518	0.487
ESPCN	2	0.250	0.728	**0.772**	**0.844**	**0.749**	**0.637**	**0.599**
BICUBIC	2	0.250	0.752	0.645	0.829	0.694	0.576	0.532
ESPCN	3	0.167	0.737	0.606	0.816	0.665	0.540	0.499
BICUBIC	3	0.167	0.749	0.527	0.804	0.619	0.492	0.448
ESPCN	4	0.125	0.735	0.462	0.788	0.568	0.436	0.396
BICUBIC	4	0.125	**0.755**	0.388	0.777	0.512	0.387	0.344

TABLE 2. Quantitative results for test region 2.

Model	Scale	Resolution	Precision	Recall	Overall Acc	F1-score	Kappa	Jaccard
LR	1	0.500	0.645	0.514	0.787	0.572	0.432	0.400
ESPCN	2	0.250	0.727	**0.680**	**0.840**	**0.703**	**0.594**	**0.542**
BICUBIC	2	0.250	0.748	0.577	0.828	0.651	0.540	0.483
ESPCN	3	0.167	0.749	0.560	0.826	0.641	0.529	0.472
BICUBIC	3	0.167	0.756	0.509	0.818	0.608	0.495	0.437
ESPCN	4	0.125	0.751	0.406	0.798	0.527	0.413	0.358
BICUBIC	4	0.125	**0.765**	0.373	0.794	0.501	0.390	0.335

TABLE 3. Quantitative results for test region 3.

Model	Scale	Resolution	Precision	Recall	Overall Acc	F1-score	Kappa	Jaccard
LR	1	0.500	0.689	0.699	0.827	0.694	0.573	0.531
ESPCN	2	0.250	0.715	**0.749**	**0.846**	**0.732**	**0.624**	**0.577**
BICUBIC	2	0.250	**0.769**	0.640	0.845	0.699	0.595	0.537
ESPCN	3	0.167	0.741	0.609	0.830	0.668	0.556	0.502
BICUBIC	3	0.167	0.747	0.541	0.819	0.627	0.512	0.457
ESPCN	4	0.125	0.712	0.447	0.794	0.550	0.425	0.379
BICUBIC	4	0.125	0.731	0.380	0.786	0.500	0.381	0.333

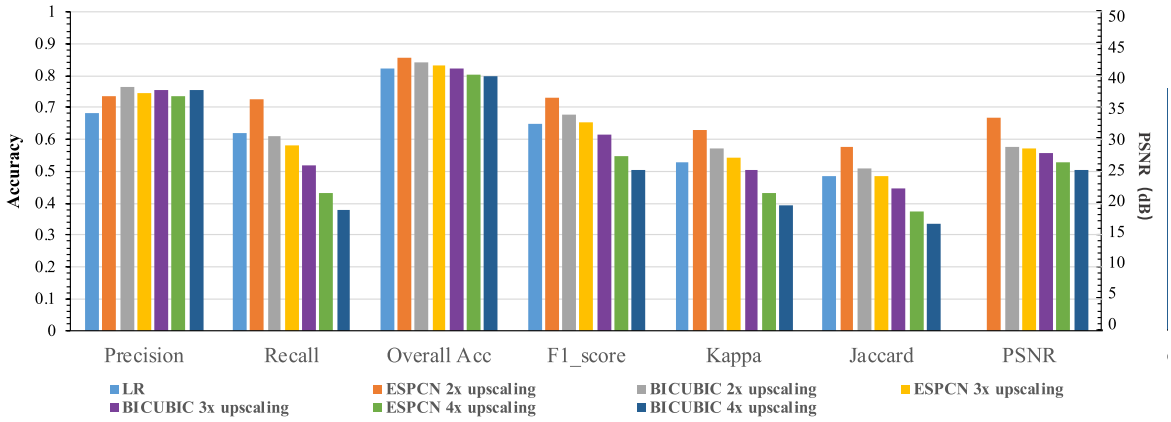
indicate that ESPCN with an upscale factor of 2 can segment large buildings with the lowest *fn*.

