Article

Super-Resolution Rural Road Extraction from Sentinel-2 Imagery Using a Spatial Relationship-Informed Network

Yuanxin Jia 1, Xining Zhang 2,3, Ru Xiang 2,3 * and Yong Ge 2,3,4,5,*

1 Academy of Forest Inventory and Planning, National Forestry and Grassland Administration, Beijing 100714, China; jiayx@lreis.ac.cn
2 State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; zhangxn@lreis.ac.cn (X.Z.); xiangru22@mails.ucas.ac.cn (R.X.)
3 University of Chinese Academy of Sciences, Beijing 100049, China
4 School of Land Engineering, Chang’an University, Xi’an 710064, China
5 Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China
* Correspondence: gey@lreis.ac.cn

Abstract: With the development of agricultural and rural modernization, the informatization of rural roads has been an inevitable requirement for promoting rural revitalization. To date, however, the vast majority of road extraction methods mainly focus on urban areas and rely on very high-resolution satellite or aerial images, whose costs are not yet affordable for large-scale rural areas. Therefore, a deep learning (DL)-based super-resolution mapping (SRM) method has been considered to relieve this dilemma by using freely available Sentinel-2 imagery. However, few DL-based SRM methods are suitable due to these methods only relying on the spectral features derived from remote sensing images, which is insufficient for the complex rural road extraction task. To solve this problem, this paper proposes a spatial relationship-informed super-resolution mapping network (SRSNet) for extracting roads in rural areas which aims to generate 2.5 m fine-scale rural road maps from 10 m Sentinel-2 images. Based on the common sense that rural roads often lead to rural settlements, the method adopts a feature enhancement module to enhance the capture of road features by incorporating the relative position relation between roads and rural settlements into the model. Experimental results show that the SRSNet can effectively extract road information, with significantly better results for elongated rural roads. The intersection over union (IoU) of the mapping results is 68.9%, which is 4.7% higher than that of the method without fusing settlement features. The extracted roads show more details in the areas with strong spatial relationships between the settlements and roads.

Keywords: spatial relationship; deep learning; super-resolution mapping; rural road extraction; feature enhancement

1. Introduction

Road information is essential geographic information that constitutes land use and land cover [1], playing an important role in emergency responses, traffic navigation, and urban–rural planning [2–4]. With the help of advanced Earth observation technologies, automatic road extraction from different types of remote sensing images is becoming increasingly dominant, saving people from time-consuming and laborious traditional road surveying methods [5]. For decades, the majority of road extraction research has been focused on using high-resolution (HR) or very high-resolution (VHR) satellite or aerial images to extract road information. Simler [6] suggested the use of RGB VHR aerial images of 0.5 m for building and road detection in dense urban areas, using a three-class support vector machine. Li et al. [7] proposed a hierarchical method based on a binary partition tree for urban road extraction and successfully applied it to a 0.5 m-resolution Pleiades-B image.
of Wuhan, China, and a 0.6 m-resolution Quickbird image of Enschede, the Netherlands. Zhu et al. [8] proposed a global context-aware and batch-independent network for complete and continuous road extraction and tested its performance on two public VHR road datasets and one of their own large-scale satellite image data with a resolution of 0.5 m. Xu et al. [9] utilized densely connected convolutional networks with local and global attention units for road extraction from a publicly available dataset composed of screenshots taken from Google Earth, and the spatial resolution of images in this dataset is 1.2 m. Tao et al. [10] adopted three datasets with resolutions of 0.5 m, 0.6 m, and 1.2 m for validating their spatial information inference net in road extraction.

However, the primary limitation of the utilization of HR or VHR images is that their utility in large-scale road extraction is greatly limited by their relatively small coverage, low revisit period, and expensive cost [11,12]. In addition, both the application scenarios for the above methods and most of the road extraction datasets are mainly concentrated on urban areas, ignoring rural areas. In fact, with the rapid development of the economy, the construction and renovation of roads are increasing, especially in some developing countries such as China [13]. Under the influence of poverty alleviation policies and rural revitalization strategies, road updates in impoverished rural areas are becoming frequent. Obtaining accurate road distribution is vital for government decision-making and management related to rural economic development [2,3]. Therefore, there is an urgent need for methods to monitor road changes and quickly update road information.

Considering the free availability and global scale coverage with a higher revisit cycle than HR or VHR images, recent researchers have turned to relying on Sentinel-2 images to detect roads [14], providing us with the opportunity to extract rural roads from Sentinel-2 imagery. Nevertheless, rural roads are narrow, generally less than 10 m in width, and can reach subpixel size in Sentinel-2 images, resulting in the mixed pixel problem [15]. This thoroughly increases the difficulty of the road extraction task.

Super-resolution mapping (SRM) is a downscaling technique and has long been an effective way of super-resolving the mixed pixel in coarse remote sensing images to obtain a finer scale classification map [16,17]. Under the rule of spatial correlation, super-resolution mapping can be converted into the problem of sub-pixel allocation, with the constraint of pixel class proportion [18–20]. Over the decades, various SRM models have been developed, such as the spatial attraction model [21], subpixel swapping model [22], object-based model [23], and the vectorial boundary-based model [24], and have successfully improved many applications of remote sensing information extraction, such as finer flood extraction [25], land use/land cover updating [26], and burned-area extraction [27].

