FSRSS-Net: High-Resolution Mapping of Buildings from Middle-Resolution Satellite Images Using a Super-Resolution Semantic Segmentation Network

Tao Zhang 1,2,3, Hong Tang 1,3,* , Yi Ding 2, Penglong Li 2, Chao Ji 1,3 and Penglei Xu 1,3

1 State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; zhangtaobnu@mail.bnu.edu.cn (T.Z.); jichao@mail.bnu.edu.cn (C.J.); 201921051188@mail.bnu.edu.cn (P.X.)
2 Chongqing Geographic Information and Remote Sensing Application Center, Chongqing 401121, China; dy@dl023.net (Y.D.); lpl@dl023.net (P.L.)
3 Beijing Key Laboratory for Remote Sensing of Environment and Digital Cities, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
* Correspondence: tanghong@bnu.edu.cn; Tel.: +86-10-5880-6401

Abstract: Satellite mapping of buildings and built-up areas used to be delineated from high spatial resolution (e.g., meters or sub-meters) and middle spatial resolution (e.g., tens of meters or hundreds of meters) satellite images, respectively. To the best of our knowledge, it is important to explore a deep-learning approach to delineate high-resolution semantic maps of buildings from middle-resolution satellite images. The approach is termed as super-resolution semantic segmentation in this paper. Specifically, we design a neural network with integrated low-level image features of super-resolution and high-level semantic features of super-resolution, which is trained with Sentinel-2A images (i.e., 10 m) and higher-resolution semantic maps (i.e., 2.5 m). The network, based on super-resolution semantic segmentation features is called FSRSS-Net. In China, the 35 cities are partitioned into three groups, i.e., 19 cities for model training, four cities for quantitative testing and the other 12 cities for qualitative generalization ability analysis of the learned networks. A large-scale sample dataset is created and utilized to train and validate the performance of the FSRSS-Net, which includes 8597 training samples and 766 quantitative accuracy evaluation samples. Quantitative evaluation results show that: (1) based on the 10 m Sentinel-2A image, the FSRSS-Net can achieve super-resolution semantic segmentation and produce 2.5 m building recognition results, and there is little difference between the accuracy of 2.5 m results by FSRSS-Net and 10 m results by U-Net. More importantly, the 2.5 m building recognition results by FSRSS-Net have higher accuracy than the 2.5 m results by U-Net 10 m building recognition results interpolation up-sampling; (2) from the spatial visualization of the results, the building recognition results of 2.5 m are more precise than those of 10 m, and the outline of the building is better depicted. Qualitative analysis shows that: (1) the learned FSRSS-Net can be also well generalized to other cities that are far from training regions; (2) the FSRSS-Net can still achieve comparable results to the U-Net 2 m building recognition results, even when the U-Net is directly trained using both 2-meter resolution GF2 satellite images and corresponding semantic labels.

Keywords: satellite mapping of buildings; Sentinel-2A images; super-resolution semantic segmentation; feature deconvolution

1. Introduction

Human settlements are one of the most important elements on the earth surface, and they play an important role in urban management, development planning and disaster emergency response. With the launch of a large number of remote sensing satellites, human settlement mapping from satellite remote sensing images has been one of the
hottest research topics during the last decades [1–7]. Satellite images with different spatial resolutions can be utilized to reflect different viewpoints of human settlements. For instance, images with a resolution of hundreds or tens of meters, e.g., MODIS, Landsat or Sentinel images, are often utilized to identify large-scale built-up areas on the earth surface [1,6]. In contrast, each building within a city might be delineated from meter- or sub-meter-resolution images, e.g., WorldView, QuickBird or UAV aerial images [8]. For the sake of simplifying expression, images with a resolution of hundreds or tens of meters, and meter- or sub-meter-resolution images are termed as middle- and high-resolution images in this paper, respectively.

In the last decade, some worldwide products of built-up areas have been produced from middle-resolution images, such as FROM-GLC [9–11], the global classification of ASTER data at 15 m and 30 m resolution [12,13], GlobalLand30 [1], Global Urban Footprint (GUF) production [14], Urban-land [6], GHS-built-up data of 40 years [8,15,16], FROM-GLC10 [7] and so on. As for building extraction in large regions, the automatic detection of buildings from VHR data covering the whole of Europe has been finished by applying textural and multi-scale morphological image features inside the associative inference framework [17]. In recent years, Microsoft [18] and Facebook [19] have carried out building extraction based on convolutional neural networks and deep learning. However, a whole worldwide product of buildings is still in progress [20,21], since it is still rather difficult to identify and extract individual buildings from high-resolution and large-scale remote sensing images. The major difficulties might consist of: (1) the complexity of classification between buildings and non-buildings in terms of image features; (2) the amount of labor to collect and label the massive image dataset; and (3) the intensive computation demand. Recently, convolutional neural networks have been shown to be a potential and effective way to deal with the complexity of classification by deep-learning of multiscale image features [4,22]. However, the other two difficulties still remain silently.

Whether can we directly generate large-scale building maps from middle-resolution satellite images in order to significantly reduce the demand of both huge labor and intensive computation? To answer the question, we utilized a novel concept, i.e., super-resolution semantic segmentation. In other words, given a set of satellite images, super-resolution semantic segmentation produces semantic maps whose spatial resolution is higher than that of the images. Specifically, we proposed a deep-learning approach to delineate high-resolution maps of buildings from middle-resolution satellite images.

To achieve super-resolution semantic segmentation, one needs to solve two key problems: (1) how to achieve super-resolution of satellite images and (2) how to make a reliable prediction for each pixel on the higher-resolution image plane. The deep learning technologies has been shown to be an effective way to learn the corresponding relationship between input images and output semantic maps. This motivates us to use deep learning to realize the corresponding relationship between middle-resolution images and high-resolution maps of buildings. By combining image super-resolution strategies (i.e., enhancing the resolution of images) [23] and CNN semantic segmentation, we propose a super-resolution semantic segmentation network to realize the mapping from middle spatial resolution satellite images to high spatial resolution maps of buildings.

