remote sensing

MDPI

Article

DA-CapsUNet: A Dual-Attention Capsule U-Net for Road Extraction from Remote Sensing Imagery

Yongfeng Ren [1,*], Yongtao Yu [2] and Haiyan Guan [3]

1 School of Energy and Power Engineering, Nanjing Institute of Technology, Nanjing 211167, China
2 Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian 223003, China; allennessy@hyit.edu.cn
3 School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; guanhy.nj@nuist.edu.cn
* Correspondence: yfren@njit.edu.cn

Received: 2 August 2020; Accepted: 2 September 2020; Published: 4 September 2020



Abstract: The up-to-date and information-accurate road database plays a significant role in many applications. Recently, with the improvement in image resolutions and quality, remote sensing images have provided an important data source for road extraction tasks. However, due to the topology variations, spectral diversities, and complex scenarios, it is still challenging to realize fully automated and highly accurate road extractions from remote sensing images. This paper proposes a novel dual-attention capsule U-Net (DA-CapsUNet) for road region extraction by combining the advantageous properties of capsule representations and the powerful features of attention mechanisms. By constructing a capsule U-Net architecture, the DA-CapsUNet can extract and fuse multiscale capsule features to recover a high-resolution and semantically strong feature representation. By designing the multiscale context-augmentation and two types of feature attention modules, the DA-CapsUNet can exploit multiscale contextual properties at a high-resolution perspective and generate an informative and class-specific feature encoding. Quantitative evaluations on a large dataset showed that the DA-CapsUNet provides a competitive road extraction performance with a precision of 0.9523, a recall of 0.9486, and an F-score of 0.9504, respectively. Comparative studies with eight recently developed deep learning methods also confirmed the applicability and superiority or compatibility of the DA-CapsUNet in road extraction tasks.

Keywords: road extraction; capsule network; capsule U-Net; capsule attention network; remote sensing imagery; deep learning

1. Introduction

Periodically monitoring and updating the transportation infrastructures is a routine work so as to facilitate the smoothness and guarantee the security of the transportation-related activities. As a typical transportation infrastructure, roads play a significant and irreplaceable role in providing the passage for vehicles and pedestrians. The types and the coverage densities of the road network within an area reflect the population density, the living environment, and the economic development level. In addition, the road network information also provides essential ingredients to a number of applications, such as map navigation, land use planning, traffic flow monitoring, etc. The coverage rate, the network structure, and the condition of roads greatly affect the convenience, security, and behaviors of the related activities. Therefore, periodically and accurately updating the road network database is significantly favorable for making rational decisions, conducting efficient route planning, and directing correct driving behaviors. However, as a result of road maintenance, road constructions, and the influences of natural factors or traffic accidents, the road network is always changing. The changes

in the roads will definitely lead to information incompleteness or inaccuracy of the current road network database. Traditionally, road network information updating is usually carried out based on field investigations by well-trained workers or using mobile mapping systems mounted with video cameras/laser scanners and global navigation satellite system (GNSS) antennas. However, such means are quite inefficient, labor-intensive, and difficult to rapidly provide for the up-to-date road network database. As it requires a long time and considerable labor to collect the road data covering an entire state or a whole country. The huge amount of data collected by the mobile mapping systems also require considerable computation overhead. Sometimes, the measurement accuracy cannot be well guaranteed due to the incomplete coverage of some road sections or the inoperable conditions of some road segments.