In recent years, to more accurately describe the complex distributions of geographical objects, deep learning (DL) has emerged as an alternative approach for SRM from coarse remote sensing images. According to the manner of obtaining the finer classification map, there are two kinds of DL-based SRM. Inspired by the success of DL for single image super-resolution in the computer vision community, the first takes coarse resolution fraction images as the input and employs DL models such as the commonly-used convolutional neural network (CNN) to estimate the fine resolution indicator images for each class, and then the super-resolution classification map is produced by these finer indicator images via class-allocation algorithms [28–31]. In these methods, however, the coarser fraction of each class is produced by spectral unmixing, where the uncertainty can be inevitably propagated in the final result. In this context, the second takes the original coarse remote sensing imagery as input and jointly optimizes the spectral unmixing and SRM process [32–34]. In this way, more reasonable finer spatial patterns of geographic objects can be generated by establishing the complex nonlinear relationship between the input coarse imagery and the finer classification map [12]. This approach utilizes the original image as much as possible during optimization and avoids the uncertainty propagation from spectral unmixing to SRM [32], thus it has been attracting more attention in DL-based SRM. However, few DL-based SRM methods are suitable for rural road extraction research. This limitation arises mainly due to the fact that these DLSRM methods rely primarily on spectral features.
derived from remote sensing images, which act as the sole input source. Unfortunately, rural roads are often made of diverse materials such as asphalt, gravel, and dirt [3], making it challenging to distinguish them from their surroundings based solely on multi-spectral features. Furthermore, rural roads tend to be narrow, which makes their recognition even more difficult when using only medium-resolution Sentinel-2 imagery.

Very recently, there has been a growing focus on research concerning knowledge-guided DL for spatiotemporal statistical analysis and remote sensing information extraction [35–37]. Notably, Ge et al. [37] provided a comprehensive summary of geoscience knowledge and geoscience features that are advantageous in extracting valuable information from remote sensing images. This emerging field highlights the significance of incorporating domain-specific knowledge to enhance the capabilities of DL models in various remote sensing studies and applications, including tasks such as DL-based SRM. Since SRM is an information extraction technique using remote sensing images at a finer scale and belongs to a specific geographic application, there are various geoscience knowledge or geoscience features that can be used. Several researchers have already made successful explorations in this area. For example, Zhang et al. [38] proposed a two-stream network to hierarchically incorporate prior spatial transition features into image features which was proven to be an effective means for producing high-quality SRM results. He et al. [39] used the semantic information modulated module to parametrically incorporate the semantic prior into the feed-forward network to reinforce the representation of spatial context features.

Inspired by these studies, this paper proposes a spatial relationship-informed super-resolution mapping network (SRSNet) to extract rural roads from Sentinel-2 imagery. It needs to be clarified here that the rural roads in this paper refer to roads located in rural areas, typically used to connect farmland, villages, and other rural facilities. Highways and overpasses are not included. The construction of this SRSNet is based on the spatial relationship knowledge that rural roads often lead to rural settlements, and the co-existence relation of road and rural settlements, which is calculated through a proposed feature enhancement module, is used to strengthen the road-relevant features from the remote sensing images. Finally, fine-grained road maps at 2.5 m are generated from Sentinel-2 imagery. Two contributions are obvious: (1) A new rural road dataset is established with co-registered pairs of 10 m Sentinel-2 images and well-annotated rural road labels on 2.5 m Google Earth images and contains a total of circa 0.23 billion annotated pixels. This dataset fills the deficiency of the standard dataset in the rural road extraction research field. (2) A special DL model called the SRSNet is proposed for rural road extraction. In the SRSNet, a feature enhancement module is used to guide the feature selection related to rural roads by taking the relative position relation between roads and rural settlements. Experiments demonstrate that this is a promising rural road extraction method from medium-resolution remote sensing imagery which gives a new perspective for incorporating spatial relationship knowledge into DL models.

The remainder of the paper is organized as follows: Section 2 elaborates on the proposed method, Section 3 presents the experimental dataset and experimental parameters, Section 4 demonstrates the experimental results, followed by further discussion and analysis provided in Section 5, and finally, Section 6 provides some concluding remarks.

2. Method

2.1. Overview of the SRSNet

The SRSNet is a dual CNN model. One CNN network extracts features from remote sensing imagery, and the other CNN network extracts spatial features from rural settlement data. The architecture of the SRSNet is shown in Figure 1. The SRSNet mainly consists of four parts: BaseNet for extracting remote sensing image features, POINet for extracting settlement information, a feature enhancement module for fusing the deep features of images and settlements, and an up-sampling classification module for the super-resolving process and classification.
with a stride of 1, a batch normalization layer, and a rectified linear unit (ReLU) activation mechanism. The input LR (low-resolution) remote sensing image of the same color in the model mean that they are at the same level. SA denotes the spatial attention enhancement module; and features of images and settlements, and an up-sampling classification module for the super-resolving process and classification.

tracting settlement information, a feature enhancement module for fusing the deep features at different scales and optimize the feature extraction results. U-Net is a classical semantic segmentation network that has been widely used in various image-processing fields, including medical imaging [40], remote sensing [41], and others [42,43]. However, the model used in this paper is slightly different from the traditional U-Net structure. In addition to the classical encoder and decoder parts, a dilated convolution [44] part with expansion coefficients of 1, 2, and 3 is applied in this module to extract deep features at different scales and optimize the feature extraction results.

The encoder part uses a combination of three identical CB (the combination of a convolutional layer and a batch normalization (BN) layer) modules followed by a dropout layer [45] and max pooling layer, where the last CB module does not follow the max pooling layer. Each CB module consists of a convolutional layer using a $3 \times 3$ convolutional kernel with a stride of 1, a batch normalization layer, and a rectified linear unit (ReLU) activation function. The output size of each convolutional layer is kept consistent with the input. A CB module can be represented by Equation (1):

$$CB(x) = \text{BN}(\text{ReLU}(Wx + b), \gamma, \beta)$$  \hspace{1cm} (1)

where $x$ is the input feature map; $W$ and $b$ are the weight and bias, respectively; ReLU is the activation function; and $\text{BN}$ denotes the BN layer, where $\gamma$ and $\beta$ are the learnable parameters.

