 doi:10.20944/preprints202104.0209.v1

Article

Making Low-Resolution Satellite Images Reborn: A Deep Learning Approach for Super-Resolution Building Extraction

Lixian Zhang [1], Runmin Dong [1], Shuai Yuan [2], Weijia Li [3], Juepeng Zheng [1] and Haohuan Fu [1]*

[1] Department of Earth System Science, Tsinghua University, Beijing, 100084, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
[3] CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong, Hong Kong, China
* Correspondence: haohuan@tsinghua.edu.cn

Abstract: Existing methods for building extraction from remotely sensed images strongly rely on aerial or satellite-based images with very high resolution, which are usually limited by spatiotemporally accessibility and cost. In contrast, relatively low-resolution images have better spatial and temporal availability but cannot directly contribute to fine- and/or high-resolution building extraction. In this paper, based on image super-resolution and segmentation techniques, we propose a two-stage framework (SRBuildingSeg) for achieving super-resolution (SR) building extraction using relatively low-resolution remotely sensed images. SRBuildingSeg can fully utilize inherent information from the given low-resolution images to achieve high-resolution building extraction. In contrast to the existing building extraction methods, we first utilize an internal pairs generation module (IPG) to obtain SR training datasets from the given low-resolution images and an edge-aware super-resolution module (EASR) to improve the perceptional features, following the dual-encoder building segmentation module (DES). Both qualitative and quantitative experimental results demonstrate that our proposed approach is capable of achieving high-resolution (*e.g.* 0.5 m) building extraction results at 2×, 4× and 8× SR. Our approach outperforms 8 other methods with respect to the extraction result of mean Intersection over Union (mIoU) values by a ratio of 9.38%, 8.20% and 7.89% with SR ratio factors of 2, 4, and 8, respectively. The results indicate that the edges and borders reconstructed in super-resolved images serve a pivotal role in subsequent building extraction and reveal the potential of the proposed approach to achieve super-resolution building extraction. Our code is available at https://github.com/xian1234/SRBuildSeg.

Keywords: remote sensing imagery; building extraction; super-resolution; deep learning.

1. Introduction

With rapid urbanization in recent years, high-resolution building extraction plays an increasingly essential role in urban planning, change monitoring, and population estimation [1–4]. With a rich set of remotely sensed images, it is possible to infer and distinguish buildings from background objects at pixel level [5]. Such a process is defined as building segmentation or building extraction [6].

In terms of data source, very high resolution (VHR) remotely sensed images were viewed as an essential data source for producing high-resolution building extraction in previous studies, such as 0.1 m airborne images [7,8] and 0.5 m space-borne images [9]. Nevertheless, those VHR images are restricted to a limited spatial extent and temporal availability, thus making the methods which demand VHR images as data source difficult to apply in large area. In contrast, relatively low-resolution images such as satellite-based images of WorldView series (1.2 m-2.4 m) and Planet series (3 m)[10] have better spatiotemporal availability. Notwithstanding, it has been proven by Mariana Belgiu and Lucian Drăguţ [11] along with Huiping Huang et al. [12], that remotely sensed images with relatively lower resolution could generally lead to lower accuracy and coarser boundaries in segmentation results. The resolution inconsistency

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

between the remotely sensed images and the building extraction target greatly impacts the segmentation results. Ryuhei Hamaguchi and Shuhei Hikosaka [13] pointed out that deep learning models that were trained using low-resolution images could hardly extract buildings with significantly different high-resolution. Juan M Haut et al. [14] pointed out that the resolution of remotely sensed images significantly affects the distribution of the spatial features, which is important in distinguishing the pixels of buildings from those of the background. Therefore, it remains challenging to develop an automated framework for achieving super-resolution building extraction results using relatively low-resolution remotely sensed images [15–17].

Despite the difficulties, achieving super-resolution building extraction results from relatively low-resolution remotely sensed images can be of great value. First, for long-term building change detection research, relatively low-resolution remotely sensed images are irreplaceable and exclusive, especially for the late 20th century and early 21th century [18–20]. In these cases, relatively low-resolution remotely sensed images are the only choice for building extraction. In addition, with the rich diversity of satellites and remote sensor technologies, it is common to observe inconsistent spatial resolutions in source datasets and target results for a certain task. For example, B Chen et al. [21] transferred collected low-resolution training material into remotely sensed image pixel classification of another resolution version, making it possible to generate building segmentation results over large areas [11,22,23] or long time periods [24–26] via spatiotemporally available low-resolution remotely sensed images.

To conduct such a task, the simplest and most widely used solution is to interpolate all the resolution-inconsistent images into one desired resolution in the preprocess, for example, by bilinear interpolation and bicubic interpolation [27]. However, interpolation-based methods, for which the generated pixels are calculated by adjacent pixels, suffer the loss of spatial information, especially in the edges and high-frequency regions where interpolation will generate insufficient gradients [27,28].

Alternatively, super-resolution (SR) methods aim at reconstructing low-resolution images into high-resolution versions with finer spatial details [29]. SR provides a promising alternative to map remotely sensed images with inconsistent resolution into a version with uniform resolution for high-resolution building extraction. However, existing SR approaches in remote sensing require a number of external high-resolution images to obtain training datasets. Juan Mario Haut et al. [26] retrieved 2,100 external high-resolution images for training, while Zhenfeng Shao et al. [25] collected more than 100,000 image patches for training. Developing a novel SR approach with no need for external high-resolution images remains challenging but valuable. Moreover, previous studies mainly focus on the perceptual improvements of super-resolved images, with no evaluation regarding how much the improvement of image perceptual quality can be transferred into the improvement of subsequent building extraction.

Hereby, we propose the edge-aware super-resolved building segmentation network (SRBuildingSeg) as a novel framework to achieve super-resolution building extraction. The major contributions in this paper are as follows:

- We propose a two-stage framework for attaining super-resolution building extraction, named SRBuildingSeg, which can make use of the extracted features of the given low-resolution images to improve the performance of building extraction in high-resolution representation.
- Considering the self-similarity between each building in remotely sensed images, we develop an internal pairs generation module (IPG) and an edge-aware super-resolution module (EASR). Using the two proposed modules, we can fully utilize the internal information of the given images to improve the perceptional features for subsequent building segmentation without any external high-resolution images.
- We propose a dual-encoder integration module (DES) for building segmentation tasks which enables our approach to attain super-resolution building extraction by fully utilizing the texture features and enhanced perceptional features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

- We demonstrate that the reconstructed high frequency information of the super-resolved image can be transferred into the improvement of the super-resolution building extraction task. The assessment results reveal that our proposed approach ranks the best among all 8 compared methods.

The rest of the paper is organized as follows. In Section 2, we introduce the related work, including the existing deep learning-based building extraction methods and single image super-resolution techniques. In Section 3, we provide a detailed description of the proposed approach. Experimental results and discussion are given in Sections 4 and 5. We present our conclusions in Section 6.