In recent decades, the rapid development of optical remote sensing techniques has increased great potentials and convenience to a large variety of applications ranging from intelligent transportation systems, environment monitoring to disaster analysis based on remote sensing images. As the optical remote sensing sensors upgrade increasingly in resolutions, flexibilities, and qualities, it is quite efficient and cost-effective with a small amount of time and manpower expenditures to collect high-resolution, high-quality remote sensing images covering large areas of the earth surface by using satellite sensors or unmanned aerial vehicle (UAV) systems. The size of the resultant remote sensing images is remarkably smaller compared to that of the mobile mapping systems, thereby causing less computation overhead for processing these images. The up-to-date properties and the topologies of the roads can be well presented in the images. Thus, due to the advantageous properties of remote sensing images, they have been an important data source for assisting road network updating tasks. However, owing to the special bird-view image capturing pattern, the roads in remote sensing images often exhibit different levels of occlusions caused by the road-side high-rise objects (e.g., buildings and trees). In addition, the considerable topology variations of the roads in shapes, lengths and widths, the spectral and textural diversities of the roads caused by different road surface materials paved and varying illumination conditions, and the complex surrounding scenarios also affect the accurate identification and localization of the roads in the remote sensing images. Moreover, the shadows cast on the roads, the presence of on-road vehicles and pedestrians, and the painted road markings severely change the spectral and texture consistencies of the roads in the remote sensing images. Thus, it is still challenging to realize fully automated and highly accurate extraction of roads from remote sensing images. Exploiting advanced and high-performance techniques to further upgrade the road extraction efficiency and accuracy is greatly meaningful and urgently demanded by a wide range of applications.

The rapid development in deep learning techniques has burst a great number of breakthroughs on efficiency and accuracy in remote sensing image analysis tasks [1,2]. Compared to traditional feature descriptors, deep learning models have the superiorities of automatically exploiting multilevel, representative, and informative features in a fully end-to-end manner. Consequently, great attention and efforts have been paid to conduct road extraction from remote sensing images by using deep learning techniques [3]. Dai et al. [4] combined a multiscale deep residual convolutional neural network (MDRCNN) with a sector descriptor to detect roads. Specifically, the sector descriptor-based tracking connection was leveraged to improve the road integrity. Eerapu et al. [5] proposed a dense refinement residual network (DRR Net) for alleviating the class imbalance problems presenting in the remote sensing image-based road extraction tasks. To effectively handle multiscale roads in complex scenarios, Li et al. [6] designed a hybrid CNN (HCN) model by taking advantage of the multi-grained features. The HCN was composed of three parallel backbones for, respectively, providing the coarse-grained, medium-grained, and fine-grained road segmentation maps, which were finally fused to generate an accurate road segmentation result. Liu et al. [7] developed a multi-task CNN model, named RoadNet, to simultaneously extract road surfaces, edges, and centerlines. Tao et al. [8] put forward a spatial information inference net (SII-Net), which showed advantageous performance in inferring occlusions and preserving the continuities of the extracted roads. Senthilnath et al. [9] combined deep transfer learning with an ensemble classifier and proposed a road extraction framework

called Deep TEC. In this framework, the road segments extracted by the deep transfer learning models were refined to obtain the roads via the majority voting-based ensemble classifier. To make full use of multilevel semantic features, Gao et al. [10] designed a multiple feature pyramid network (MFPN). The MFPN integrated a feature pyramid module and a tailored pyramid pooling module for extracting, recalibrating, and fusing the multilevel features. Likewise, a fully convolutional network-based ensemble architecture, named SC-FCN, was used in [11] for solving the class imbalance issues. In addition, encoder-decoder networks [12–14] and U-Net architectures [15,16] were also exploited to fuse multilevel and multiscale features. To obtain accurate boundaries of roads, Zhang et al. [17] developed a multi-supervised generative adversarial network (MsGAN), which was jointly constructed using the spectral and topology features of the roads. In addition, to solve the problem of small-size samples, a lightweight GAN was designed in [18] for the road extraction task. Other deep learning models, such as the DeepWindow [19] and coord-dense-global (CDG) models [20] were also constructed for extracting roads. Despite the advanced achievements obtained by deep learning models, they still face the problems of requiring large-volume data for model training. The quality and the amount of the training data also significantly affect the robustness and the effectiveness of the designed deep learning models.