The middle part of the model is a group of dilated convolutions with expansion coefficients of 1, 2, and 3. The convolution kernel size, step size, and padding scheme are the same as the previous convolution in the encoder part. The features extracted by the different expansion factors are combined using concatenation. This component can be expressed by Equation (2):

$$\text{Dconv}(x) = \text{Con}(\text{ReLU}(Wx + b), d)$$ \hspace{1cm} (2)

Figure 1. The proposed SRSNet’s architecture. ① is BaseNet; ② is POINet; ③ is the feature enhancement module; and ④ is the up-sampling classification module. In BaseNet, blocks and layers of the same color in the model mean that they are at the same level. SA denotes the spatial attention mechanism. The input LR (low-resolution) remote sensing image $Y$ and the settlement data $P$ have dimensions of $W \times H \times C$ and $W \times H \times 1$, respectively, where $C$ is the number of bands. The size of the output HR (high-resolution) road map $X$ is $(W \times S) \times (H \times S)$, where $S$ is the scale factor. $f_{\text{max}}(\cdot)$ is the function used to obtain the final classification for each subpixel by selecting the maximum membership probability.

2.2. BaseNet

BaseNet adopts the classical U-Net framework to obtain higher-level feature representations. U-Net is a classical semantic segmentation network that has been widely used in various image-processing fields, including medical imaging [40], remote sensing [41], and others [42,43]. However, the model used in this paper is slightly different from the traditional U-Net structure. In addition to the classical encoder and decoder parts, a dilated convolution [44] part with expansion coefficients of 1, 2, and 3 is applied in this module to extract deep features at different scales and optimize the feature extraction results.
where $D_{conv}$ denotes dilated convolutional layer; $Conv$ denotes the concatenate operation; ReLU, $x$, $w$, and $b$ have the same meaning as before; and $d$ denotes the expansion factor.

The decoder part, which is for data scaling, uses two combinations of DB (the combination of a transposed convolutional layer and a BN layer), CB, and dropout layers, where an additional CB module is inserted into the middle of the second combination. The model still retains the classical skip connection, which is used to combine the feature maps from the corresponding scale of the encoding part and decoding part, and the combined features are inserted before each DB module. Taking the first module as an example, one DB module can be represented by Equation (3):

$$
DB(x) = BN(ReLU(W(SC(x, x_{ei})) + b)^T(s), \gamma, \beta)
$$

where $SC$ denotes the skip connection; $x$ is the current input feature map; $x_{ei}$ denotes the feature map of $i$th down-sampling in the encoder part; $T(s)$ denotes a transposed convolution operation with scale $s$; and the other parameters are the same as above.

### 2.3. POINet and Feature Enhancement Module

The POINet part is used to extract the deep features of the points of interest (POI)-based rural settlement data. The part uses three consecutive dilated convolution layers and each of them is followed by ReLU layers. The three dilated convolutional layers have expansion factors of 2, 4, and 8 and use a $3 \times 3$ convolutional kernel with a stride of 1. The size of the layer output is consistent with the input.

The feature enhancement module is used to enhance the deep image feature extracted by BaseNet through the introduction of settlement features extracted by POINet. The feature enhancement module applies the structure of the position attention module [46], which is shown in Figure 2. Firstly, the deep feature map of settlement data and remote sensing imagery are reshaped into two dimensions, where $N$ denotes $(W \times H)$ and an extra transpose is performed for the settlement features. Then, two reshaped feature maps are multiplied and normalized to between 0 and 1 by the Softmax function. The normalized result is restored to the same size as the input image feature using the reshape operation. Finally, it is superimposed with the deep feature map of the original image via the multiply operation to obtain the enhanced feature maps. The Softmax function is implemented through Equation (4).

$$
\text{Softmax}(\text{feature}(S, I)) = \frac{\exp(\text{feature}(S, I)_i)}{\sum_{i=1}^{N} \exp(\text{feature}(S, I)_i)}
$$

where $\text{feature}(S, I)$ denotes the extracted feature map after a multiply operation on the reshaped feature maps of settlement and image; $S$ and $I$ denote the feature map of the settlement and image, respectively; and $N$ denotes the number of the element in $\text{feature}(S, I)$.
2.4. Up-Sampling Classification Module

The up-sampling classification module focuses on the upscaling and final classification of the enhanced feature maps. The module employs two consecutive combinations of DB, CB, and dropout layers to achieve the mapping of low-resolution feature maps to high-resolution feature maps. In the first combination, a spatial attention mechanism [47] is first employed to further enhance the feature map, followed by a DB module to achieve a rise in scale. A CB module is then used, followed by a dropout layer, where the parameters of the DB and CB modules remain the same as above. In this paper, the scale factor is set to 2, and a high-resolution feature map with a four-fold increase in scale is obtained by scaling twice.

At the end of the network, a Softmax layer is used as the model’s classifier. This layer converts the high-resolution feature map into a probability distribution map, providing the probabilities of each class for every pixel in the feature map. The Softmax function is calculated in the same way as Equation (4). The final output size is \((W \times S) \times (H \times S) \times C\), where \(C\) represents the number of classes. After training the model, the classification result for each sub-pixel is obtained by selecting the class with the highest probability. This process is mainly achieved through Equation (5).

\[
C_{pi} = f_{\text{max}}(P^c_i), \quad \text{where} \begin{cases} 
  i = 1, 2, 3 \cdots (W \times S) \\
  c = 1, 2, 3 \cdots C
\end{cases}
\] (5)

where \(C_{pi}\) denotes the class of the \(i\)th subpixel in the feature map; \(f_{\text{max}}(\cdot)\) is used to obtain the class corresponding to the max probability value; and \(P^c_i\) denotes the probability value of subpixel \(i\) belonging to class \(c\).