The impact of the resolution on the segmentation of large buildings was analyzed in greater detail by selecting a few representative large buildings with a simple roof texture and that are surrounded by wide open areas for comparison purposes. As shown in Figure 10, although the size of large buildings in LR imagery is comparable with that of small buildings

TABLE 4. Quantitative results for test region 4.

Model	Scale	Resolution	Precision	Recall	Overall Acc	F1-score	Kappa	Jaccard
LR	1	0.500	0.711	0.655	0.862	0.682	0.594	0.517
ESPCN	2	0.250	0.766	**0.711**	**0.886**	**0.738**	**0.665**	**0.584**
BICUBIC	2	0.250	**0.781**	0.569	0.867	0.659	0.578	0.491
ESPCN	3	0.167	0.752	0.551	0.858	0.636	0.55	0.466
BICUBIC	3	0.167	0.766	0.500	0.853	0.605	0.520	0.434
ESPCN	4	0.125	0.748	0.422	0.838	0.540	0.450	0.370
BICUBIC	4	0.125	0.766	0.369	0.832	0.498	0.412	0.332

(a)



(b)

Model	Scale	Resolution	PSNR	Precision	Recall	Overall Acc	F1-score	Kappa	Jaccard
LR	1	0.500	-	0.684	0.622	0.820	0.651	0.529	0.484
ESPCN	2	0.250	**33.353**	0.734	**0.728**	**0.854**	**0.730**	**0.630**	**0.576**
BICUBIC	2	0.250	28.880	**0.762**	0.608	0.842	0.676	0.572	0.511
ESPCN	3	0.167	28.471	0.745	0.582	0.832	0.652	0.544	0.485
BICUBIC	3	0.167	27.934	0.754	0.519	0.823	0.615	0.505	0.444
ESPCN	4	0.125	26.487	0.736	0.434	0.804	0.546	0.431	0.376
BICUBIC	4	0.125	25.209	0.754	0.378	0.797	0.503	0.392	0.336

FIGURE 11. Average performance of SR reconstruction and segmentation for the four test regions using different methods. (a) Bar diagram for performance comparison. The x- and y-axis represent the assessment criteria and corresponding values, respectively. (b) Table for performance comparison. For each assessment criterion, the highest values are highlighted in bold.

in the training HR imagery, the unclear contour of buildings in LR imagery is prone to misclassification and produces results with a large *fn* value. Such results reflect the importance of aligning the resolution between training and testing data from the side, as well as the effects of SR integrated method on semantic segmentation in satellite imagery.

The detailed results provided in Table 4 confirm the aforementioned conclusion. Large buildings in LR imagery can be detected by the model trained on HR imagery with relatively high accuracy; however, in contrast with the ESPCN integrated method, which contains high-frequency information, the performance remains poor.

V. DISCUSSION

Section IV presented comprehensive qualitative and quantitative results of the segmentation of buildings, which are located in various areas and which differ in terms of their density, shape, texture, size, and usage. The discussion we provide in this section aims to further demonstrate the feasibility of SR-integrated segmentation methods. First, the average quantitative results for the four regions are used to indicate the robustness of the proposed method. Then, we take reconstruction quality as a reference to show the relationship between segmentation and SR. It should be noted that since the HR satellite imagery is not available, we utilize the reconstruction quality generated in training procedure by HR aerial imagery to represent that of SR satellite imagery. Finally, selected qualitative results of important land features other than buildings are shown. Besides, poor results are briefly analyzed and discussed.

The average segmentation performance obtained by adopting different methods is shown in Figure 11. In general, the ESPCN-integrated method with an upscale factor of 2 significantly outperforms the other methods including LR imagery, improving the overall accuracy from 0.802 to 0.854, the F1-score from 0.651 to 0.730, kappa from 0.529 to 0.630, and the Jaccard index from 0.484 to 0.576. These quantitative results indicate that the model trained by HR imagery cannot