The main contributions of this paper are summarized as follows: (1) we designed a new neural network, i.e., the FSRSS-Net, which is shown to be an effective way to extract high-resolution buildings from middle-resolution images; and (2) we created a large-scale and public available sample dataset for super-resolution semantic segmentation of buildings from Sentinel-2A images.

The rest of this paper is organized as follows. In Section 2, we review related work on super-resolution semantic segmentation. The main idea and details of the proposed method, i.e., FSRSS-Net, is described in Section 3. Both experimental data and results are given in Section 4 and 5, respectively. Some conclusions are drawn in Section 6.
2. Related Work on Super Resolution

Some studies related to the super-resolution semantic segmentation are summarized in this section, e.g., image super-resolution using CNN and feature deconvolution used in image semantic segmentation.

2.1. Super-Resolution Semantic Segmentation

The super-resolution (SR) semantic segmentation model needs to achieve resolution enhancement and pixel level classification. Because there is rarely relevant research on super-resolution semantic segmentation, we summarize the strategies of super-resolution semantic segmentation based on empirical knowledge. From the perspective of the relationship among images, their features and semantics, there are three kinds of approaches to super-resolution semantic segmentation, i.e., (1) SR semantic segmentation of up-sampled images, (2) up-sampling of semantic segmentation results, and (3) SR semantic segmentation of up-sampled features.

As shown in Figure 1a, lower-resolution images are firstly up-sampled into higher-resolution images, followed by a semantic segmentation network of the up-sampled images. This kind of approach is based on the image super-resolution results, which are only suitable to the situation that exists in a lot of both low- and high-resolution image pairs for training. As for large-scale satellite image mapping, one could not collect enough pairs of low- and high-resolution images of the same region in the same period of time. Thus, this method is very difficult due to the lack of enough pairs of low- and high-resolution images. From another perspective, if there are enough high-resolution images, high-resolution semantic segmentation maps are directly generated from the high-resolution images.

Figure 1. The three kinds of approaches to super-resolution semantic segmentation. The “Encoder-Decoder” represents the process of semantic segmentation using a convolutional neural network, e.g., U-Net. The up-sampling with the shape of ‘V’ denotes the process of enlarging the size of inputting images, features or semantic maps.
As shown in Figure 1c, the segmentation result is generated from the image by using a semantic segmentation network, and then the higher-resolution semantic maps could be obtained by up-sampling or deconvolution of the lower-resolution semantic segmentation results. Consequently, the quality of the lower-resolution semantic maps would dominate the final results, since the rich image features are out of the up-sampling processing. Therefore, this way might be trivial because the up-sampling from the low-resolution semantic results to the high-resolution semantic results discards the information of the image.

As shown in Figure 1b, the super-resolution semantic segmentation can be achieved by using the learning features of images. This approach can make full use of the information from both low-resolution image features and high-resolution semantic maps. The proposed method in this paper belongs to this kind of approach to super-resolution semantic segmentation.

2.2. Image Super-Resolution Based on Deep Learning

It is well known that image super-resolution (SR) is an important class of image processing technique to enhance the resolution of images and videos in computer vision. The key point of image SR is how to perform up-sampling (i.e., generating high-resolution images from low-resolution images). The existing studies of up-sampling SR techniques with deep learning can be roughly grouped into four kinds of frameworks [24], i.e., pre-up-sampling SR, post-up-sampling SR, progressive up-sampling SR and iterative up-and-down sampling SR. The SRCNN [25] is the first way to image super-resolution by using the CNN. In the SRCNN, features are, firstly, extracted from low-resolution images, then mapped to high-resolution images, and finally the mapped features are restored to high-resolution images. Based on the SRCNN, the FSRCNN [26] have been proposed to enhance and accelerate the effect of image super-resolution. The ESPCN [27] proposes an efficient method to extract features directly from low-resolution images and calculate high-resolution images. Based on residual network [28], the VDSR [29] takes the interpolated low-resolution image as the input of the network, and then adds the image and the residual learned from the network to obtain the final output of the network. In the past two years, with the success of generative adversarial network in image feature learning and image domain mapping, the SRGAN was proposed [30], which uses perception loss and discriminator loss to improve and restore the authenticity of images.

There are many related research projects on the generation of high-resolution images from low-resolution images [25,27,29–33]. However, research on the generation of high-resolution semantic maps from low-resolution images is very rare. As for high-resolution semantic maps, there is more detailed and abundant information compared with low-resolution. If a model can generate a high-resolution semantic map from a low-resolution image, it will show the contour of the target more precisely. There are some difficulties in obtaining a large number of high-resolution remote sensing images, but the low- and middle-resolution images (Landsat, Sentinel, etc.) can be downloaded freely and widely. For extracting buildings or other ground objects from remote sensing images, it is a great breakthrough to produce fine high-resolution results from low- or middle-resolution images, which might have great practical significance.

2.3. Feature Deconvolution

With the rapid development of CNN, CNNs have shown great advantages in image classification [34], semantic segmentation [35], object detection [36], and other issues. As for semantic segmentation, there is a lot of research on feature up-sampling, such as FCN [37], DeconvNet [38], SegNet [39], DeepLab [40], U-Net [41] and so on.

These classical semantic segmentation neural networks are mainly based on the idea of an encoder-decoder. The encoders of these architectures often consist of a set of layered convolution and pooling for features extraction. However, the decoders have different structures, such as bilinear interpolation up-sampling in the FCN, features skip-connection and deconvolution in the U-Net. The task of the decoder is to project lower-resolution
discriminative features learned by the encoder into a higher-resolution image plane to finish the dense classification.