3. Experiment

3.1. Overview of Methodology

The main goal of this research is to propose a super-resolution road mapping network that incorporates the relative position relation between roads and rural settlements. Figure 3 shows the flowchart of this study. First, the model is iteratively trained with the training data set to simulate the relationship between high-resolution road maps, images, and residential areas. The trained model is then evaluated on the test set. The evaluation of the model includes four aspects: the comparison of the road extraction and mapping results with other models; the ablation analysis of different modules in our model; the comparison of the super-resolution mapping results using different up-sampling methods; and the visualization analysis of the feature maps of different layers of the proposed model.

![Figure 3. Flowchart of the methodology.](image-url)
3.2. Study Area and Datasets

Chongyang County, once a poor county in Hubei Province, China, and its surrounding area are selected as the study area for this experiment (Figure 4). Chongyang County is located in southeastern Hubei Province, at the border of Hubei, Hunan, and Jiangxi Provinces, and has an average altitude of 400 m and is surrounded by mountains with a small basin in the middle. It is located in a subtropical monsoon climate zone with an average annual temperature of 18.3 degrees and an average annual rainfall of 1300 mm. The Sentinel-2 image map of the study area is shown in Figure 4.

![Figure 4. Sentinel-2 image map of Chongyang County.](image)

This experiment uses Sentinel-2 multi-spectral remote sensing data with a spatial resolution of 10 m and rural settlement distribution data. The Sentinel-2 image (10 m) includes four bands of red, green, blue, and near-infrared (NIR). In this study, two scenes of Sentinel-2 images from November 2017 within the study area were downloaded through the Copernicus Open Access Hub platform (https://scihub.copernicus.eu/dhus/#/home (accessed on 8 June 2023)) when vegetation occlusion will have less impact on the road, and the outline of the road will be clearer. The rural settlement distribution data are expressed by POI data which is provided by the Google Map Open Platform, and the POI-based settlement data covers all villages in the study area. The distribution of all settlements is shown in Figure 5a.

![Figure 5. Distribution of POI-based settlement data and label data patches. (a) The distribution of POI-based settlement data. (b) The distribution of label data patches.](image)
For remote sensing images, we mainly use Arcmap and ENVI software for preprocessing such as splicing and cloud removal. For the distribution data of rural settlements, we created a 50 m rectangular buffer and then converted the buffer into a raster so that the data of rural settlements could be input into the neural network.

The label data for this experiment are higher resolution road data, which is produced as follows: firstly, according to the distribution of roads and settlements, the high-resolution image is cropped into 881 image patches of size 128 × 128, which are consistent with the input Sentinel-2 image; their spatial location distribution is shown in Figure 5b. Then, all road annotation is carefully vectorized by QGIS on Google Earth images downloaded at 2.5 m spatial resolution, and the vectorized data are transformed into raster road data with a resolution of 2.5 m as label data for the experiment.

In this study, the actual dataset obtained contains 881 samples after pre-processing the image, settlement, and label data. In the experiment, four bands of the Sentinel-2 image are selected as part of the input to the model, with a count-in dimension of 128 × 128 × 4 and a resolution of 10 m per pixel, so that a single sample actually represents a spatial extent of 1280 m × 1280 m. The input dimension of the rural settlement data is 128 × 128 × 1. The labels of the model are one-hot encoded, and the output dimension is 512 × 512 × 2, corresponding to a resolution result of 2.5 m.

3.3. The Comparison Methods and Evaluation Metrics

To better illustrate the effectiveness of the proposed method, four existing DL-based SRM methods—SPMCNN-ESPCN [32], SRMCNN [33], CASNet [12], and SCNet [38]—are selected for comparison. SPMCNN-ESPCN is a very simple network that has a small model complexity and a number of parameters. CASNet is built based on a stack of global context attention modules embedded into stacked residual blocks to reinforce global representation. Both SRMCNN and SCNet are encoder–decoder structures, while SCNet has two branches and they originally extract features from remote sensing images and prior soft information, respectively. Here, the prior information is replaced by rural residential distribution for the purpose of this paper. For the above methods that do not provide source code, we rigorously reproduced these networks following the description of the network structure and parameters in the corresponding references. As for ablation studies focusing on analyzing the effects of rural settlement information and the feature enhancement operation, two methods were used as comparison methods in this study. The first one only uses features extracted from multi-spectral images via BaseNet and outputs the road extraction results using the up-sampling classification module. This approach is referred to as SRM\_CNN and its structure is shown in Figure 6a; the second approach uses a dual CNN model to extract rural settlement and multi-spectral image features separately. The two features are directly superimposed and up-sampled using the up-sampling classification module to obtain the super-resolution load map. This approach of directly overlaying POI data features is referred to as SRM\_CNN\_op, and its structure is shown in Figure 6b.

Since the road information occupies a relatively small area of the whole image, calculating global information in the evaluation will reduce the objectivity of the road extraction assessment results. Given that road extraction is essentially a binary classification problem, the producer’s accuracy (PA) and the user’s accuracy (UA) of the road are considered as two evaluation indices in this study. In addition, this study adds the intersect over union (IoU) [48], a commonly used image segmentation evaluation metric in the field of computer vision, as an evaluation criterion, which is calculated as shown in Equation (6):

\[
\text{IoU} = \frac{R_p \cap R_r}{R_p \cup R_r}
\]

where \(R_p\) denotes the road pixel region predicted by the model and \(R_r\) is the actual reference road pixel region. The IoU uses a ratio value calculation, which satisfies non-negativity and homogeneity. So, it is insensitive to the scale of the object and it better reflects the shape consistency of the predicted results with the reference results.
The two features are directly superimposed and up-sampled using the up-sampling classification module to obtain the super-resolution load map. This approach of directly overlaying POI data features is referred to as SRMCNN_op, and its structure is shown in Figure 6b.