2. Related Work

2.1. Building Extraction Using Deep Learning Approaches

Since Sakrapee Paisitkriangkrai et al. [30] first proposed a CNN based framework to extract buildings in multispectral images, deep learning based building extraction approaches were proposed and have proven to be effective using VHR images [31–33]. Despite the great success of deep learning approaches in building semantic segmentation, only a few discussions focus on building extraction in which the given images and the target results differ in spatial resolution. Mariana Belgiu and Lucian Drăguţ [11], Ryuhei Hamaguchi and Shuhei Hikosaka [13] compared the building segmentation results of several different approaches using multi-resolution remotely sensed images. They found that the accuracy of extraction results differs with respect to each building size and each specific resolution. Philipp Schuegraf and Ksenia Bittner [34] proposed a hybrid deep learning network for obtaining high-resolution (0.5 m) building extraction results using low-resolution (2 m) multi-spectral and panchromatic images, but their experimental results only show slight improvement in extracting buildings of small size. Zhiling Guo et al. [35] proposed a framework to extract buildings from relatively low-resolution remotely sensed images while using relatively high-resolution images as training material. Nevertheless, their proposed framework could only generate low-resolution segmentation results from the given high-resolution training material. In addition, they used 0.5 m remotely sensed images as "low-resolution images", and their extraction accuracy rapidly declines as the ratio of unaligned resolutions enlarges. Thus, it remains a challenge to obtain fine and high-resolution building extraction from low-resolution remotely sensed images.

2.2. Single Image Super-Resolution

Single image super-resolution (SISR), which aims at reconstructing the image into a high-resolution version while providing finer spatial details than those of the original version [29], has emerged as a promising alternative in mapping low-resolution remotely sensed images into versions of higher resolution [36–38]. Although super-resolution (SR) can reconstruct essential details of land features from the original datasets into a specific desired spatial resolution, it also generally requirsd tremendous external paired high-resolution images for training [25,39]. Moreover, the reconstructed images generated by those SR models strongly relied on the external information provided by training material, which made the collection of training samples more difficult [40]. At the same time, the SR based models trained in an unsupervised way, *e.g.*, the unsupervised Generative Adversarial Networks (GAN) model for SR [41], have emerged as practical alternatives. However, the performances of those unsupervised algorithms usually are unsatisfactory in the high-frequency region as compared with the supervised approaches [42,43]. Another unsupervised SR model, zero-shot super-resolution (ZSSR) [40], requires thousands of gradient updates in image reconstruction. In addition, remotely sensed images are usually large in size, but the ZSSR model is designed for natural images of small size. Thus, it is still challenging to generate fine super-resolved images without using external high-resolution images. Furthermore, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

contributions of the SR methods for the subsequent building extraction lack qualitative evaluation and discussion.

3. Methodology

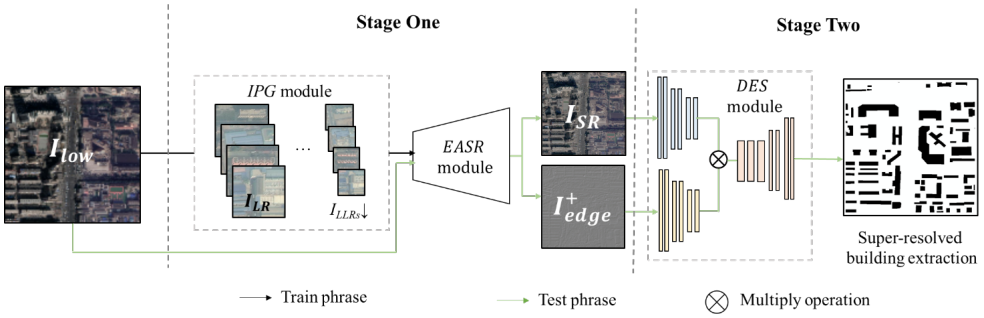

Figure 1. The overall workflow of proposed building extraction using the LR images approach.

In this paper, we aim to utilize the given low-resolution remotely sensed images to achieve building extraction in high-resolution representation. As illustrated in Fig. 1, the overall framework SRBuildingSeg is a two-stage architecture. Stage one focuses on reconstructing a high-resolution version from the given low-resolution images. We first propose an internal pairs generation module (IPG) to construct LR-HR training pairs, which can improve the model trained in an unsupervised way. Hereafter, we reconstruct the super-resolved images using an edge-aware super-resolution module (EASR) which is trained on the constructed training dataset. Stage two exploits the dual-encoder building segmentation module (DES) to achieve building extraction in high-resolution representation, which takes both super-resolved images and enhanced perceptional features as input in order to improve the segmentation performance.

We will elaborate on the details of SRBuildingSeg in the following sections. In Section 3.1 and 3.2, we respectively introduce the IPG module and EASR module. The description of DES is presented in Section 3.3, the assessment criteria are presented in Section 3.4, and the loss fuction is presented in Section 3.5.

3.1. Internal Pairs Generation Module

Existing supervised SR methods in the remote sensing domain require a large number of LR-HR pairs as training material. In addition, the performance of supervised approaches strongly relies on the external information extracted from LR-HR pairs, *e.g.*, the representativeness of the training dataset. Considering the limitations of supervised SR approaches, we propose an internal pairs generation module (IPG) to obtain LR-HR training pairs without any external high-resolution images. Different from existing supervised approaches, the IPG can fully exploit the self-similarity of the remotely sensed image, which generally covers a large area and thus contains buildings of nearly all various colors, shapes, surroundings, materials, heights, and forms. The internal information of the remotely sensed images is a generalized and representative information source, which proved its effectiveness in the training of the SR model [24,44].

The proposed IPG consists of four steps to generate the HR and its corresponding LR (LR-HR) training pairs from the given low-resolution image I_{low}. First, we obtain the "HR training pairs" by simply splitting and cropping the given low-resolution image I_{low}. In other words, the "HR training pairs" I_{LR} is actually presented in relatively low resolution, which is considered as a high-resolution information source in the process of training dataset generation. The corresponding LR training pairs $I_{LLRs\downarrow}$ are then obtained by downscaling each image in I_{LR}, where the s represents the desired SR scale factor. The "LR training pairs" $I_{LLRs\downarrow}$ is actually a lower-resolution version of the given I_{low}. The generated I_{LR} and $I_{LLRs\downarrow}$ consist of many LR-HR image pairs, which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

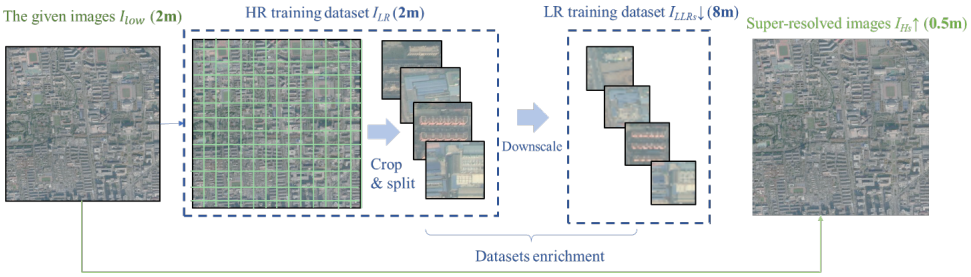

Figure 2. An example workflow of the proposed LR-HR training pairs generation with a scaling factor of 4.

can be used as input and target in the training process of the subsequent SR module. Furthermore, for the sake of robustness, as well as to enrich the diversity of building sizes, we generate many versions of the LR-HR image pairs using a random upscale factor. Finally, the training material is enriched by randomly transforming each image in LR-HR pairs using 4 rotations (0°, 90°, 180°, 270°), mirror reflecting in the vertical and horizontal directions, flipping, resampling and adding Gaussian noise.