The decoding process is the inverse operation of encoding. The features of the encoder are up-sampled step by step, and a segmentation map with the same size as the original image is obtained finally. As shown in Figure 2, there are three kinds of methods for up-sampling on the decoder, i.e., un-pooling, interpolating and deconvolution. The un-pooling is often used in CNN to represent the reverse operation of the max-pooling [42]. In practical research, the un-pooling is rarely used because of its inflexibility in size recovery. The interpolation method can also realize the size recovery of an image or feature. On the basis of the original image pixels, the nearest neighbor interpolation, bilinear interpolation and cubic convolution interpolation are often used to insert new elements between the pixel values. Because interpolation uses a few local pixel values and the interpolation method is fixed, it is rigid in the process of model learning. Therefore, it might be not the best choice in practice.

The deconvolution is the inverse process of convolution, which constructs the mapping relationship of pixel values by learning for the size recovery of features [43]. With the successful application of deconvolution in neural network visualization, it has been widely adopted in more and more research [44]. In this paper, the deconvolution is used to implement features up-sampling.

3. The Proposed Method for Super-Resolution Semantic Segmentation

The structure of neural network used in the proposed method, coupled with its key components, is described in this section.
3.1. The Structure of FSRSS-Net

As shown in Figure 3, there are two methods of feature deconvolution before classification in the FSRSS-Net. The first one aims for deconvoluted low-level features, e.g., spectral, geometric and morphological features of buildings, which are learned from the inputting image. The second one aims to extract high-level semantic features using a typical encoder-decoder neural network, followed by a deconvolution of the features. Since the low-level features and the high-level semantic features are used together, the network is called FSRSS-Net.

Inspired by the idea of encoding and decoding for semantic segmentation, we design the structure of FSRSS-Net as shown in Figure 4. For learning the corresponding relationship between middle-resolution (i.e., 10 m Sentinel-2A) images and high-resolution (i.e., 2.5 m buildings) semantic maps, the resolution of images is twofold improved and pixel level semantic segmentation is achieved at the up-sampled image planes.

As shown in Figure 4, the input of the FSRSS-Net is a Sentinel-2A image with the size 64×64. On the one hand, the low-level features with size of 64×64 are directly deconvoluted into feature maps with the size of 128×128. On the other hand, they go through the processing of a variant of the U-Net [41], so that high-level features can be extracted by encoding and decoding in the variant of U-Net. Then, both low-level and high-level features are concatenated and furthermore deconvoluted to features with size of 256×256. Finally, the convolution layer and softmax layer are added to realize pixel level classification. In general, the FSRSS-Net structure makes full use of the primary low-level features and high-level semantic features of images, and after two deconvolution
operations, the spatial resolution is upgraded twice, i.e., from 10-meter images to 2.5-meter semantic maps.

The specification of the FSRSS-Net is given in Figure 5, where the input is of 64×64×4, since three visible and one Near-Infrared bands of Sentinel-2A images are used in this paper. First of all, the inputting image is transformed into 64 feature maps with the size of 64×64 by a 3×3 convolution operation with 1 zero-padding. Then, a batch normalization layer is followed by another convolution. After that, it goes through two parallel processing for low-level and high-level feature extraction and deconvolution as shown in Figure 3. At last, the softmax is used to finish the classification for each pixel on the up-sampled image plane. Please note that there are six kinds of operations in the network, as shown in the legend of Figure 5.

![Figure 5](image)

**Figure 5.** The specification of the FSRSS-Net.

### 3.2. A Variant of the U-Net Used in This Paper

Among many semantic segmentation networks, the U-Net [41] stands out for its unique design and excellent performance. In the process of up-sampling, the feature maps of down-sampling are used, which make full use of the primary features of the shallow layer in the deep convolution. Consequently, this makes the input of convolution more abundant, and the results can reflect the original information of images more. Specially, we have made the following two changes, as shown in Figure 5, comparing with the original U-Net: (1) In the last layer, two convolution layers are added, which makes the depth of the network deeper, and meanwhile, a batch-normalization layer and a dropout layer are added to prevent the over-fitting of the model. (2) In the process of down-sampling and feature extraction, the second and fourth convolutions are modified into dilated convolution with dilation rate of 2 [45], which makes the range of receptive field larger.

To reveal the characteristics of the FSRSS-Net, the variant of the U-Net in Figure 5 is used as a baseline method for the sake of comparison in this paper. Specifically, five convolution layers and a softmax layer are added into the variant U-Net, in order to generate 10 m semantic segmentation results from 10 m Sentinel-2A images. The modified
U-Net generates 10 m semantic segmentation results from 10 m images using the network shown in Figure 6.

![Figure 6. The structure of the modified U-Net.](image)

It should be emphasized that this paper does not compare the advantages and disadvantages of FSRSS-Net and U-Net, nor does it attempt to prove that FSRSS-Net is better than U-Net. In this paper, we hope to prove that FSRSS-Net can generate high-resolution semantic maps from middle-resolution images, and can realize more precise building recognition of remote sensing images.

4. Experimental Data

For super-resolution semantic segmentation of buildings, due to the lack of directly available datasets, we created a set of experimental data covering 35 cities in China. This section introduces the data source, preprocessing of images and building ground-truth in detail. The experimental data covering the 35 cities are partitioned into three groups: model training area, quantitative accuracy evaluation area and qualitative generalization performance analysis area. Finally, we describe the samples for model training and the way to select these samples.

4.1. Data Source and Preprocessing

Figure 7 shows a Sentinel-2 image and its corresponding semantic map in the city of Beijing. These data originate from the Google Earth Engine (GEE) platform [46,47] and the Chinese map-world known as “Tiandi Map”, respectively.

Sentinel-2 is a wide-swath, middle-resolution, multispectral imaging mission with a global 5-day revisit frequency. The Multispectral Instrument (MSI) samples 13 spectral bands: visible and near infrared (NIR) at 10 m, red edge and short-wave infrared (SWIR) at 20 m, and atmospheric bands at 60 m spatial resolution. The images used in our experiments are the Level-2A of Sentinel-2 (Sentinel-2A) with red, green, blue and near infrared spectrum (NIR) bands. The images have been screened according to imaging time and cloud amount, and preprocessed by splicing, cloud removal and cutting.
The Chinese map-world known as ‘Tiandi Map’ is a comprehensive geographic information service website built by the National Platform for Common Geospatial Information Service of China (https://www.tianditu.gov.cn/ accessed on 10 June 2021). Digital line drawing map data can be obtained from the platform. The semantic map of buildings is extracted by color value filtering. Then, the 10 m and 2.5 m building ground-truth maps are derived according to nearest neighbor resampling of the digital maps. It can be seen from Figure 7d,e that the 2.5 m ground-truth map exhibits more precise boundaries of buildings than the 10 m map.