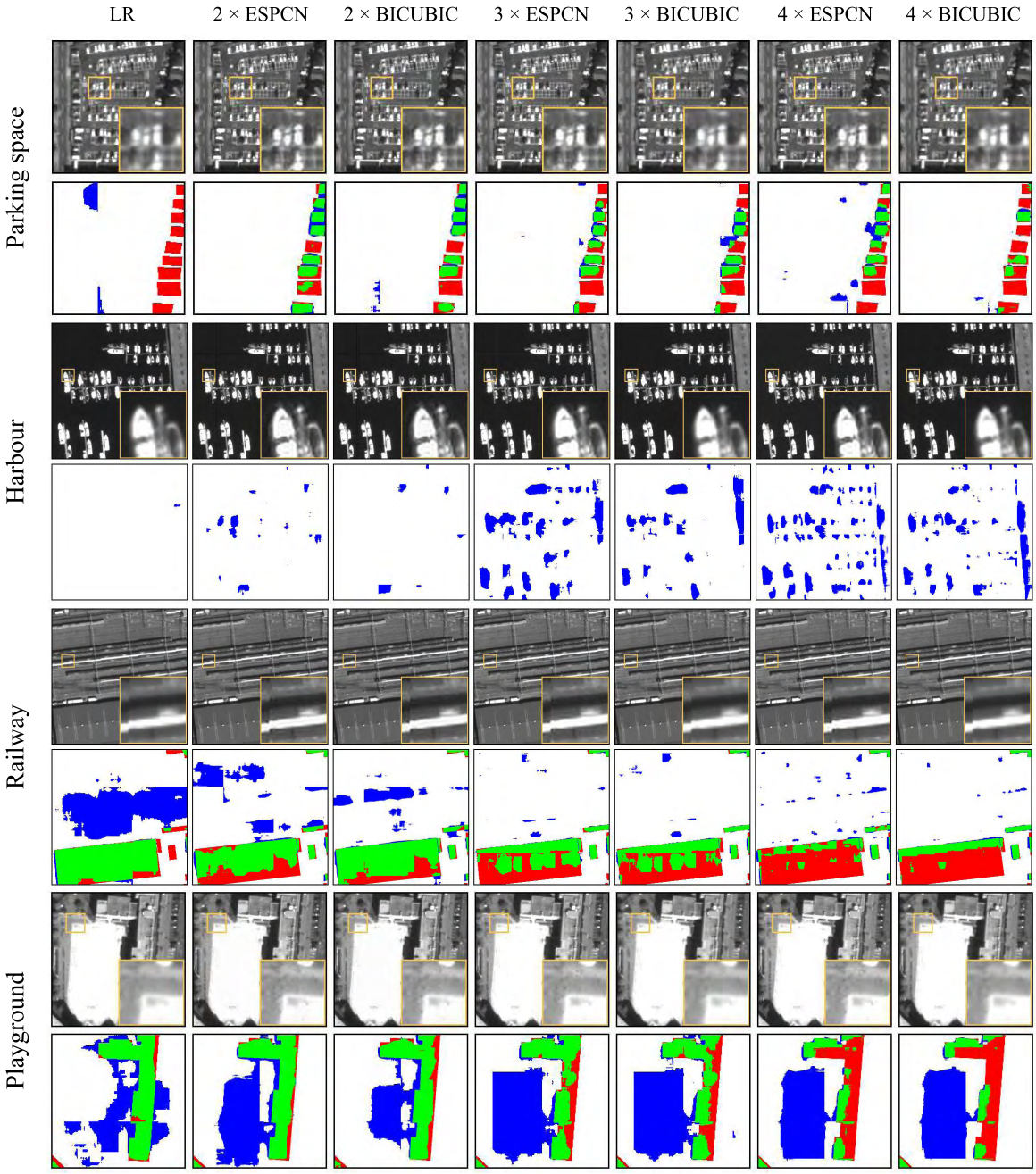


FIGURE 12. Results for important land features.

detect small buildings with high accuracy and that increasing the resolution of LR imagery could enlarge the building size by providing more pixels, which would align the size of buildings in the training data to a certain degree. This point of view would also be supported by the results generated via bicubic interpolation with an upscale factor of 2. Apart from increasing the image resolution, compared with simple interpolation, SR-based methods would also reconstruct finer spatial details with higher PSNR, which would yield improved segmentation results for the same upscale factor.