Taking the training pairs generated with a scale factor of 4 as an example, as illustrated in Fig. 2, we first generate HR training pairs I_{LR} (2 m) via cropping and splitting the given images I_{low} (2 m). We then downscale each image in I_{LR} and generate the corresponding LR training pairs $I_{LLRs\downarrow}$ (8 m). After dataset enrichment, the generated LR-HR pairs are used as training material for SR model. Finally, we use the properly trained the SR model and the given images I_{low} (2 m) as input to generate super-resolved images $I_{Hs\uparrow}$ (0.5 m).

3.2. Edge-Aware Super-Resolution Module for Reconstructing High-Resolution Images

Considering that our LR-HR training pairs are generated using only the given low-resolution images and contain no external information, the high-frequency information of reconstructed images, which plays a vital role in subsequent building extraction [45], remains to be improved. We employ an edge-aware super-resolution module (EASR) to better reconstruct the high frequency of any given low-resolution remotely sensed images. EASR integrates the initial generative adversarial subnetwork and gradient-based enhancement subnetwork. In the training phase, the EASR utilizes the constructed I_{LR} and $I_{LLRs\downarrow}$ as training material. In the test phase, the EASR takes the given LR image I_{low} as input and outputs super-resolved images $I_{Hs\uparrow}$ with given scale factor s as follows:

$$I_{Hs\uparrow} = EASR(I_{low}) \tag{1}$$

The proposed EASR network is illustrated in Fig. 3. EASR is a GAN-based architecture consisting of a generator and a discriminator.

The generator, which aims to reconstruct HR image $I_{Hs\uparrow}$ from given LR image I_{low} with given scale factor s, consists of an initial reconstruction subnetwork and gradient-based enhancement subnetwork. The reconstruction process contains the following three steps:

The first step reconstructs an initial SR image I_{init} using the initial generative adversarial subnetwork composed of several residual blocks and a reconstruction layer as decoder for generating the intermediate HR result I_{init} , which thus helps to achieve overall performance for the reconstruction of $I_{Hs\uparrow}$.

The second step focuses on the reconstruction of high-frequency information I^{+}_{edge}, which plays an important role in distinguishing the borders and edges of buildings in remotely sensed images. In this step, we first utilize gradient guidance operation to detect and extracts gradient information from I_{init}, which is intuitively useful for better inferring the local intensity of sharpness. In addition, a frame branch and mask branch

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

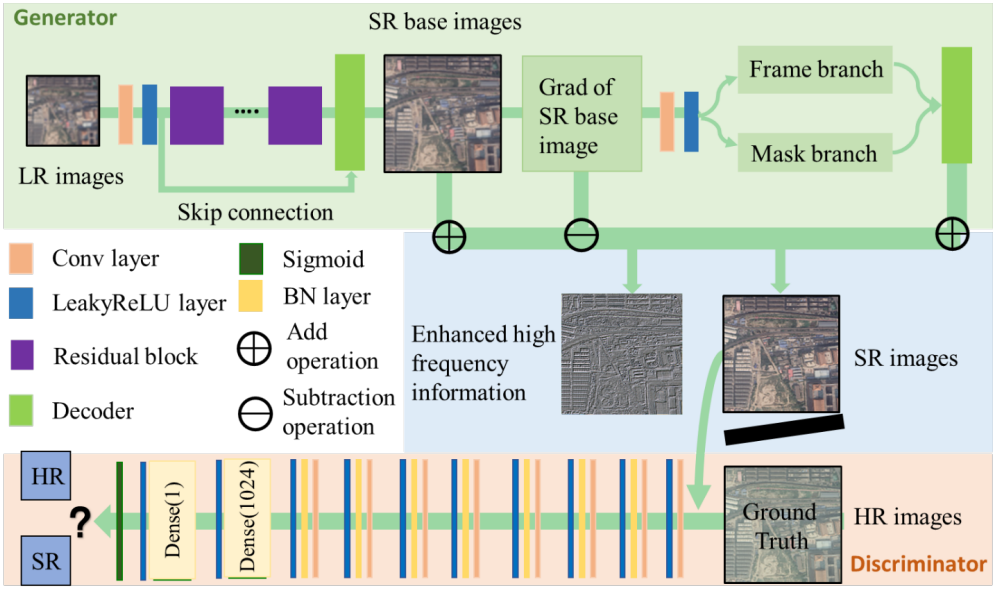

Figure 3. The architectures of the proposed EASR module.

is utilized to extract fine edge maps from the gradient information. These two branches are utilized to learn the noise mask through the attention mechanism so that the network can focus on the real edge information to achieve the purpose of removing noises and artifacts. Specifically, the mask branch consists of three convolutional layers, which aims to adaptively learn specific weight matrices with soft attention to the relevant information. The frame branch contains several residual blocks to infer and extract the sharp edge information. Therefore, the gradient-based enhancement subnetwork reconstructs I^{+}_{edge} as follows:

$$I^{+}_{edge} = GE(I_{init}) \tag{2}$$

where $GE(\cdot)$ denotes the mapping function of the gradient-based enhancement subnetwork, which consists of gradient calculator, frame branch and masks branch. The enhancement subnetwork can reconstruct the edge while reducing the noises and maintaining sharpness.

The third step concatenates the initial SR image I_{init} and enhanced I^{+}_{edge}, and produces the final enhanced SR images I_{SR} as follows:

$$I_{SR} = I_{init} - I_{edge} + I^{+}_{edge} \tag{3}$$

While the generator module is dedicated to reconstructing an SR image which is similar to the ground truth HR image, and the discriminator module aims to distinguish the reconstructed SR images from ground truth HR images. For the discriminator, we take the architectural design in [45] as a reference but use the maximum pooling to replace the strided convolution.

3.3. Segmentation Network for Building Extraction

Using the reconstructed HR image $I_{Hs\uparrow}$ and corresponding building footprint label as training material, we train a dual-encoder segmentation module (DES) for building extraction in stage two. The proposed DES is a modified version of DlinkNet, which was firstly proposed by Lichen Zhou et al. [47] and proved to be effective in several recent studies [48–50]. The proposed DES contains two encoder submodules and one decoder submodule. As discussed above, the high-frequency information of reconstructed images

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

can help define the building boundaries. Hence, we append an extra encoder module which takes the reconstructed high-frequency information I^{+}_{edge} as input to assist the segmentation module in distinguishing building area from background. The final building segmentation Seg is produced as follows:

$$Seg = DES(I_{Hs\uparrow}, I^{+}_{edge}) \tag{4}$$

Each encoder of the proposed DES uses a ResNet-34 pre-trained weight on the ImageNet dataset as an initial parameter. In addition, we employ dilated convolution layers with dilation rates of 1, 2, 4, and 8 to improve the global and local representativeness of the buildings. The other submodule is the decoder of the segmentation network, which is in accordance with the decoder in U-net. The decoder uses transposed convolution layers to upscale the feature map to the same size as the size of input images.

3.4. Loss Function

In stage one, we utilize commonly used loss functions for SR methods, including reconstruction loss L_{rec}, adversarial loss L_{adv} [45], content loss L_{cont} [46], and a total variation (TV) regularization L_{tv} [51] to constrain the smoothness of I_{SR}. The overall loss is defined as:

$$L_{total} = L_{rec} + \alpha L_{adv} + \beta L_{cont} + \gamma L_{tv} \tag{5}$$

Where α, β and γ denote the weights of each loss.