4.2. Dataset

As shown in Figure 8, Sentinel-2A images covering 35 cities are collected from the GEE platform after cloud removal, radiation correction and other preprocessing. These images are grouped into three subsets according to the distribution of these cities. Figure 8 shows the geographic location of these images.
The first group, used to train and validate the models, consists of 19 cities and represents the situation of the major cities under different climates and landforms in China. Buildings in China are mainly located in the big cities, and the big cities are mainly located in the coastal areas, or beside the big rivers and lakes (e.g., the Yangtze River, the Yellow River and the Lake Taihu). Therefore, the selection of 19 training areas mainly considers the eastern and southeastern coastal cities, whose climate type represents the most typical tropical and subtropical monsoon climate in China. Among the 19 training areas, Zhengzhou has the largest area of 3957.75 km² and Guangzhou has the smallest area of 268.73 km².

The second group, including four cities, is utilized to quantitatively evaluate accuracy for extracting buildings based on the learned models. The four cities include Chongqing, Wuhan, Qingdao and Shanghai with approximately the same area of 203.48 km², and the image pixel size of each city is 1416 × 1437. The four accuracy evaluation regions represent the different landforms of China. Chongqing is a typical mountain city, in which the buildings are distributed on the slope because of the large topographic relief. Qingdao is a seaport city in the North China Plain, in which the buildings are arranged in order. Shanghai, located in the plain of the middle and lower reaches of the Yangtze River, is a typical international metropolis with densely distributed buildings and numerous tall buildings. Wuhan is located in the central hills, with a dense water network, and the buildings are distributed near the small mountains and rivers.

The third group, consisting of 12 cities, is used for qualitative analysis of the generalization ability of the learned model, in which some cities are far from the cities in the first group. The image coverage area of each city in the qualitative analysis area is close to that of each city in the quantitative accuracy evaluation area. There is no corresponding ground-truth with the images in the 12 qualitative analysis cities. The 12 cities can be divided into four regions, representing different geographical divisions: (1) In the eastern and southeastern regions (Wenzhou, Hong Kong and Yangzhou), the climate type is a tropical and subtropical monsoon climate. The urbanization of this region is well developed and the buildings are densely distributed. (2) In the north and northeast regions (Hohhot, Changchun and Harbin), the climate type is a temperate monsoon climate, which belongs to industrial and agricultural base, and there are many large buildings in these cities. (3) In the southwest and central regions (Changsha, Chengdu and Nanning), the climate type is a subtropical monsoon climate, and the urbanization process is fast; there are many buildings demolished and newly built. (4) In the west and northwest (Lhasa, Urumqi and Xining), the climate is a temperate continental climate and plateau alpine climate. The buildings are relatively low and distributed on both sides of the valley and river.

4.3. Training Samples

Figure 9 shows two typical training samples from three kinds of situation, i.e., high density buildings in cities, low density buildings in cities and sparse buildings in rural areas. Each sample is related to three subfigures, i.e., one 10 m image with size of 64 × 64, one 10 m building ground-truth with size of 64 × 64 and one 2.5 m building ground-truth with size of 256 × 256.

Since the mapping time of ‘Tiandi Map’ might not match with the imaging time of Sentinel-2A images, the labels of buildings cannot be directly used as ground-truth for training. We carefully selected 8597 samples to train both the U-Net and FSRSS-Net. These 8597 training samples have been made public, and can be downloaded from the link: https://drive.google.com/open?id=1-ui_KshUbCgaQINiZ0nvQ3k55HOhO9wj accessed on 10 June 2021. In order to keep the balance between buildings and non-buildings in term of the number of pixels, the number of samples is in proportion to the ration of buildings. Specifically, among all of samples, the proportion of building pixels is 83%, and the proportion of background pixels is 17%. Each sample has more than one building and there is no pure background samples for training. There are 47 samples with more than 50% buildings, 211 samples with 40-50% buildings, 3577 samples with 20-40% buildings, 4762 samples with 0-20% buildings.
Figure 9. The partial training samples (each sample includes one 10 m image with 64×64 size, one 10 m building ground-truth with 64×64 size, and one 2.5 m building ground-truth with 256×256 size).

In addition, two images approximately containing 10% cloud in Kunming and Nanjing regions are added to the training datasets. As shown in Figure 10, the shape outline of the thick cloud is drawn by visual interpretation and handwork from the image, and then the area with cloud is regarded as the background in the ground-truth. Therefore, for the pixels with clouds in the image, they are all considered as background pixels in ground-truth, regardless of whether the actual category of those pixels are background or building.

Figure 10. Images with clouds and ground-truth: (a) shows the whole Nanjing area, and (b,c) are the 10 m and 2.5 m ground-truth, respectively, for one small region.

5. Experiment and Discussion

In this section, firstly, the training process of the models is described. Then, both the trained FSRSS-Net and U-Net models are used to identify buildings from Sentinel-2A images, and five accuracy evaluation indexes are applied to evaluate the quantitative accuracy and to compare the results of building identification of 10 m and 2.5 m. Finally, the transfer generalization performance of the model is qualitatively analyzed.