In principle, regarding the alignment of the resolution of HR with that of LR imagery, upscaling the resolution of LR imagery with a factor of 3 to $0.167m$ by ESPCN would match that of HR imagery to a great extent to generate the best segmentation results. However, because of the ill-posed problem, reconstructing high-quality SR imagery from LR space with a large upscale factor would be highly challenging. According to the IQA criteria shown in Figure 11b, an increase in the upscale factor from 2 to 4 causes the PSNR of ESPCN-based SR imagery to drastically decline

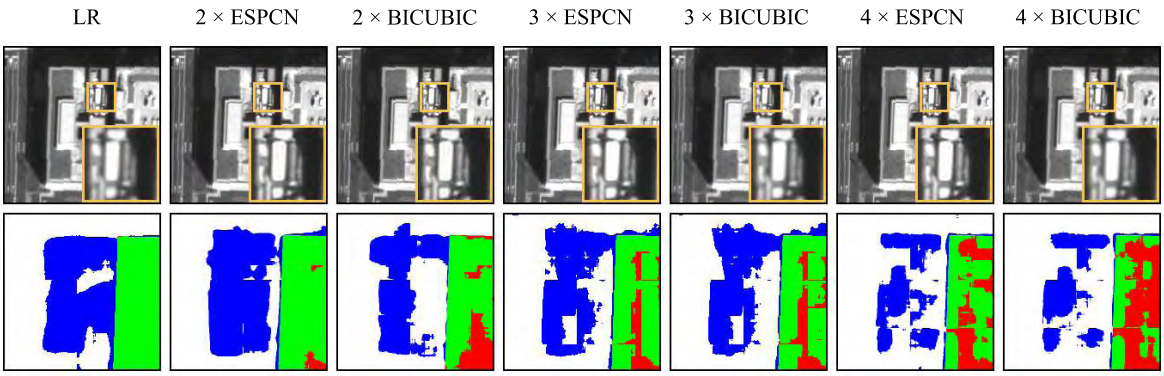


FIGURE 13. Bad results caused by annotation.

from 33.353 to 28.471 to 26.487. Although the resolution could match that of the training data, the reconstruction quality also severely impacts the correct representation of high-frequency information. Thus, low-quality SR imagery even with an appropriate resolution would worsen the segmentation performance. Ultimately, maintaining a balance between resolution and reconstruction quality is of great importance.

Figure 12 shows the semantic segmentation results of other representative land features. The first row shows a parking lot on which cars of different categories are distributed. As shown in the enlarged yellow window, by adopting SR, the shape and textural information of each car becomes much more refined. Because cars are common land features in HR aerial imagery, abundant training data for cars would enable the value of *fp* to be effectively decreased. In contrast, as shown in the third and fourth rows, the barges moored in the harbor are likely to be misclassified as buildings after the resolution is upscaled. This problem is caused by insufficient training samples for boats in HR aerial imagery. In terms of railways, trains and tracks are presented by a simple stripe-like feature, and increasing the resolution would enlarge the distance between adjacent strips to improve the performance. Finally, as shown in the last two rows, some polygon-like land features with simple textures such as playgrounds are prone to be misclassified as buildings at a different resolution, indicating that the problem is caused by the UNet model rather than the proposed SR integrated method.

Especially, Figure 13 shows a large region in which all methods misclassify non-building areas; after carefully analyzing the original LR image, we believe that the problem is caused by an imperfect ground truth. The large building, which could be well segmented by adopting ESPCN with an upscale factor of 2, further demonstrates the feasibility of the proposed method.

It should be noted that the investigation of the feasibility of the SR-integrated method for processing multi-source remote sensing imagery is difficult because these images differ in terms of data acquisition methods, resolution, and color space. However, the testing results confirm that the accuracy and robustness of the proposed SR-integrated method is considerably higher than those of the other methods, and that it can achieve comparably accurate building semantic segmentation results using the provided study materials.