The reconstruction loss L_{rec} is utilized to preserve the consistency of image content between the super-resolved image I_{SR} and HR image I_{LR}, which is defined as:

$$L_{rec} = \sqrt{(I_{SR} - I_{LR})^2} \tag{6}$$

The content loss enforces the generator to generate an intermediate I_{SR} image similar to I_{init}, which is defined as:

$$L_{cont} = \sqrt{(I_{init} - I_{LR})^2} \tag{7}$$

The adversarial loss helps the network to improve the perceptual quality of generated images. The discriminator and the generator are optimized as follows:

$$L_{adv} = -log(D(G(I_{LR}))) \tag{8}$$

$$L_{adv-D} = -log(D(I_{LR})) - log(1 - D(I_{LR})) \tag{9}$$

The total variation (TV) loss aims to constrain the smoothness of I_{SR}, which is defined as:

$$L_{tv} = ||\nabla(I_{SR})||_2 \tag{10}$$

Where $\nabla(\cdot)$ denotes the gradient operator among the horizontal and vertical directions.

In stage two, we utilize commonly used binary cross entropy loss for the segmentation task.

4. Experiments

4.1. Study Area and Data

The study area contains the main city zone of three megacities in China, including Beijing, Shanghai and Xi'an. The study area cover a total of approximately 1860 km^2 and contains multiple building types. As shown in Fig. 4, the study areas cover variable types, forms, and shapes of buildings, including the most modern buildings in developed areas and factories under development, all of which make the selected areas representative and remarkable for this study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

Figure 4. Examples of buildings with various shapes, types, and forms in study area.

As for datasets, versions of remotely sensed images of several resolutions are retrieved from Google Earth for the year 2018, including resolutiosn of 0.5 m, 1.0 m, 2.0 m, and 4.0 m, which are used as given material with scale ratio of 1, 2, 4, and 8 for each experimental case. For the training period in stage one, training and evaluation material were generated via segmenting the input images into patches of size 128×128 pixels. For the training period in the segmentation stage, images were cropped into 1024×1024 pixel patches. We extract a total of 36, 000 group images with size of 1024×1024 as our datasets. The datasets were further divided into training set, test dataset, and validation set according to 7:2:1 proportions.

With regard to the annotated dataset, the building footprint was annotated manually with the referenced remotely sensed images retrieved in 2018. The annotated dataset contains spatial coordinates of all annotated building polygons, and a further rasterization process was conducted in the QGIS platform to generate ground truth labels with the corresponding resolution for each baseline case. Note that a few mismatch cases are inevitable between annotation results and the actual 'ground truth' as a result of limitations in human-based interpretation and minor time inconsistency between retrieved images and referenced images for annotating buildings.

4.2. Implementation Details

Two experiments are conducted for verifying the effectiveness of our proposed two-stage SRBuildingSeg. In the first experiment, we compare the building footprint segmentation results of varied unsupervised SR approaches and the same segmentation approach. The other experiment compares the building extraction performance using the proposed SR methods and varied segmentation approach.

In the training phase, our method is implemented in PyTorch. All the networks in this paper are trained by mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and the weight decay of 0.0005. The learning rate of the super-resolution stage is initialized as 0.001 and the learning rate of the segmentation stage is initialized as 0.0001. We utilize a reduced rate of 0.9 after every 5 epochs for both stages. Our network converges in 100 epochs in both stages, and the batch size is set to 5. An NVIDIA 2080Ti GPU is used for training.

4.3. The Effect of Super-Resolution in Building Extraction

In this section, we focus on comparing the effects contributed by the super-resolution stage in achieving super-resolution building extraction. Therefore, we train each DlinkNet [47] for the segmentation stage under the same conditions while using different SR methods in the super-resolution stage. Considering that the IPG module can help train the SR techniques in an unsupervised way, we select 2 unsupervised SR approaches (i.e., TSE [52], ZZSR [40]) as well as 2 supervised SR methods (i.e., SRGAN [45], EEGAN [46]) which are trained on the dataset generated by our proposed IPG module. All segmentation networks are trained under the same conditions using 0.5 m super-resolved images. According to the scale ratios in generating those SR images, our building segmentation experiment consists of 3 cases, including ratio x2 (the resolution of LR images is 1 m),

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

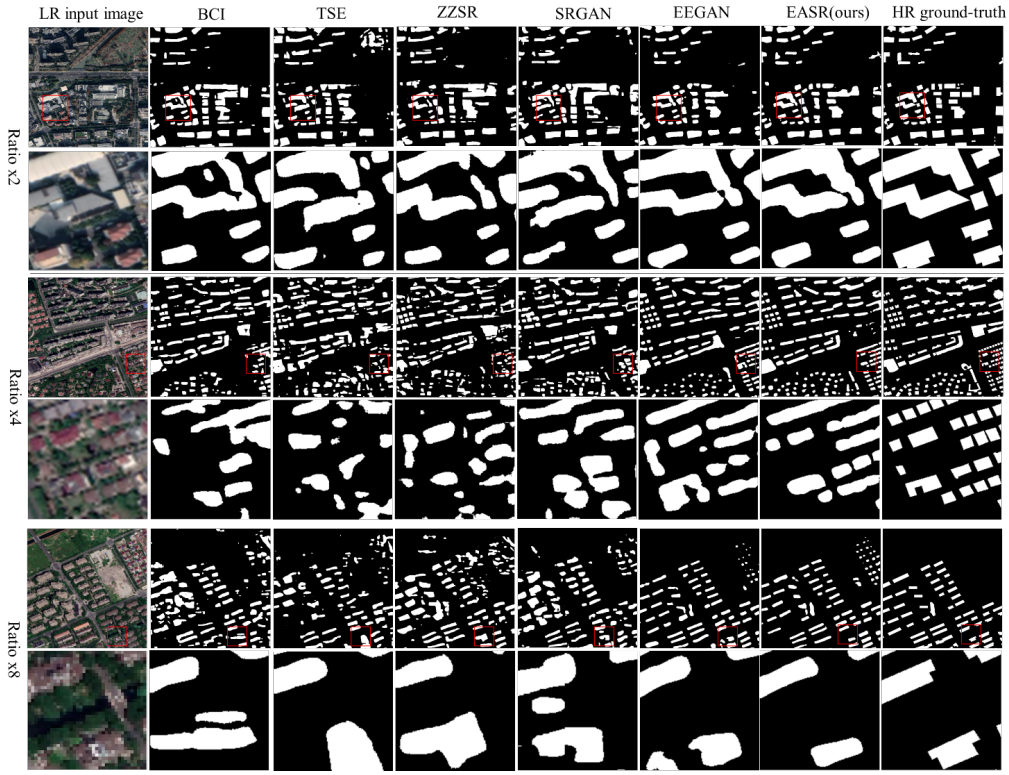

Figure 5. Qualitative examples of segmentation results of each method with ratio factors of 2, 4 and 8.

ratio x4 (the resolution of LR images is 2 m), and ratio x8 (the resolution of LR images is 4 m). Note that the case of using the bicubic interpolated remotely sensed images to train the segmentation model (BCI) is viewed as baseline in this experiment.

Table 1. Quantitative evaluation results on the super-resolution stage.