5.1. Training Model

5.1.1. Experimental Setting

Although the balance between buildings and non-buildings has been considered during sample selection, the number of pixels in a batch of samples might still be imbalanced.
In order to avoid this problem, we designed a selectively weighted loss function given by Equation (1). For each batch of samples, if the background pixel ratio is higher than 50%, the loss function uses the weighted-binary cross entropy, otherwise it uses the binary cross entropy. It is important to note that this formula only applies to the case where each sample must have building pixels and background pixels.

\[
\text{Loss} = \begin{cases} 
-\left[ y \times \ln p + (1 - y) \times \ln(1 - p) \right], & Bp \leq 0.5 \\
\left[ y \times \ln p \times \frac{Bp}{1 - Bp} + (1 - y) \times \ln(1 - p) \right], & Bp > 0.5 
\end{cases}
\] (1)

where \(y\) is the label of a sample, i.e., 1 for the building pixel, and 0 otherwise. And \(p\) represents the probability that the sample pixel is predicted to be the building. \(Bp\) is the ratio of background pixels in each batch of samples.

In our experiments, 8 NVIDIA Tesla K80 GPUs are utilized to train both the U-Net and the FSRSS-Net. The batch size was set to 32 and the epoch was set to 400. We use cross entropy defined in Equation (1) as the loss function, and use the stochastic gradient descent (SGD) as the optimizer to learn the model parameters. The initial learning rate of SGD is set to 0.01. When the learning rate is no less than 0.000001, the learning rate decreases exponentially with the increase of training epoch-index. The calculation formula of the learning rate is given by:

\[
lrate = \begin{cases} 
0.000001, & lrate < 0.000001 \\
0.5 \times \frac{\ln n}{10} \times 0.01, & lrate \geq 0.000001 
\end{cases}
\] (2)

where \(n\) is the training epoch-index.

5.1.2. Training Accuracy and Loss

The 8597 training samples are applied to train both the FSRSS-Net and U-Net. Then, the accuracy and loss of training processes are calculated as shown in Figure 11. As for the U-Net, the model tends to converge after the 180\textsuperscript{th} epoch, and the value of training accuracy is stable at 0.913, but the value of training loss still slows down by inches. When the 400 epochs are completed, it probably takes 102 h. As for the FSRSS-Net, the model tends to converge after the 200\textsuperscript{th} epoch, and the value of training accuracy is stable at 0.904. When the 400 epochs are completed, it probably takes 118 h. The accuracy and loss values in the training process indicate that the FSRSS-Net and U-Net network are trained normally and well.

Figure 11. The training process of the U-Net and the FSRSS-Net.

5.2. Quantitative Accuracy Evaluation and Comparison

The data of quantitative accuracy evaluation areas, which covers four cities, is used to evaluate and compare the accuracy of 10 m and 2.5 m building recognition results.
The trained FSRSS-Net and U-Net are used to extract buildings from the images covering Chongqing, Qingdao, Shanghai and Wuhan. First, we map the spatial distribution of the building extraction results of the four cities.

Figures 12–15 show the results of building identification in Chongqing, Qingdao, Shanghai and Wuhan, respectively. In each figure, the left part is the whole range of the quantitative accuracy evaluation city, and the right part is the enlarged display of a local small area in the red box on the left, which used to reflect the details of building identification. The ground-truth is represented in dark green, and the identified buildings are represented in bright red. On a large scale, the results of both 10 m and 2.5 m are fairly good. It is obvious that there is no case in which urban green land and wetland are incorrectly classified into buildings. In addition, the roads between buildings have been clearly distinguished. From the view of local small area, the building recognition results of 2.5 m are more precise than those of 10 m, and the outlines of the buildings are better depicted.

Figure 12. The building extraction results in Chongqing, the mountain city of China. The subfigure (a) represents the large region of Chongqing, and the subfigure (b) is the enlarged display of a local small area in the red box of subfigure (a1). The subfigures (1), (2) and (3) represent the Sentinel-2A 10 m image, the 10 m building recognition result of U-Net model and the 2.5 m building recognition result of FSRSS-Net model.

From Figures 12–15, both FSRSS-Net and U-Net were trained successfully, and the two trained models can recognize buildings. Most importantly, the FSRSS-Net can really produce high-resolution semantic segmentation results from middle-resolution images. In this paper, the U-Net inputs 10 m resolution images and outputs 10 m resolution building recognition results, and the FSRSS-Net inputs 10 m resolution images and outputs 2.5 m resolution building recognition results. The output spatial resolutions of the FSRSS-Net and U-Net model are different, so it is inappropriate to compare the performance of FSRSS-Net and U-Net. We pay more attention to the differences in the spatial distribution of building recognition results, and then highlight the advantages of 2.5 m building recognition results.
Figure 13. The building extraction results in Qingdao, the seaport city of China. The subfigure (a) represents the large region of Qingdao, and the subfigure (b) is the enlarged display of a local small area in the red box of subfigure (a1). The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, 10 m building recognition result of U-Net model and 2.5 m building recognition result of the FSRSS-Net model.

Figure 14. The building extraction results in Shanghai, the coastal plain in the lower reaches of the Yangtze river. The subfigure (a) represents the large region of Shanghai, and the subfigure (b) is the enlarged display of a local small area in the red box of subfigure (a1). The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, 10 m building recognition result of the U-Net model and 2.5 m building recognition result of the FSRSS-Net model.
Figure 15. The building extraction result of Wuhan, the hill and plain in center China. The subfigures (a, b) represent the large region of Wuhan and the enlarged display in the red box of subfigure (a1). The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, the 10 m result of the U-Net model and the 2.5 m result of the FSRSS-Net model, respectively.

From the details of the local small area, the results of 10 m cannot identify some small buildings. Some individual buildings with dense distribution can be identified as contiguous built-up areas, which is shown in the yellow circle of subfigures (b2) in Figures 12–15. On the contrary, as shown in subfigure b3 in Figures 12–15, the 2.5 m building recognition results can distinguish small houses, and the building contour is more accurate. Therefore, from the spatial distribution map of building extraction results, it can be significantly concluded that the building recognition results of 2.5 m are better than that of 10 m.

In order to quantitatively evaluate the accuracy of the 2.5 m and 10 m building recognition results, we evaluate the accuracy based on the overall accuracy (OA), recall (Rec), precision (Pre), F1-score (F1) and intersection over union (IOU). In particular, in order to make the accuracy evaluation scientific and reasonable, for the 2.5 m results of FSRSS-Net, the ground-truth of 2.5 m is used for accuracy evaluation, while for the 10 m results of U-Net, the ground-truth of 10 m is used for accuracy evaluation.