VI. CONCLUSION

In this paper, we presented a novel SR-integrated method for building semantic segmentation of multi-source remote sensing imagery of different resolution. The experimental results demonstrate the potential and the capability of the proposed method to solve the problem caused by the resolution of the training data being unaligned with that of the testing data. In particular, the proposed SR-integrated method could achieve considerably higher accuracy and more precise segmentation results than the other methods, which also indicates the feasibility of our proposed method. In addition, it is important to carefully consider the color influence on multi-source remote sensing imagery, investigate the method of balancing resolution and reconstruction quality to enhance the segmentation to a maximum extent, optimize the robustness of both segmentation and SR models, and explore the effectiveness of proposed method in other study areas with buildings in different types, which we aim to study in future.

REFERENCES

[1] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 1873–1876.

[2] S. Saito and Y. Aoki, "Building and road detection from large aerial imagery," *Proc. SPIE*, vol. 9405, Feb. 2015, Art. no. 94050K.

[3] Y. Yan, G. Liu, S. Wang, J. Zhang, and K. Zheng, "Graph-based clustering and ranking for diversified image search," *Multimedia Syst.*, vol. 23, no. 1, pp. 41–52, 2017.

[4] Y. Wei, Z. Zhao, and J. Song, "Urban building extraction from high-resolution satellite panchromatic image using clustering and edge detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, vol. 3, Sep. 2004, pp. 2008–2010.

[5] B. Fulkerson, A. Vedaldi, and S. Soatto, "Class segmentation and object localization with superpixel neighborhoods," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 670–677.

[6] M. Song and D. Civco, "Road extraction using SVM and image segmentation," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 12, pp. 1365–1371, 2004.

[7] O. Teboul, L. Simon, P. Koutsourakis, and N. Paragios, "Segmentation of building facades using procedural shape priors," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3105–3112.

[8] A. Sampath and J. Shan, "Segmentation and reconstruction of polyhedral building roofs from aerial LiDAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1554–1567, Mar. 2010.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[11] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. Van-Den Hengel, "Effective semantic pixel labelling with convolutional networks and conditional random fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 36–43.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[14] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2016, pp. 1–9.

[15] Z. Guo, Q. Chen, G. Wu, Y. Xu, R. Shibasaki, and X. Shao, "Village building identification based on ensemble convolutional neural networks," *Sensors*, vol. 17, no. 11, p. 2487, 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[18] Y. Liu, D. M. Nguyen, N. Deligiannis, W. Ding, and A. Munteanu, "Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery," *Remote Sens.*, vol. 9, no. 6, p. 522, 2017.

[19] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," 2017, *arXiv:1709.05932*. [Online]. Available: https://arxiv.org/abs/1709.05932

[20] G. Wu, X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, Y. Xu, and R. Shibasaki, "Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks," *Remote Sens.*, vol. 10, no. 3, p. 407, 2018.

[21] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raska, "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 172–181.

[22] K. Zhao, J. Kang, J. Jung, and G. Sohn, "Building extraction from satellite images using mask R-CNN with building boundary regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 242–246.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[24] R. Delassus and R. Giot, "CNNs fusion for building detection in aerial images for the building detection challenge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 237–241.

[25] V. I. Iglovikov, S. Seferbekov, A. V. Buslaev, and A. Shvets, "TernausNetV2: Fully convolutional network for instance segmentation," 2018, *arXiv:1806.00844*. [Online]. Available: https://arxiv.org/abs/1806.00844

[26] M. Dickenson and L. Gueguen, "Rotated rectangles for symbolized building footprint extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 225–228.

[27] W. Li, C. He, J. Fang, and H. Fu, "Semantic segmentation based building extraction method using multi-source gis map datasets and satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, Jun. 2018, pp. 233–236.

[28] G. Wu, Y. Guo, X. Song, Z. Guo, H. Zhang, X. Shi, R. Shibasaki, and X. Shao, "A stacked fully convolutional networks with feature alignment framework for multi-label land-cover segmentation," *Remote Sens.*, vol. 11, no. 9, p. 1051, 2019.

[29] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.

[30] Q. Chen, L. Wang, Y. Wu, G. Wu, Z. Guo, and S. L. Waslander, "TEMPORARY REMOVAL: Aerial imagery for roof segmentation: A large-scale dataset towards automatic mapping of buildings," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 42–55, Jan. 2019.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[32] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6792–6810, Nov. 2018.