Since the road information occupies a relatively small area of the whole image, calculating global information in the evaluation will reduce the objectivity of the road extraction assessment results. Given that road extraction is essentially a binary classification problem, the producer's accuracy (PA) and the user's accuracy (UA) of the road are considered as two evaluation indices in this study. In addition, this study adds the intersect over union (IoU) [48], a commonly used image segmentation evaluation metric in the field of computer vision, as an evaluation criterion, which is calculated as shown in Equation (6):

\[
\text{IoU} = \frac{R_p \cap R_r}{R_p \cup R_r}
\]

where \(R_p\) denotes the road pixel region predicted by the model and \(R_r\) is the actual reference road pixel region. The IoU uses a ratio value calculation, which satisfies non-negativity and homogeneity. So, it is insensitive to the scale of the object and it better reflects the shape consistency of the predicted results with the reference results.

Figure 6. Illustration of the two comparison methods.

3.4. Experimental Details

In this study, a total of 881 samples from the experimental dataset are used, of which 850 samples are used as the training data for the model and 31 samples are used as the test data for the model. The training data and test data are randomly obtained from the dataset. In the training set, 20% of the data are used for model validation. The model is trained on Tensorflow 2.1, operating on Windows 10, with a 2080ti 11 GB accelerating training. During the actual training, the weights and biases of the model are first generated randomly, and the model is optimized using Adam [49], with the learning rate set to 0.001 for all models except CASNet, which is set to 0.0001. The training batch size is set to 12, and the model undergoes 50 epochs. For the SRSNet and all of its comparisons, the IoU is employed as the loss function, which is calculated as shown in Equation (6). The metrics used to monitor the training process include the loss value of the model, the validation loss value, and the accuracy rate.

4. Experimental Results

4.1. Comparison with Other Methods

A visual comparison of the results of the SRSNet and the four existing methods is shown in Figure 7. It can be seen that rural road mapping is very difficult due to the elongated road shapes. SPMCNN-ESPCN can capture many traces of road distribution, but these traces are extremely scattered and discontinuous. The mapping results of SRMCNN and CASNet are clearer and more continuous, while SCNet is able to extract some narrow and short roads that are difficult to recognize due to the introduction of rural residential distribution, thus improving the road extraction effect. It is obvious that the SRSNet generates more accurate rural road extraction results than the other SRM methods, both in terms of identifying small-sized roads and maintaining road integrity and continuity. This is mainly attributed to the more effective integration of rural settlements that coexist with rural roads. The feature enhancement module in the SRSNet can guide image feature selection through the relative position relation between rural roads and settlements, rather than just using a simple combination of two features as SCNet does.
Remote Sens. 2023, 15, 4193

The mapping accuracy of the SRSNet is significantly better than the other four methods, which is consistent with the visual comparison in Figure 7. The SRSNet achieves the best values on all metrics, with improvements of up to 10%, 21.7%, and 19.5% on PA, UP, and IoU, respectively, compared to SPMCNN-ESPCN. The SRSNet also achieves 7.2%, 2.5%, and 7.3% amelioration on PA, UP, and IoU compared to SCNet, which also incorporates features from rural settlements but without feature selection and enhancement. Based on the above observations, it can be concluded that for the rural road SRM task, the integration of rural settlements is effective, and the feature fusion strategy is also crucial. The SRSNet leverages the feature enhancement module to reinforce the key features beneficial for rural road extraction, resulting in an impressive performance.

Table 1. The results of various indicators for rural road extraction using the three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA (%)</th>
<th>UA (%)</th>
<th>IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPMCNN-ESPCN</td>
<td>65.9</td>
<td>66.4</td>
<td>49.4</td>
</tr>
<tr>
<td>SRMCNN</td>
<td>73.7</td>
<td>72.1</td>
<td>57.4</td>
</tr>
<tr>
<td>CASNet</td>
<td>74.9</td>
<td>74.7</td>
<td>59.8</td>
</tr>
<tr>
<td>SCNet</td>
<td>68.7</td>
<td>85.6</td>
<td>61.6</td>
</tr>
<tr>
<td>SRSNet</td>
<td>75.9</td>
<td>88.1</td>
<td>68.9</td>
</tr>
</tbody>
</table>

4.2. Ablation Studies on the Feature Enhancement Module

Table 2 presents the results of road extraction using SRM\textsubscript{CNN}, SRM\textsubscript{CNN, op}, and SRSNet by comparison with the reference image at the pixel level. The CN indicates the number of correctly extracted road pixels, the MN represents the number of road pixels that are missed by the methods, and the EN represents the number of road pixels classified as errors in the results. From the table, it can be observed that the SRM\textsubscript{CNN} method has the lowest CN value, while the SRM\textsubscript{CNN, op} method has a higher CN value compared to SRM\textsubscript{CNN}, indicating that utilizing two branches of the network improves the capability of rural road extraction. Notably, the proposed SRSNet achieves the highest CN value, which suggests the effectiveness of the feature enhancement module in effectively utilizing the spatial knowledge of the relative position relation between the road and rural settlements for accurate rural road identification and extraction. In terms of the MN value, the SRSNet shows the lowest value compared to the other two methods. This further demonstrates its capability to capture road pixels effectively. However, it is worth noting that the EN value of the proposed method slightly increases compared to that of the SRM\textsubscript{CNN} method. This
implies that the SRSNet may occasionally misclassify some pixels as roads. Therefore, a comprehensive analysis and discussion are necessary to evaluate the overall effectiveness of the proposed method for rural road extraction.