Table 1 shows the relationship between prediction results and ground-truth. As defined in Equation (3), OA represents the ratio of the number of correctly predictive pixel to that of all of prediction. As shown in Equation (4), Rec is the ratio of the number of predicted building pixels to that of actual building pixels. For Equation (5), Pre denotes the ratio of the number of actual building pixel to that of all of predicted building pixels. In Equation (6), F1 is the harmonic average of precision and recall, and IOU is the ratio of building intersection and building union, as shown in Equation (7).
Table 1. The representation of confusion matrix between prediction and ground-truth.

| Ground Truth | Prediction | Sum               |
|--------------|------------|-------------------|
|              | Background | Building          |
| Background   | True Negative (TN) | False Positive (FP) | Actual Background (TN+FP) |
| Building     | False Negative (FN) | True Positive (TP) | Actual building (FN+TP) |
| Sum          | Predicted Background (TN+FN) | Predicted building (FP+TP) | TN+ TP+ FN+ FP |

The OA, Rec, Pre, F1 and IOU are calculated using Equations (3)–(7), respectively:

\[
OA = \frac{TP + TN}{TP + TN + FN + FP} \times 100\% \quad (3)
\]

\[
Rec = \frac{TP}{TP + FN} \times 100\% \quad (4)
\]

\[
Pre = \frac{TP}{TP + FP} \times 100\% \quad (5)
\]

\[
F1 = \frac{2 \times Rec \times Pre}{Rec + Pre} \times 100\% \quad (6)
\]

\[
IOU = \frac{TP}{TP + FN + FP} \times 100\% \quad (7)
\]

where the meanings of TP, TN, FN and FP are shown in Table 1.

In order to evaluate the quantitative accuracy objectively and truly, 766 samples are selected from the four cities by visual judgment. There are 236 samples in Chongqing, 206 samples in Qingdao, 225 samples in Shanghai and 99 samples in Wuhan. In the samples of the four cities, the proportions of building pixels are 17.22%, 20.21%, 21.71% and 22.40%, respectively. In particular, the 10 m recognition results of U-Net are directly up-sampled to 2.5 m, based on the nearest neighbor interpolation, which is used to compare with the 2.5 m recognition results of FSRSS-Net. Table 2 displays the accuracy of 10 m building recognition results, the accuracy of 10 m results interpolation up-sampling to 2.5 m, and the accuracy of 2.5 m building recognition results over the four cities.

On the whole, the OAs in these four cities are higher than 76%, but it is obviously lower than the training accuracy of 19 cities in the model training areas, because the images of the four cities are not ever seen by the model during training. It can be seen from Table 2 and Figure 16 that there is little difference between the accuracy of 10 m results and 2.5 m results. However, the accuracy of 10 m results interpolation up-sampling to 2.5 m is generally lower than the accuracy of 10 m results and the accuracy of 2.5 m results. Finally, we can conclude that: based on the 10 m Sentinel-2A image, the FSRSS-Net can achieve super-resolution semantic segmentation and produce 2.5 m building recognition results. The 2.5 m building recognition results are better than the 10 m building recognition results interpolation up-sampling to 2.5 m.
Table 2. The accuracy evaluation results of the four regions.

| Evaluation Regions | OA    | Rec   | Pre   | F1    | IOU  |
|--------------------|-------|-------|-------|-------|------|
| **Chongqing**      |       |       |       |       |      |
| (236 samples,      | 82.26 | 42.50 | 44.09 | 43.28 | 27.93|
| 17.22% building    |       |       |       |       |      |
| pixels)            |       |       |       |       |      |
| U-Net 10 m result  | 81.77 | 41.09 | 42.67 | 41.87 | 26.80|
| up-sampling to 2.5 |       |       |       |       |      |
| m                  | 82.80 | 39.63 | 44.80 | 42.06 | 26.94|
| FSRSS-Net 2.5 m    |       |       |       |       |      |
| result             | 79.02 | 37.18 | 43.46 | 40.07 | 25.63|
| Qingdao            |       |       |       |       |      |
| (206 samples,      | 78.95 | 36.62 | 42.98 | 39.54 | 25.15|
| 20.21% building    |       |       |       |       |      |
| pixels)            |       |       |       |       |      |
| U-Net 10 m result  | 78.95 | 36.62 | 42.98 | 39.54 | 25.15|
| up-sampling to 2.5 |       |       |       |       |      |
| m                  | 80.48 | 39.15 | 47.55 | 42.94 | 27.80|
| FSRSS-Net 2.5 m    |       |       |       |       |      |
| result             | 80.83 | 53.31 | 53.08 | 53.20 | 36.35|
| Shanghai           |       |       |       |       |      |
| (225 samples,      | 79.79 | 50.75 | 50.53 | 50.64 | 34.09|
| 21.71% building    |       |       |       |       |      |
| pixels)            |       |       |       |       |      |
| U-Net 10 m result  | 79.79 | 50.75 | 50.53 | 50.64 | 34.09|
| up-sampling to 2.5 |       |       |       |       |      |
| m                  | 81.91 | 51.56 | 55.95 | 53.67 | 36.70|
| FSRSS-Net 2.5 m    |       |       |       |       |      |
| result             | 76.61 | 41.46 | 45.01 | 43.16 | 27.72|
| Wuhan              |       |       |       |       |      |
| (99 samples,       | 76.08 | 40.20 | 43.74 | 41.90 | 26.72|
| 22.40% building    |       |       |       |       |      |
| pixels)            |       |       |       |       |      |
| U-Net 10 m result  | 76.08 | 40.20 | 43.74 | 41.90 | 26.72|
| up-sampling to 2.5 |       |       |       |       |      |
| m                  | 77.04 | 39.07 | 45.54 | 42.06 | 26.88|
| FSRSS-Net 2.5 m    |       |       |       |       |      |
| result             | 82.26 | 42.50 | 44.09 | 43.28 | 27.93|
| **Figure 16.** The quantitative accuracy evaluation over the four cities.