Stage one	Stage two	Scale	IoU	Precision	Recall	F1 score	Kappa
BCI	DlinkNet	2	0.6206	0.7420	0.7914	0.7659	0.6684
TSE			0.6403	0.7720	0.7896	0.7807	0.7157
ZZSR			0.6492	0.7760	0.7989	0.7873	0.7066
SRGAN*			0.6538	0.7638	0.8195	0.7907	0.7537
EEGAN*			0.6633	0.7805	0.7986	0.7976	0.7438
EASR(ours)			**0.6721**	**0.7965**	**0.8288**	**0.8039**	**0.7771**
BCI	DlinkNet	4	0.6069	0.7499	0.7221	0.7554	0.6069
TSE			0.6184	0.7693	0.7592	0.7642	0.6969
ZZSR			0.6224	0.7786	0.7563	0.7673	0.6972
SRGAN*			0.6263	0.7787	0.7842	0.7702	0.7160
EEGAN*			0.6336	0.7875	0.7643	0.7757	0.6940
EASR(ours)			**0.6413**	**0.7919**	**0.7916**	**0.7814**	**0.7361**
BCI	DlinkNet	8	0.5616	0.7206	0.7180	0.7193	0.6152
TSE			0.5822	0.6978	**0.7785**	0.7359	0.6866
ZZSR			0.5925	0.7366	0.7518	0.7441	0.6906
SRGAN*			0.5863	0.7225	0.7568	0.7392	0.6589
EEGAN*			0.6060	0.7547	0.7547	0.7547	0.7041
EASR(ours)			**0.6237**	**0.7818**	0.7551	**0.7682**	**0.7214**

* indicates that this supervised SR method is trained on datasets generated by our proposed IPG module.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

Table 1 presents the quantitative evaluation of the super-resolved building extraction with the scale factors of 2, 4, and 8 by using those methods. Note that it presents the average results of all images collected for testing to provide a global view. All original high-resolution images in the test area are used for testing. According to the quantitative assessment reported in Table 1, the proposed EASR achieves better performance over other methods, which reveals that the EASR module helps to improve the representativeness of reconstructed images distinguishing from background. In comparison with unsupervised methods, the building extraction achieved via integration of the supervised SR method and proposed IPG module achieve better performance, indicating that our proposed IPG module offers an advantage in helping supervised SR methods to fully utilize the internal information of low resolution remotely sensed images, which can also be proven through qualitative evaluation of Fig. 5. As shown in Fig. 5, the proposed approach exhibits a great advantage in extracting borders and primary structures of buildings in remotely sensed images.

As the SR scale ratio enlarges, the IPG module contributes less improvement and even regresses in achieving super-resolution building extraction. The reason for this phenomenon is that the IPG module takes LR images as input to extract useful information for reconstructing edges and borders of super-resolved images. Meanwhile, as the SR scale ratio enlarges, the LR images tend to contain more noise and blurring, which makes it difficult to extract useful information.

4.4. The Effect of the Segmentation Module in Building Extraction

In this section, we compare our proposed DES module with four other state-of-the-art segmentation methods (i.e., Unet [53], DeepLabv3p [54], PSPNet [55], and DlinkNet [47]). For a fair comparison, we train each segmentation network using the same super-resolved images reconstructed via our proposed EASR. For a fair comparison, all segmentation networks are trained under the same conditions using 0.5 m super-resolved image.

Table 2. Quantitative evaluation result on the segmentation stage.

Stage One	Stage Two	Scale	IoU	Precision	Recall	F1 score	Kappa
EASR(ours)	Unet	2	0.6081	0.7546	0.7580	0.7563	0.6654
	DeepLabv3p		0.6536	0.8162	0.7664	0.7905	0.7160
	PSPNet		0.6522	0.7991	0.7801	0.7895	0.7125
	DlinkNet		0.6721	0.7805	0.8288	0.8039	0.6684
	DES(ours)		**0.7070**	**0.8265**	**0.8305**	**0.8278**	**0.7761**
EASR(ours)	Unet	4	0.5889	0.7037	0.7830	0.7413	0.6585
	DeepLabv3p		0.6336	0.7716	0.7643	0.7757	0.6940
	PSPNet		0.6385	0.7646	0.7947	0.7794	0.6296
	DlinkNet		0.6413	0.7787	0.7842	0.7814	0.7206
	DES(ours)		**0.6595**	**0.7875**	**0.8195**	**0.7948**	**0.7361**
EASR(ours)	Unet	8	0.5414	0.7404	0.6683	0.7025	0.6045
	DeepLabv3p		0.6155	0.7849	0.7313	0.7620	0.6786
	PSPNet		0.6317	0.7854	0.7634	0.7743	0.6974
	DlinkNet		0.6237	0.7818	0.7551	0.7682	0.7214
	DES(ours)		**0.6346**	**0.7955**	**0.7682**	**0.7765**	**0.7310**

We demonstrate the qualitative evaluation via visualizing the results between predicted results and ground-truth labels. As demonstrated in Fig. 6, the segmentation results of the proposed DES can maintain the main structures and borders of buildings, while others fail to extract buildings, especially in large scale factor of building extraction tasks (the ratio x8 cases in Fig. 6), which reveals that our approach can significantly improve feature representativeness in the process of building extraction, especially in the region of borders of buildings. Furthermore, our proposed approach shows its robustness in extracting buildings of variable density, height, textures, and forms, while

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

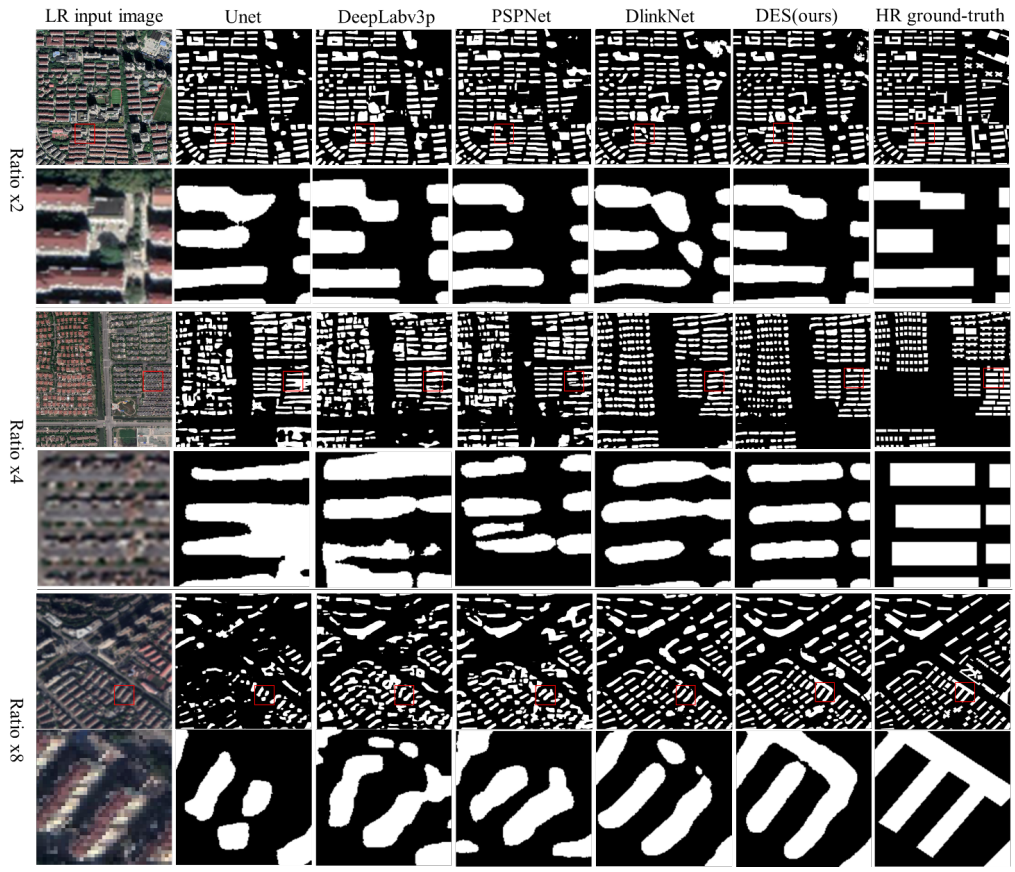

Figure 6. Qualitative examples of segmentation results of each method with ratio factors of 2, 4 and 8.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