Table 2. The results of various indicators for rural road extraction using the three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>CN</th>
<th>MN</th>
<th>EN</th>
<th>PA (%)</th>
<th>UA (%)</th>
<th>IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRM CNN</td>
<td>108556</td>
<td>24196</td>
<td>36436</td>
<td>74.9</td>
<td>81.8</td>
<td>64.2</td>
</tr>
<tr>
<td>SRM CNN_op</td>
<td>114732</td>
<td>18020</td>
<td>38484</td>
<td>74.9</td>
<td>86.4</td>
<td>67.0</td>
</tr>
<tr>
<td>SRSNet</td>
<td>117016</td>
<td>15736</td>
<td>37188</td>
<td>75.9</td>
<td>88.1</td>
<td>68.9</td>
</tr>
</tbody>
</table>

To evaluate the efficacy of the proposed method in rural road extraction, the PA, UA, and IoU are computed for a better evaluation of the three methods. As shown in Table 2, the SRSNet has the highest PA value. Compared with SRM CNN, the PA of the proposed method increased by 1%. SRM CNN_op has the same PA value as SRM CNN, but its UA value is about 5% higher than SRM CNN. The UA of the SRSNet is 88.1%, which is more than 6% higher than that of SRM CNN. The IoU index of the SRSNet method is the highest, about 5% higher than that of SRM CNN and about 2% higher than SRM CNN_op. This indicates that the spatial relationship knowledge considered can assist in the extraction of rural road information, and the proposed method can effectively use the feature of settlements extracted by POINet, which has a better result in extracting roads. Although Table 2 shows that the addition of settlement features results in some of the pixels being misclassified as roads, the proposed method has the highest index values across all metrics, indicating that fusing the settlement features can improve the overall accuracy of road extraction.

Figure 8 shows the Sentinel-2 images, the extracted road mapping results with 2.5 m resolution using the three methods, and the reference in three regions, in which the red areas represent roads. As shown in the first area, the model demonstrates a significantly better performance in extracting rural roads compared to the other methods. It can extract more information about the road and its extraction results are closer to the reference results. For example, it can be observed that the SRSNet significantly outperforms other methods in extracting the rural roads in the second and third areas. However, the extracted roads are fragmented and discontinuous compared to the reference image. This is because some rural roads in the original image are occluded by trees. From a pixel-level perspective, the pixel represents trees rather than roads, making it difficult for the model to recognize them as roads. For instance, in the top right corner of the fourth area, it can be observed from the original image that the road exhibits a fragmented state due to tree occlusion, directly limiting the model’s effectiveness in road extraction.

In addition, the widths of the rural road extracted by the three methods are a bit larger than the actual reference. This essentially means that the model classifies some off-road areas as roads. Although the SRSNet is good at extracting the rural roads, it also increases the results of misclassification to some extent as a result. Furthermore, the SRSNet may incorrectly identify some regions as roads. For example, in the middle part of the third region, it is clear that the model misclassifies linear structures as roads and misclassifies a higher number of pixels than the other two methods; all of these analyses are consistent with the results shown in Table 2. Overall, the SRSNet proposed in this paper demonstrates better recognition and extraction capabilities for roads.

4.3. Analysis of Improvements for Road Extraction with Settlement Information

In order to demonstrate the role of settlements and the effectiveness of the model in utilizing settlement information, three areas with a strong spatial relationship between roads and settlements are selected, and their original Sentinel-2 images, extracted road mapping results (2.5 m) of the three methods, and reference images are shown in Figure 9. The yellow square area in the remote sensing image indicates the settlements used in the study.
5.1. Comparison of Different Up-Sampling Methods

By comparing the results of SRMCNN and SRMCNN\(_{op}\), it can be observed that in some areas, the two methods incorrectly classify many small areas as roads, and their road mapping results appear fragmented. This is because some rural roads in the original image are occluded by trees. From a pixel-level perspective, the pixel represents trees rather than roads, making it difficult for the model to recognize them as roads. For instance, in the top right corner of the fourth area, it can be observed that the SRSNet approach outperforms the other two approaches. The SRSNet approach extracts richer road information, mainly because the model is able to make full use of the relationship between settlements and roads, thus making full use of the settlement features to enhance the model’s extraction of roads; therefore, the road extraction results of the SRSNet are better than those without or directly overlaying the settlement feature methods. However, as the settlement feature enhances the representation of the spatial relationship knowledge can be used as auxiliary information to assist in the extraction of the third region, it is clear that the model misclassifies linear structures as roads and settlements are selected, and their original Sentinel-2 images, extracted road map—indicates the settlements. On the other hand, by comparing the results of the SRSNet, SRMCNN, and SRMCNN\(_{op}\), respectively, where the green dashed box in (d) Mapping results of road extraction using SRMCNN\(_{op}\), and SRSNet, respectively. (e) The reference road map.

Figure 9. Comparison of the road mapping results in the vicinity of the settlement area. (a) The Sentinel-2 remote sensing image using RGB bands, where the yellow square area in the remote sensing image indicates the settlements. (b–d) Mapping results of road extraction using SRMCNN, SRMCNN\(_{op}\), and SRSNet, respectively, where the green dashed box in (a), (b) and (c) is the area near the settlements. (e) The reference road map.
By comparing the results of $SRM_{CNN}$ and $SRM_{CNN,op}$, it can be observed that in some areas, the road extraction results of $SRM_{CNN,op}$ are inferior to those of $SRM_{CNN}$. For example, $SRM_{CNN}$ extracts more road information in the second area marked by the green dashed line. This indicates that the extraction of road results using a direct overlay of the settlement features does not fully exploit the settlement information. The model is unable to capture the weight of the relationship between settlements and roads, resulting in settlements not playing a role as auxiliary information in some areas, and the results extracted by $SRM_{CNN,op}$ being inferior to those of $SRM_{CNN}$.