5.3. Qualitative Analysis of the Generalization Ability

The learned models are used to identify buildings from the third group of images covering the 12 cities. As shown in Figures 17–20, both 10 m and 2.5 m recognition results of buildings in most of cities look very good.
Figure 17. The building extraction results of the eastern and southeastern region in China. The subfigures (a–c) represent the Wenzhou, Hong Kong and Yangzhou, respectively. The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, the 10 m result of the U-Net model, and the 2.5 m result of the FSRSS-Net model, respectively.

Figure 18. The building extraction results of the north and northeast region in China. The subfigures (a–c) represent the Hohhot, Changchun and Harbin, respectively. The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, the 10 m result of the U-Net model and the 2.5 m result of the FSRSS-Net model, respectively.
These results show that the learned models can be well generalized to some cities in China. However, for some cities, e.g., Lhasa and Hohhot, the results are not as good as other cities, since the two cities are located on the plateau and grassland Gobi, and therefore the land surface of these regions is very different from the cities in the training dataset.

It can be seen from the yellow circled area in subfigures on the second rows of Figures 17–20 that the results of 2.5 m are more precise and accurate, especially in the cities of Wenzhou, Yangzhou, Chengdu, Changsha, Chengdu and Nanning.
In addition, as shown in the suburbs and villages in subfigures (a1) and (b1) in Figure 21, buildings are sparsely distributed, and the learned models can still accurately identify buildings. When the image is covered by clouds, as shown in Figure 21 (c1), the cloud cannot be mistaken as buildings. This shows that the learned model can deal with the variation caused by imaging environment, to some extension.

![Building extraction results of images in suburb and cloud area. The subfigures (a–c) represent the suburb with small buildings, the suburb with big buildings and the image with cloud, respectively. The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, the 10 m result of the U-Net model and the 2.5 m result of the FSRSS-Net model, respectively.](image)

**Figure 21.** The building extraction results of images in suburb and cloud area. The subfigures (a–c) represent the suburb with small buildings, the suburb with big buildings and the image with cloud, respectively. The subfigures (1), (2) and (3) represent the 10 m Sentinel-2A image, the 10 m result of the U-Net model and the 2.5 m result of the FSRSS-Net model, respectively.

5.4. Discussion

In order to further investigate the characteristics of 2.5 m building recognition results of the FSRSS-Net, we discussed two situations compared with 2.5 m results in this subsection. The first one is the U-Net training with 2 m high-resolution images from GF2 satellite and 2 m building ground-truth. The second one is two deconvolution layers are directly added at the end of U-Net, that is, super-resolution semantic segmentation with deconvolution of high-level semantic features.

5.4.1. Comparison between the 2.5 m Results of FSRSS-Net and the 2 m Recognition Results from GF2 Image

The U-Net in Figure 6 is also trained with high-resolution images, e.g., 2 m-resolution GF2 images for generating the 2 m building extraction results. Specifically, the 2 m-resolution images of GF2 and the 2 m ground-truth of the same regions are used to train the U-Net model. After 400 epochs of training, the training accuracy reaches 94% when the model converges and 89% on the test samples. As shown in Figure 22 (a1) and (b1), the two small areas in Kunming represent cities and suburbs, respectively. In the subfigures (a1), (b1), (a3) and (b3) in Figure 22, there are the Sentinel-2A images and building identification results, achieved by using the FSRSS-Net. And the subfigures (a4), (b4), (a5) and (b5) in Figure 22 is GF2 images and the predicted 2 m resolution building results by U-Net. It can be seen that the results of 2.5 m are obviously better than those of 10 m, but slightly worse than those of 2 m. In the results of 2 m, the outline of the building is more accurate, and the edge of the house is straight, while in the results of 2.5 m, the edge of the house is a little smooth. This is because the information contained in the 10 m image and the 2.5 m ground-truth must be less than that of the 2 m image and the 2 m ground-truth.
Figure 22. The comparison between 10 m, 2.5 m results from Sentinel-2A images and 2 m results from GF2 images. The subfigures (a, b) represent the urban area in Kunming and the suburb area in Kunming. The subfigures (1), (2), (3), (4) and (5) represent the 10 m Sentinel-2A image, 10 m result of U-Net model, 2.5 m result of FSRSS-Net model, 2 m GF2 image and 2 m result of U-Net model, respectively.

5.4.2. Comparison between FSRSS-Net and U-Net+2SR

As stated in Section 3.1, the unique structure of the FSRSS-Net network is to integrate both the up-sampling of primary image features and the up-sampling of high-level semantic features. In order to further validate its rationality, we design the network structure as shown in Figure 23. Two deconvolution layers and three convolution layers are added at the end of the U-Net variant, and the final classification layer still uses the softmax. Based on the U-Net variant, the network is deconvoluted and up-sampled twice. It is called “U-Net+2SR” in the following, since there are two deconvolution layers after U-Net. There is no up-sampling of primary image features in the “U-Net+2SR” network, which is the only difference from the FSRSS-Net. The same 8597 samples are used to train the U-Net+2SR network. In addition, the training hyper-parameters and loss functions are also the same as the FSRSS-Net. In addition, the total training epoch is 400. After about the 220th epoch, the training accuracy of the model is stable at 0.893, which is 0.014 lower than that of the FSRSS-Net model.
Taking Shanghai as an example, the trained U-Net+2SR model is used to identify 2.5 m buildings, which is compared with the results of FSRSS-Net. In order to make a clearer comparison, three small areas with different kinds of shapes and densities of buildings are selected and shown in Figure 24. In the three small areas, the results of FSRSS-Net are significantly better than those of U-Net+2SR. As shown in the yellow circle, the results of U-Net+2SR have obvious missed detection and misclassification. The U-Net+2SR model mistakenly identifies multiple backgrounds as buildings, and cannot distinguish the neighboring buildings accurately.