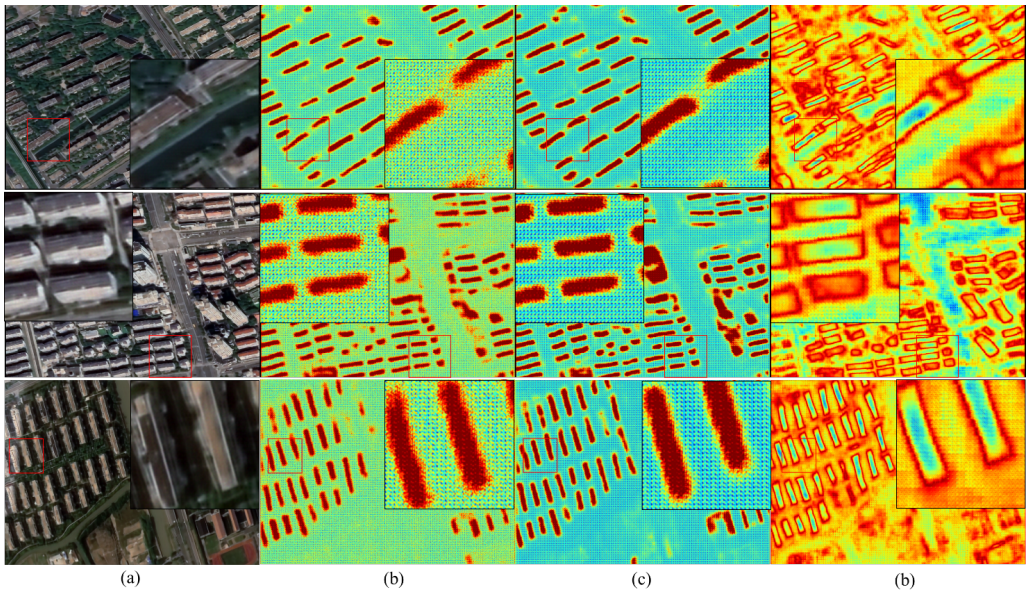

Figure 7. The visualizations of the penultimate CNN layer (better visualized in color) in (a) DlinkNet and (b) DES. (c) is the variance between (a) and (b), which denotes the enhanced information in the segmentation task.

the others result in the unclear contour of buildings (the ratio x4 cases in Fig. 6). The quantitative evaluation, as shown in Table 2, indicates that the proposed approach could achieve better performance over other methods with regard to IoU, recall, F1 score, overall accuracy, and kappa coefficient. This signifies that our proposed approach can enhance the comprehensive features and information of super-resolved remotely sensed images with an appropriate SR scale factor. It could be inferred from the segmentation details, as shown in Fig. 6, that the extracted results outperform other methods with fewer false positive cases, especially in the vicinity of building boundaries.

5. Discussion

Since the above experimental results show the potential of the proposed approach in achieving high-resolution building extraction using low-resolution images, the mechanism and limits of the proposed approach necessitate further discussion. In this section, we primarily discuss 2 topics: 1) how enhanced high frequency information influences the super-resolution building extraction and 2) what the limits of the proposed approach are in conducting super-resolution building extraction tasks.

5.1. The Effectiveness of High Frequency Information in Building Extraction

As demonstrated in Table 1 and Table 2, it seems feasible to achieve high-resolution building extraction via integrating super-resolution and building segmentation methods. In comparison to the building extraction results using bicubic interpolated images, all SR-integrated methods achieve better performance with the scaling ratios of 2, 4 and 8. In addition, the details of reconstructed images using our approach are well-maintained in comparison to those of other integrated methods. A simple but vital question is: how does enhanced high frequency information influence the super-resolution building extraction?

As demonstrated in Fig. 7, we visualized the penultimate CNN layer of DlinkNet (Fig. 7 (a)) and our proposed DES (Fig. 7 (b)). Specifically, we visualized the variance between the penultimate CNN layer of DlinkNet and DES (Fig. 7 (c)) , which highlights the enhanced high frequency information in the segmentation task. Owing to the fusion of enhanced details and input images, it is clear that the features in edges and borders of buildings are well represented, resulting in better building extraction results. Meanwhile,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

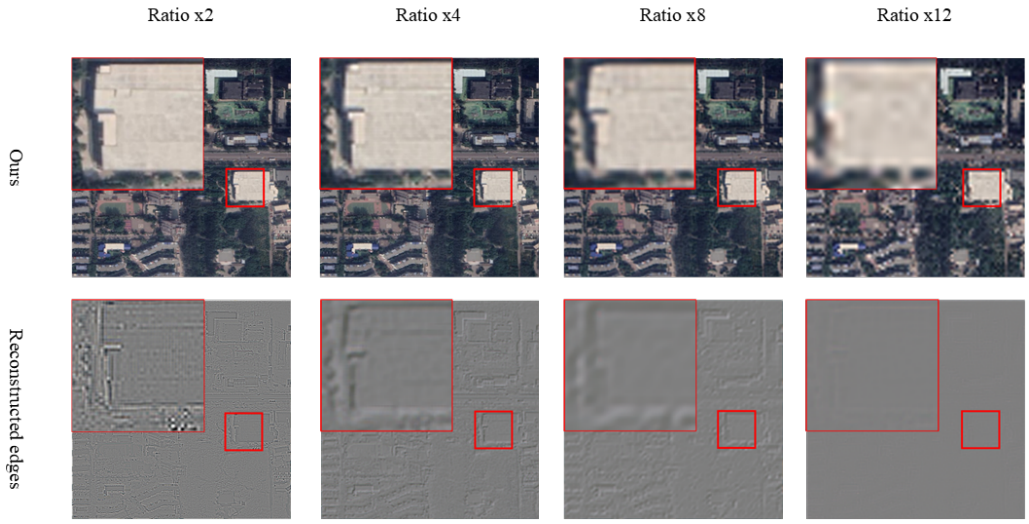

Figure 8. Qualitative assessment of the improvements contributed by the enhancement module with scale ratios of 2 (column I), 4 (column II), 8(column III), and 12(column IV).

the super-resolved images generated via other SR methods maintain better edges and borders in comparison with that of bicubic interpolated images, thus resulting in better performance. This reveals that edges and borders of buildings serve pivotal roles in building segmentation from remotely sensed images.