On the other hand, by comparing the results of the SRSNet, $SRM_{CNN}$, and $SRM_{CNN,op}$ in the vicinity of settlements (the green dashed area in the figure), it can be clearly observed that the SRSNet approach outperforms the other two approaches. The SRSNet approach extracts richer road information, mainly because the model is able to make full use of the spatial knowledge of the relative position relation between the road and rural settlements. There is a neighboring spatial relationship between settlements and roads, and spatial relationship knowledge can be used as auxiliary information to assist in the extraction of road information. However, the direct overlay of settlement information ignores this. In contrast, the feature enhancement module in the SRSNet is able to better capture the relationship between settlements and roads, thus making full use of the settlement information to enhance the model’s extraction of roads; therefore, the road extraction results of the SRSNet are better than those without or directly overlaying the settlement feature methods. However, as the settlement feature enhances the representation of the relevant features, the model can also incorrectly extract some areas as roads. For example, in the green dashed box in the second area, it can be seen that the enhanced road information from the surrounding rural settlements increases the information extracted by the model for roads, but in the area above it, the model also identifies some linear grounds as roads.

5. Discussion

5.1. Comparison of Different Up-Sampling Methods

To further analyze the performance of the proposed method in super-resolution mapping, we compare the SRSNet with the classic up-sampling methods, such as the nearest and the bilinear interpolation methods. Comparison experiments are conducted by replacing the Conv2DTranspose layer in the SRSNet with the nearest interpolation layer and bilinear interpolation layer. A comparison of the mapping results in three areas is shown in Figure 10. It can be seen that the results of the nearest interpolation and bilinear interpolations are generally worse than the SRSNet. The two methods incorrectly classify many small areas as roads, and their road mapping results appear fragmented. This is because both up-sampling methods take into account the smoothing effect of the image. When there are some small linear features in an image, these features may be smoothed into larger regions after up-sampling and, thus, are misclassified. Secondly, the two methods have an insufficient ability to extract road features, and the discontinuity of the road is strong, such as the road on the left side of the second area and the road near the river in the third area. This is because after the up-sampling layer is replaced by the nearest interpolation and bilinear interpolation, the trainable parameters of the model become smaller than before, and the performance of the model decreases, so the ability of the model to capture road features decreases. Compared with the bilinear interpolation, there are more small areas misclassified as roads in the road mapping results of the nearest up-sampling method, which can be seen in the top part of the mapping results of the three areas. This is because the nearest method only considers the nearest neighbor pixel value, which cannot capture the details of the image and easily amplifies the noise. This may incorrectly assign background pixels as roads in some areas.
because both up-sampling methods take into account the smoothing effect of the image. When there are some small linear features in an image, these features may be smoothed into larger regions after up-sampling and, thus, are misclassified. Secondly, the two methods have an insufficient ability to extract road features, and the discontinuity of the road is strong, such as the road on the left side of the second area and the road near the river in the third area. This is because after the up-sampling layer is replaced by the nearest interpolation and bilinear interpolation, the trainable parameters of the model become smaller than before, and the performance of the model decreases, so the ability of the model to capture road features decreases. Compared with the bilinear interpolation, there are more small areas misclassified as roads in the road mapping results of the nearest up-sampling method, which can be seen in the top part of the mapping results of the three areas. This is because the nearest method only considers the nearest neighbor pixel value, which cannot capture the details of the image and easily amplifies the noise. This may incorrectly assign background pixels as roads in some areas.

Table 3 shows the accuracy comparison of the mapping results of two up-sampling methods. It can be seen that the accuracy of PA, UA, and IOU using the nearest and bilinear up-sampling methods is lower than that of the SRSNet. The PA accuracy of the nearest is 73.9%, which is 2% lower than that of the SRSNet but higher than that of bilinear by 3.3%. This indicates that the road pixels extracted by the nearest up-sampling are larger than that of the bilinear. This may be because the nearest algorithm itself considers nearby pixels for up-sampling. This not only increases the number of pixels of road features but also increases the number of pixels of non-road features. Therefore, its UA is 4.3% lower than that of the bilinear method. At the same time, the use of nearest up-sampling increases the data noise, making the IOU of the result lower than that of the bilinear, which is consistent with the previous analysis results.

Table 3. The results of various indicators for rural road extraction using the three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA (%)</th>
<th>UA (%)</th>
<th>IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRSNet-nearest</td>
<td>73.9</td>
<td>74.2</td>
<td>58.7</td>
</tr>
<tr>
<td>SRSNet-bilinear</td>
<td>70.6</td>
<td>78.5</td>
<td>59.2</td>
</tr>
<tr>
<td>SRSNet</td>
<td>75.9</td>
<td>88.1</td>
<td>68.9</td>
</tr>
</tbody>
</table>

5.2. Visualization Analysis of Feature Maps at Different Layers

In this section, we compare the proposed SRSNet with the other two methods in terms of road feature extraction. We make a comparison using Grad-CAM [50] by visualizing the features extracted by the three methods. The results are shown in Figure 11.
Table 3. The results of various indicators for rural road extraction using the three methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>PA (%)</th>
<th>UA (%)</th>
<th>IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRSNet-nearest</td>
<td>73.9</td>
<td>74.2</td>
<td>58.7</td>
</tr>
<tr>
<td>SRSNet-bilinear</td>
<td>70.6</td>
<td>78.5</td>
<td>59.2</td>
</tr>
<tr>
<td>SRSNet</td>
<td>75.9</td>
<td>88.1</td>
<td>68.9</td>
</tr>
</tbody>
</table>

5.2. Visualization Analysis of Feature Maps at Different Layers

In this section, we compare the proposed SRSNet with the other two methods in terms of road feature extraction. We make a comparison using Grad-CAM [50] by visualizing the features extracted by the three methods. The results are shown in Figure 11.

Figure 11. Grad-CAM visualization of the feature maps of the different layers of the three methods. (a) The original Sentinel-2 image overlay with settlements. (b) The road mapping of the reference. (c) The feature map of the layer extracted by BaseNet. (d) The feature map of the layer after fusing features from settlements, which is the feature map of SRMCNN before up-sampling. (e) The feature map of the last convolutional layer. (f) The weighted superimposition of the feature map of the last convolutional layer on the original image.