Figure 24. Comparison of the extraction results of FSRSS-Net and U-Net+2SR, taking Shanghai as an example (three small areas (a–c) are selected for detailed comparison).
6. Conclusions

Motived by the idea of image super-resolution technologies and CNN semantic segmentation, this paper proposes a super-resolution semantic segmentation network, i.e., FSRSS-Net, which makes full use of the rich details in the high-resolution ground-truth, and aims to learn the corresponding relationship between the middle-resolution images and the high-resolution semantic segmentation maps. The main experimental results in this paper are as follows: (1) There is little difference between the accuracy of 10 m results by U-Net and 2.5 m results by FSRSS-Net. However, the accuracy of 10 m results interpolation up-sampling to 2.5 m is generally lower than the accuracy of 10 m results and the accuracy of 2.5 m results. (2) In the 12 cities for generalization ability evaluation, the recognition results of buildings are pretty good, indicating that the migration and generalization performance of the models are strong. (3) Comparing 2.5 m results of FSRSS-Net with the results of 2 m-resolution GF2 images, the results of 2.5 m are obviously better than those of 10 m, but slightly worse than those of 2 m. Finally, two main conclusions can be drawn from the above experimental results: (1) Based on the 10 m Sentinel-2A image, the FSRSS-Net can achieve super-resolution semantic segmentation and produce 2.5 m building recognition result. The 2.5 m building recognition results are better than the 10 m building recognition results interpolation up-sampling to 2.5 m. (2) The learned FSRSS-Net can be also well generalized to other cities that are far from training regions.

The research of this paper is an important step in super-resolution semantic segmentation for satellite mapping of buildings, which would be used as a reference for high-resolution mapping of buildings from middle-resolution satellite images. In this paper, a large number of samples are produced automatically based on existing images and online data, which is of great significance to improve the efficiency and effectiveness of building extraction, and can save a lot of manpower and time.

Although there are some innovative research contents and achievements in this paper, there are still some shortcomings and limitations. The following two points need to be explored and studied: (1) In the super-resolution semantic segmentation model-FSRSS-Net, what is the role of higher-resolution building ground-truth in training model? How the mapping relationship between 10 m image and 2.5 m ground-truth learned from the model be understood and visualized? (2) If we design a network with multiple deconvolution layers, when the resolution of image and semantic segmentation results are different many times, such as 10 m image and 1 m building ground-truth, can we still extract buildings better?

According to the shortcomings and limitations of this paper, further research can be carried out in the following aspects in the future: (1) To apply the FSRSS-Net model to global building extraction, we need to select samples from typical regions of other countries in order to train models with strong generalization performance and global applicability. (2) The network with multiple deconvolution layers is designed, and how to train and optimize the model needs to be studied, so as to realize whether the model can better extract buildings when the resolution of image and semantic segmentation results are often different.

Author Contributions: Conceptualization, H.T.; funding acquisition, H.T.; investigation, C.J. and T.Z.; methodology, T.Z.; software, T.Z. and P.X.; and supervision, H.T., Y.D. and P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2017YFB0504104) and the National Natural Science Foundation of China (No. 41971280).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.
25. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In European Conference on Computer Vision; Springer: Cham, Switzerland, 2014; pp. 184–199. [CrossRef]
26. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. Comput. Sci.-CVPR 2016, 391–407. [CrossRef]
27. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. arXiv 2016, arXiv:1609.05158.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. IEEE CVPR 2016. [CrossRef]
29. Kim, J.; Lee, K.J.; Lee, M.K. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. 2016 IEEE CVPR 2016. [CrossRef]
30. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.H.; et al. Photo-realistic single image super-resolution using a generative adversarial network. arXiv 2016, arXiv:1609.04802.
31. Pathak, H.N.; Li, X.; Minaee, S.; Cowan, B. Efficient Super Resolution for Large-Scale Images Using Attentional GAN. arXiv 2018, arXiv:1812.04821.
32. Mustafa, A.; Khan, S.H.; Hayat, M.; Shen, J.B.; Shao, L. Image Super-Resolution as a Defense Against Adversarial Attacks. arXiv 2019, arXiv:1901.01677. [CrossRef]
33. Gargiulo, M. Advances on CNN-based super-resolution of Sentinel-2 images. arXiv 2019, arXiv:1902.02513.
34. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. Acta Ecol. Sin. 2015, 28, 627–635. [CrossRef]
35. Luc, P.; Couprie, C.; Chintala, S.; Verbeek, J. Semantic Segmentation using Adversarial Networks. arXiv 2016, arXiv:1611.08408.
36. Evo, I.; Avramovi, A. Convolutional neural network based automatic object detection on aerial images. IEEE Geoscience Remote Sens. Lett. 2016, 13, 740–744. [CrossRef]
37. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. IEEE Trans. PAMI 2014, 39, 640–651. [CrossRef]
38. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. ICCV 2015, arXiv:1505.04366.
39. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. PAMI 2017, 39, 2481–2495. [CrossRef][PubMed]
40. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. PAMI 2018, 40, 834–848. [CrossRef][PubMed]
41. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. arXiv 2015, arXiv:1505.04597.
42. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. arXiv 2013, arXiv:1311.2901.
43. Zeiler, M.D.; Krishnan, D.; Taylor, G.W.; Fergus, R. Deconvolutional networks. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
44. Zeiler, M.D.; Taylor, G.W.; Fergus, R. Adaptive deconvolutional networks for mid and high level feature learning. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
45. Fisher, Y.; Vladlen, K. Multi-scale context aggregation by dilated convolutions. arXiv 2015, arXiv:1511.07122.
46. Zhang, T.; Tang, H. Evaluating the generalization ability of convolutional neural networks for built-up area extraction in different cities of china. Optoelectron. Lett. 2020, 16, 52–58. [CrossRef]
47. Zhang, T.; Tang, H. A Comprehensive Evaluation of Approaches for Built-Up Area Extraction from Landsat OLI Images Using Massive Samples. Remote Sens. 2019, 11, 2. [CrossRef]