5.2. The Limitations of the Proposed Approach

Since we can generate high-resolution building extraction results with acceptable accuracy using training material with resolution around 4 m, it seems theoretically possible that we can use even lower resolution material, such as 6 m, to achieve high-resolution building extraction results following the same approach. As shown in the experimental results of remotely sensed image SR and building extraction, the improvements contributed by SR methods rapidly decline as the ratio factor enlarges, especially in the cases with ratio factor of 8. This is primarily attributed to the unsatisfactory reconstruction results obtained for the edges and borders. Nevertheless, whether there is a limit scaling ratio in conducting high-resolution building extraction using low-resolution material remains to be determined.

Table 3. Quantitative evaluation of the results of building extraction.

Methods	Scale	IoU	Precision	Recall	F1 score	Kappa
BCI	2	0.6206	0.7420	0.7914	0.7659	0.6684
Ours		0.7070	0.8265	0.8305	0.8278	0.7761
BCI	4	0.6069	0.7499	0.7221	0.7554	0.6069
Ours		0.6595	0.7875	0.8195	0.7948	0.7361
BCI	8	0.5616	0.7206	0.7180	0.7193	0.6152
Ours		0.6346	0.7955	0.7682	0.7765	0.7310
BCI	12	0.5279	0.7270	0.6584	0.6910	0.5837
Ours		0.5414	0.7404	0.6683	0.7025	0.6045

Furthermore, we demonstrate the proposed approach in conducting high-resolution building extraction with ratio factors of 2, 4, 8, and 12. Note that the training material with ratio factor of 12 was generated from OHGT using bilinear interpolation. As shown in Table 3 and Fig. 8, two shortages emerge as the ratio factor of super-resolution in building extraction continually increases, which could lead to the theoretical ratio limits of our proposed approach. On one hand, the higher the ratio is, the harder the training process becomes. The training pairs generation module first downscales the given input

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2021 doi:10.20944/preprints202104.0209.v1

images, after which much coarser images are generated for the input of the SR module, resulting in a remotely sensed image without sufficient important details for subsequent reconstruction. The strength in providing finer details regarding output images based on the proposed SR module becomes weaker as the scale ratio of the building extraction task enlarges. On the other hand, the improvement contributed by our enhancement module leads to worse performance as the scale ratio of the building extraction task enlarges, which results from the insufficiency of details retrieved from the given low-resolution images. As demonstrated in Fig.8, the high frequency information reconstructed via the proposed EASR becomes coarse, which may even lead to a few artifacts. This indicates that the proposed approach is reaching its limits in conducting the building extraction task at the scale ratio of 12.

6. Conclusion

In this study, we propose a novel two-stage framework (SRBuildingSeg) to achieve super-resolution building extraction using relatively low-resolution remotely sensed images. SRBuildingSeg can fully utilize inherent information of the given low-resolution images to achieve relatively high-resolution building extraction. For generating LR-HR training pairs, we propose an internal pairs generation module (IPG) with no need for external high-resolution images, which can reconstruct super-resolved images with only the given low-resolution images. The edge-aware super-resolution (EASR) module then generates super-resolved images at the desired higher resolution, after which the super-resolution building extraction result is obtained using the dual-encoder building segmentation module (DES). The experimental results demonstrate the capability of the proposed approach in achieving super-resolution building extraction, which outperforms other methods in terms of both the perceptual quality of the super-resolved remotely sensed image and the building extraction accuracy for all small (x2), middle (x4), and large (x8) scale ratios. Furthermore, we demonstrate how the reconstructed high frequency information affects the subsequent building extraction. The assessment results reveal that our proposed approach ranks the best among all SR-integrated methods. In summary, we present the potential of the proposed straightforward approach in demonstrating the use of widely available low-resolution resolution data to obtain high-resolution building extraction results. This approach is practical and especially useful when extra datasets of high-resolution remotely sensed images are unavailable.

Author Contributions: Conceptualization, Lixian Zhang; Data curation, Lixian Zhang and Runmin Dong; Formal analysis, Lixian Zhang; Funding acquisition, Haohuan Fu; Investigation, Lixian Zhang; Methodology, Lixian Zhang and Runmin Dong; Software, Lixian Zhang; Supervision, Haohuan Fu; Validation, Lixian Zhang; Visualization, Lixian Zhang; Writing – original draft, Lixian Zhang; Writing – review and editing, Lixian Zhang, Runmin Dong, Shuai Yuan, Weijia Li, Juepeng Zheng and Haohuan Fu.

Acknowledgments: This research was supported in part by the fund of the National Key Research and Development Plan of China (Grant Nos. 2017YFA0604500, 2017YFB0202204 and 2017YFA0604401), the National Natural Science Foundation of China (Grant No. 51761135015), and by the Center for High Performance Computing and System Simulation, Pilot National Laboratory for Marine Science and Technology (Qingdao).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Feng, T.; Zhao, J. Review and comparison: Building extraction methods using high-resolution images. 2009 Second International Symposium on Information Science and Engineering. IEEE, pp. 419–422.
2. Huang, X.; Zhang, L.J.P.E.; Sensing, R. A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery **2011**. *77*, 721–732.
3. Huang, Z.; Cheng, G.; Wang, H.; Li, H.; Shi, L.; Pan, C. Building extraction from multi-source remote sensing images via deep deconvolution neural networks. 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Ieee, pp. 1835–1838.