[33] R. Hamaguchi and S. Hikosaka, "Building detection from satellite imagery using ensemble of size-specific detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 223–227.

[34] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, *arXiv:1902.07296*. [Online]. Available: https://arxiv.org/abs/1902.07296

[35] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 206–221.

[36] G. X. Hu, Z. Yang, L. Hu, L. Huang, and J. M. Han, "Small object detection with multiscale features," *Int. J. Digit. Multimedia Broadcast.*, vol. 2018, Sep. 2018, Art. no. 4546896.

[37] J. M. Haut, M. E. Paoletti, R. Fernandez-Beltran, J. Plaza, A. Plaza, and J. Li, "Remote sensing single-image superresolution based on a deep compendium model," *IEEE Geosci. Remote Sens. Lett.*, to be published.

[38] H. Demirel and G. Anbarjafari, "Satellite image resolution enhancement using complex wavelet transform," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 123–126, Jan. 2009.

[39] P. Thévenaz, T. Blu, and M. Unser, "Image interpolation and resampling," *Handbook Med. Imag., Process. Anal.*, vol. 1, no. 1, pp. 393–420, 2000.

[40] S. Frühwirth-Schnatter, "Data augmentation and dynamic linear models," *J. Time Ser. Anal.*, vol. 15, no. 2, pp. 183–202, 1994.

[41] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[42] Y. Zhang, D. Zhao, J. Zhang, R. Xiong, and W. Gao, "Interpolation-dependent image downsampling," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3291–3296, Nov. 2011.

[43] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, Dec. 1981.

[44] C.-Y. Yang, C. Ma, and M.-H. Yang, "Single-image super-resolution: A benchmark," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 372–386.

[45] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, May 2003.

[46] H. Zhu, X. Tang, J. Xie, W. Song, F. Mo, and X. Gao, "Spatio-temporal super-resolution reconstruction of remote-sensing images based on adaptive multi-scale detail enhancement," *Sensors*, vol. 18, no. 2, p. 498, 2018.

[47] Z. Wang, J. Chen, and S. C. H. Hoi, "Deep learning for image super-resolution: A survey," 2019, *arXiv:1902.06068*. [Online]. Available: https://arxiv.org/abs/1902.06068

[48] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1874–1883.

[49] R. G. Congalton, "A review of assessing the accuracy of classifications of remotely sensed data," *Remote Sens. Environ.*, vol. 37, no. 1, pp. 35–46, 1991.

[50] Y. Sasaki, "The truth of the F-measure," *Teach Tutor Mater*, vol. 1, no. 5, pp. 1–5, 2007.

[51] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[52] M. Polak, H. Zhang, and M. Pi, "An evaluation metric for image segmentation of multiple objects," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1223–1227, 2009.
[53] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[54] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.
[55] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3791–3799.
[56] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*. [Online]. Available: https://arxiv.org/abs/1712.04621
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[58] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*. [Online]. Available: https://arxiv.org/abs/1505.00853
[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: https://arxiv.org/abs/1502.03167
[60] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[61] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 184–199.
[62] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1646–1654.
[63] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep Laplacian pyramid networks for fast and accurate super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 624–632.
[64] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4681–4690.

ZHILING GUO received the M.S. degree from the Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan, in 2017. He is currently pursuing the Ph.D. degree with the Center for Spatial Information Science, The University of Tokyo, Kashiwa, Japan. He has published some papers in international journals such as *Remote Sensing and Sensors*. He has published some papers in international conferences such as IGARSS. His current research interests include machine learning and geographic information science.

GUANGMING WU received the M.S. degree in computational biology and medical sciences from The University of Tokyo, Tokyo, Japan, in 2017. He is currently pursuing the Ph.D. degree with the Center for Spatial Information Science, The University of Tokyo, Kashiwa, Japan. His current research interests include pattern recognition and optimization algorithms. His primary research has been in computer vision and its applications in remote sensing, people flow monitoring, and pose estimation.