It can be seen from Figure 11c that these methods do not capture the features of narrow roads, only the features of wide roads. The outline of the road can be seen in the SRM\textsubscript{CNN} feature map, but its significance is far less than that of SRM\textsubscript{CNN, op} and SRSNet. Compared with SRM\textsubscript{CNN, op}, the boundary of the wide road features extracted by the SRSNet is clearer. Figure 11d compares the feature maps of the convolutional layers of the different methods after the inclusion of the settlement data. Compared with Figure 11c, the road features extracted by the three methods are more prominent and the pixels of interest to the models are further highlighted. However, the feature maps extracted by SRM\textsubscript{CNN} did not change much. In the feature maps of SRM\textsubscript{CNN, op} and SRSNet, the features of rural roads are further emphasized. This indicates that the models pay more attention to the features of rural roads after the inclusion of settlement information. However, compared with SRM\textsubscript{CNN, op}, SRSNet extracts rural road features more clearly. For example, the red dashed section in Figure 11d. As seen in Figure 11a, these areas are mainly settlement locations. This indicates that the SRSNet is more effective in using the settlement information and thus focusing on the characteristics of rural roads. Therefore, in Figure 11e, the SRSNet extracts richer road information and is closer to the actual solution results, while SRM\textsubscript{CNN}
and SRM_{CNN,op} extract less road information and the accuracy of their mapping results is lower than that of the SRSNet. This is consistent with the analysis results in Section 4.2.

5.3. Limitations of the Study

This study effectively improves road extraction accuracy by fusing the deep features of images and settlements, which effectively demonstrates the potential of settlements in assisting road extraction and super-resolution mapping. However, there are still some limitations. Firstly, it can be seen from the previous analysis results that the model-extracted roads have discontinuities due to tree occlusion, which is mainly the effect of the image itself. This phenomenon is most obvious in the extracted rural roads. Secondly, the model still misclassifies some areas into roads, which is mainly reflected in the edge areas of the rural roads extracted by the model. Moreover, the width of the rural roads extracted by the model is much wider than the actual roads due to the influence of mixed image elements. Lastly, the study area is Chongyang County and its surrounding areas. Although a lot of time was spent on preprocessing, such as the vectorization of settlements as well as roads, the sample size of the final datasets obtained that were usable in the experiment was small. Therefore, in subsequent studies, it is necessary to select higher resolution images to further expand the sample size and use other datasets to improve the accuracy and robustness of the model.

6. Conclusions

In this study, we proposed a super-resolution road mapping network model that incorporates the relative position relation between roads and rural settlements. We selected Chongyang County and its surrounding areas as the study area and constructed the model dataset using Sentinel-2 images. We evaluated their road mapping accuracy using three metrics: PA, OA, and IoU, and we focused on comparing and analyzing the SRSNet with other models and its ablation for road mapping results and further analyzed the road extraction results in areas near settlements to explore the role of settlement information in road extraction. Furthermore, we discussed the performance of different up-sampling methods on road super-resolution mapping. Finally, we analyzed the changes in the feature maps before and after fusing the settlement features using the Grad-CAM method to explore the importance of settlement information on road extraction from the model perspective.

The above analysis demonstrates the potential of spatial relationship knowledge in assisting road extraction and super-resolution mapping. The super-resolution road mapping model incorporating settlement information can extract road information more comprehensively than traditional road extraction methods. The model can capture more road details, improving the accuracy of road mapping. Its mapping result has an IoU of 68.9%, which is 4.7% higher than the method which does not integrate settlement information. However, there are certain shortcomings in the study due to objective factors, such as the influence of tree occlusion and the small sample size of the datasets. Therefore, subsequent studies should consider factors such as the images themselves and select appropriate images. Secondly, the scope of the study should be expanded, and a large-capacity dataset should be constructed to improve the generalization ability and robustness of the model. In addition, other geological knowledge or features, such as slope data, can be incorporated into the model to enhance the sensitivity of the model to capture road features.

Author Contributions: Conceptualization, Y.J. and Y.G.; methodology, Y.J. and Y.G.; software, X.Z. and R.X.; validation, Y.J., X.Z. and R.X.; formal analysis, Y.J. and X.Z.; data curation, Y.J.; writing—original draft preparation, Y.J. and X.Z.; writing—review and editing, Y.J. and X.Z.; visualization, R.X.; supervision, Y.G.; funding acquisition, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation for Distinguished Young Scholars of China (Grant No. 41725006) and two Key Programs of the National Science Foundation of China (Grant No. 41531174 and No. 41531179).

Acknowledgments: We would like to thank Copernicus Open Access Hub for providing the Sentinel 2 MSI images and the Google Brain for providing the Tensorflow2.1 framework.

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

References

3. Liu, J.; Qin, Q.; Li, J.; Li, Y. Rural Road Extraction from High-Resolution Remote Sensing Images Based on Geometric Feature Inference. ISPRS Int. J. Geo-Inf. 2017, 6, 314. [CrossRef]
13. Li, J.; Qin, Q.; Han, J.; Tang, L.A.; Lei, K.H. Mining Trajectory Data and Geotagged Data in Social Media for Road Map Inference. Trans. GIS 2015, 19, 1–18. [CrossRef]
18. Wang, Q.; Zhang, C.; Tong, X.; Atkinson, P.M. General solution to reduce the point spread function effect in subpixel mapping. Remote Sens. Environ. 2020, 251, 112054. [CrossRef]


34. Arun, P.; Buddhiraju, K.; Porwal, A. CNN based sub-pixel mapping for hyperspectral images. Neurocomputing 2018, 311, 51–64. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.