4. Rottensteiner, F.; Briese, C. A new method for building extraction in urban areas from high-resolution LIDAR data. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences. NATURAL RESOURCES CANADA, Vol. 34, pp. 295–301.
5. Liu, F.; Jiao, L.; Hou, B.; Yang, S. POL-SAR image classification based on Wishart DBN and local spatial information. *IEEE Transactions on Geoscience Remote Sensing* **2016**, *54*, 3292–3308.
6. Wang, J.; Song, J.; Chen, M.; Yang, Z. Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine. *International Journal of Remote Sensing* **2015**, *36*, 3144–3169.
7. Ghanea, M.; Moallem, P.; Momeni, M. Building extraction from high-resolution satellite images in urban areas: recent methods and strategies against significant challenges. *International journal of remote sensing* **2016**, *37*, 5234–5248.
8. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network. *ISPRS journal of photogrammetry and remote sensing* **2019**, *151*, 91–105.
9. Feng, W.; Sui, H.; Hua, L.; Xu, C.; Ma, G.; Huang, W. Building extraction from VHR remote sensing imagery by combining an improved deep convolutional encoder-decoder architecture and historical land use vector map. *International Journal of Remote Sensing* **2020**, *41*, 6595–6617.
10. Marshall, W.; Boshuizen, C. Planet labs' remote sensing satellite system **2013**.
11. Belgiu, M.; Drăguţ, L. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS Journal of Photogrammetry Remote Sensing* **2014**, *96*, 67–75.
12. Huang, H.; Wu, B.; Fan, J. Analysis to the relationship of classification accuracy, segmentation scale, image resolution. IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477). Ieee, Vol. 6, pp. 3671–3673.
13. Hamaguchi, R.; Hikosaka, S. Building detection from satellite imagery using ensemble of size-specific detectors. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, pp. 223–2234.
14. Haut, J.M.; Paoletti, M.E.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J. Remote sensing single-image superresolution based on a deep compendium model. *IEEE Geoscience Remote Sensing Letters* **2019**, *16*, 1432–1436.
15. Jozdani, S.; Chen, D. On the versatility of popular and recently proposed supervised evaluation metrics for segmentation quality of remotely sensed images: An experimental case study of building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *160*, 275–290.
16. Na, Y.; Kim, J.H.; Lee, K.; Park, J.; Hwang, J.Y.; Choi, J.P. Domain Adaptive Transfer Attack-Based Segmentation Networks for Building Extraction From Aerial Images. *IEEE Transactions on Geoscience and Remote Sensing* **2020**.
17. Zhang, L.; Wu, J.; Fan, Y.; Gao, H.; Shao, Y. An Efficient Building Extraction Method from High Spatial Resolution Remote Sensing Images Based on Improved Mask R-CNN. *Sensors* **2020**, *20*, 1465.
18. Bagan, H.; Yamagata, Y. Landsat analysis of urban growth: How Tokyo became the world's largest megacity during the last 40 years. *Remote sensing of Environment* **2012**, *127*, 210–222.
19. Dong, L.; Shan, J. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry Remote Sensing* **2013**, *84*, 85–99.
20. Weng, Q. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment* **2012**, *117*, 34–49.
21. Chen, B.; Xu, B.; Zhu, Z.; Yuan, C.; Suen, H.P.; Guo, J.; Xu, N.; Li, W.; Zhao, Y.; Yang, J. Stable classification with limited sample: Transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. *Science Bulletin* **2019**.
22. Shrivastava, N.; Rai, P.K. Automatic building extraction based on multiresolution segmentation using remote sensing data. *Geographia Polonica* **2015**, *88*, 407–421.
23. Wang, J.; Qin, Q.; Yang, X.; Wang, J.; Ye, X.; Qin, X. Automated road extraction from multi-resolution images using spectral information and texture. 2014 IEEE Geoscience and Remote Sensing Symposium. IEEE, pp. 533–536.
24. Deng, C.; Zhu, Z. Continuous subpixel monitoring of urban impervious surface using Landsat time series. *Remote Sensing of Environment* **2020**, *238*, 110929.
25. Fu, Y.; Liu, K.; Shen, Z.; Deng, J.; Gan, M.; Liu, X.; Lu, D.; Wang, K. Mapping impervious surfaces in town–rural transition belts using China's GF-2 imagery and object-based deep CNNs. *Remote Sensing* **2019**, *11*, 280.
26. Gong, P.; Li, X.; Zhang, W. 40-Year (1978–2017) human settlement changes in China reflected by impervious surfaces from satellite remote sensing. *Science Bulletin* **2019**, *64*, 756–763.
27. Zhang, Y.; Zhao, D.; Zhang, J.; Xiong, R.; Gao, W. Interpolation-dependent image downsampling. *IEEE Transactions on Image Processing* **2011**, *20*, 3291–3296.
28. Thévenaz, P.; Blu, T.; Unser, Michael analysis. Image interpolation and resampling **2000**. *1*, 393–420.
29. Park, S.C.; Park, M.K.; Kang, M.G. Super-resolution image reconstruction: a technical overview. *IEEE signal processing magazine* **2003**, *20*, 21–36.
30. Paisitkriangkrai, S.; Sherrah, J.; Janney, P.; Hengel, V.D. Effective semantic pixel labelling with convolutional networks and conditional random fields. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36–43.

31. Lin, J.; Jing, W.; Song, H.; Chen, G. ESFNet: Efficient Network for Building Extraction From High-Resolution Aerial Images. *IEEE Access* **2019**, *7*, 54285–54294.
32. Mou, L.; Bruzzone, L.; Zhu, X.X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2018**, *57*, 924–935.
33. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 3639–3655.
34. Schuegraf, P.; Bittner, K. Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN. *ISPRS International Journal of Geo-Information* **2019**, *8*, 191.
35. Guo, Z.; Wu, G.; Song, X.; Yuan, W.; Chen, Q.; Zhang, H.; Shi, X.; Xu, M.; Xu, Y.; Shibasaki, R. Super-Resolution Integrated Building Semantic Segmentation for Multi-Source Remote Sensing Imagery. *IEEE Access* **2019**, *7*, 99381–99397.
36. Dong, X.; Sun, X.; Jia, X.; Xi, Z.; Gao, L.; Zhang, B. Remote Sensing Image Super-Resolution Using Novel Dense-Sampling Networks. *IEEE Transactions on Geoscience and Remote Sensing* **2020**.
37. Zhang, D.; Shao, J.; Li, X.; Shen, H.T. Remote Sensing Image Super-Resolution via Mixed High-Order Attention Network. *IEEE Transactions on Geoscience and Remote Sensing* **2020**.
38. Zou, F.; Xiao, W.; Ji, W.; He, K.; Yang, Z.; Song, J.; Zhou, H.; Li, K. Arbitrary-oriented object detection via dense feature fusion and attention model for remote sensing super-resolution image. *Neural Computing and Applications* **2020**, pp. 1–14.
39. Yang, D.; Li, Z.; Xia, Y.; Chen, Z. Remote sensing image super-resolution: Challenges and approaches. 2015 IEEE International Conference on Digital Signal Processing (DSP). IEEE, pp. 196–200.
40. Shocher, A.; Cohen, N.; Irani, M. "zero-shot" super-resolution using deep internal learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3118–3126.
41. Haut, J.M.; Fernandez-Beltran, R.; Paoletti, M.E.; Plaza, J.; Plaza, A.; Pla, F. A new deep generative network for unsupervised remote sensing single-image super-resolution. *IEEE Transactions on Geoscience Remote Sensing* **2018**, *56*, 6792–6810.
42. Chen, S.; Han, Z.; Dai, E.; Jia, X.; Liu, Z.; Xing, L.; Zou, X.; Xu, C.; Liu, J.; Tian, Q. Unsupervised image super-resolution with an indirect supervised path. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 468–469.
43. Lugmayr, A.; Danelljan, M.; Timofte, R. Unsupervised learning for real-world super-resolution. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). IEEE, pp. 3408–3416.
44. Socher, R.; Ganjoo, M.; Manning, C.D.; Ng, A. Zero-shot learning through cross-modal transfer. Advances in neural information processing systems, pp. 935–943.
45. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. Photo-realistic single image super-resolution using a generative adversarial network. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690.
46. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-enhanced GAN for remote sensing image superresolution. *IEEE Transactions on Geoscience Remote Sensing* **2019**, *57*, 5799–5812.
47. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet With Pretrained Encoder and Dilated Convolution for High Resolution Satellite Imagery Road Extraction. CVPR Workshops, pp. 182–186.
48. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review. *Remote Sensing* **2020**, *12*, 1444.
49. Jiang, Z.; Chen, Z.; Ji, K.; Yang, J. Semantic segmentation network combined with edge detection for building extraction in remote sensing images. MIPPR 2019: Pattern Recognition and Computer Vision. International Society for Optics and Photonics, Vol. 11430, p. 114300D.
50. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sensing* **2019**, *11*, 696.
51. Aly, H.A.; Dubois, E. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing* **2005**, *14*, 1647–1659.
52. Huang, J.B.; Singh, A.; Ahuja, N. Single image super-resolution from transformed self-exemplars. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5197–5206.
53. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.
54. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision (ECCV), pp. 801–818.
55. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.