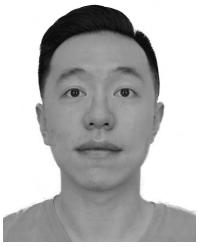
XIAOYA SONG received the M.S. degree from the Harbin Institute of Technology, Harbin, China, in 2016, where she is currently pursuing the Ph.D. degree. She is also pursuing the Ph.D. degree with the Center for Spatial Information Science, The University of Tokyo. Her current research interests include geographic information science and urban–rural planning.

WEI YUAN was born in China, in 1990. He received the B.E. degree from the Wuhan University of Science and Technology, in 2012, and the M.E. and Ph.D. degrees from The University of Tokyo, Tokyo, Japan, in 2015 and 2018, respectively, where he joined the Center for Spatial Information Science, in 2018, as a Researcher. His research interests include photogrammetry and remote sensing, GIS, and computer vision. He is a member of the American Society of Photogrammetry and Remote Sensing and the International Society of Photogrammetry and Remote Sensing.

QI CHEN was born in 1987. He received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2009, 2011, and 2015, respectively. He is currently with the School of Geography and Information Engineering, China University of Geosciences, Wuhan. His research interests include space and aerial photogrammetry, and pattern recognition from remote sensing imagery with deep-learning techniques.

HAORAN ZHANG received the B.Eng. degree and Ph.D. degrees in oil–gas storage and transportation engineering from the China University of Petroleum Beijing (CUPB), Beijing, China, in 2012 and 2018, respectively. He is currently an Assistant Professor with the Center for Spatial Information Science, The University of Tokyo. He has published 50 SCI papers in many top international journals, such as *Applied Energy*, *Energy*, and the *Journal of Computational and Applied Mathematics*. His current research interests include system optimization and data mining.

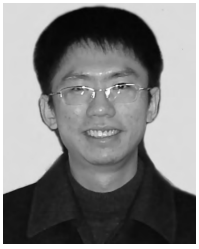
XIAODAN SHI received the B.E. and M.S. degrees in photogrammetry and remote sensing from Wuhan University, China. She is currently pursuing the Ph.D. degree with the Center for Spatial Information Science, The University of Tokyo, Kashiwa, Japan. Her main research interests include computer vision and its applications in trajectory prediction, multi-objects tracking, and remote sensing image segmentation.



MINGZHOU XU received the B.S. degree from the University of California, Santa Barbara, in 2014. He is currently a Visiting Researcher with the Center for Spatial Information Science, The University of Tokyo, Kashiwa, Japan. His current research interests include deep learning and data visualization.

YONGWEI XU received the Ph.D. degree in biomechanical mechanical engineering from The University of Tokyo, in 2013, with a major in agricultural robots, image processing, and computer vision in agricultural mechanization. He is currently a Project Researcher with the Center for Spatial Information Science, The University of Tokyo. He is currently involved in data assimilation analysis based on multi-sensor for people flow model extraction and anomaly detection based on sparse coding.

RYOSUKE SHIBASAKI is currently a Professor with the Center for Spatial Information Science, The University of Tokyo. His major research interests include the integration of data and models based on GIS to reconstruct spatial temporal dynamics of objects, microsimulation modeling, the 3-D mapping of urban areas, human behavior understanding and modeling, the analysis of mobile phone data, and urban informatics and their applications. He has been a Project Leader of ISO/TC211, an international committee to establish international standards of GIS, since 1996. He was also the former President of the Asian GIS Association and the GIS Association of Japan, the Board Member of the Japanese Society of Photogrammetry and Remote Sensing and the Infrastructure Implementation Board and Group of Earth Observations, and a member of the Scientific Committee of World Data System, the International Council of Scientific Union, and the Space Strategic Policy Committee of Japanese Government (Cabinet Office of Prime Minister).



XIAOWEI SHAO received the B.E. and Ph.D. degrees in electronic engineering and information science from the University of Science and Technology of China, Hefei, China, in 1999 and 2006, respectively. From 2004 to 2012, he was with the Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan, where he has been a Project Assistant Professor, since 2008. Since 2012, he has been with the Earth Observation Data Integration and Fusion Research Initiative, The University of Tokyo, as a Project Associate Professor. His research interests include machine intelligence, pattern recognition, and spatial data processing.