Recently, capsule networks have shown strong superiorities and advantageous performances on feature abstraction and representation capabilities. The concept of capsules was first pioneered by Hinton et al. [21]. Different from traditional scalar neuron-based deep learning models, capsule networks leverage vectorial capsule representations to characterize information-rich features [22]. Due to the superior feature encoding properties, capsule networks have been intensively exploited in many detection, classification, segmentation, and prediction tasks. Yu et al. [23] proposed a convolutional capsule network for detecting vehicles from remote sensing images. This network took the local image patches, which were generated based on a superpixel segmentation strategy, as the input to recognize the existence of vehicles. Gao et al. [24] designed a multiscale capsule network to conduct change detection from synthetic aperture radar (SAR) images. Specifically, a multiscale capsule module was employed to characterize the spatial relationship of features, and an adaptive fusion convolution module was applied to enhance the feature robustness. In addition, capsule networks were also exploited to recognize road markings [25] and traffic signs [26] from mobile light detection and ranging (LiDAR) data. Paoletti et al. [27] leveraged a spectral-spatial capsule network to perform hyperspectral image classification. Differently, Xu et al. [28] designed a multiscale capsule network with octave convolutions for classifying hyperspectral images. In this network, the octave convolutions adopted a parallel architecture and functioned to extract multiscale deep features. Yu et al. [29] proposed a hybrid capsule network for classifying land covers from multispectral LiDAR data. This network consisted of a convolutional capsule subnetwork and a fully connected capsule subnetwork for, respectively, extracting local and global features. LaLonde and Bagci [30] developed a fully convolutional capsule network, named SegCaps, to segment medical images. The SegCaps involved an encoder for extracting multiscale capsule features and a decoder for recovering a high-resolution feature representation to generate the segmentation map. Yu et al. [31] designed a capsule feature pyramid network (CapFPN) to segment building footprints from remote sensing images. In this network, the multiscale capsule features were integrated to improve both the localization accuracy and the feature semantics. Besides, interesting work for Chinese word segmentation from ancient Chinese medical books was also conducted in [32] by using capsule networks. Ma et al. [33] combined deep capsule networks with nested long-short term memory (LSTM) models to forecast traffic transportation network speeds. Specifically, the capsule network was used to extract the spatial features of traffic networks, and the nested LSTM structure was used to capture the hierarchical temporal dependencies. Liu et al. [34] proposed a multi-element hierarchical attention capsule network for predicting stocks. The attention module functioned to quantify information importance. In addition, capsule networks have also been exploited for image super-resolution [35], sentiment analysis [36], and machinery fault diagnosis [37] tasks.

In this paper, we propose a novel dual-attention capsule U-Net (DA-CapsUNet) for extracting road regions from remote sensing images. The DA-CapsUNet employs a capsule-based U-Net architecture, which involves a contracting pathway for extracting different-level and different-scale capsule features, and a symmetric expanding pathway for recovering a high-resolution, semantically strong, and informative feature representation for predicting the road region map. Specifically, a multiscale context-augmentation module and two types of feature attention modules are designed and integrated into the DA-CapsUNet to further enhance the quality and the representation robustness of the output features. The proposed DA-CapsUNet obtains a promising and competitive performance in processing remote sensing images containing roads of different conditions. The main contributions of this paper are listed as follows:

1. A deep capsule U-Net architecture is constructed to extract and aggregate multilevel and multiscale high-order capsule features to provide a high-resolution and semantically strong feature representation to improve the pixel-wise road extraction accuracy;
2. A multiscale context-augmentation module is designed to comprehensively exploit and consider the multiscale contextual information with different-size receptive fields at a high-resolution perspective, which behaves positively in enhancing the feature representation capability;
3. A channel feature attention module and a spatial feature attention module are proposed to force the network to emphasize the channel-wise informative and salient features and to concentrate on the favorable spatial features tightly associated with the road regions, thereby effectively suppressing the impacts of the background and providing a high-quality, robust, and class-specific feature representation.

The remainder of this paper is organized as follows: Section 2 presents the architecture of the proposed DA-CapsUNet. Section 3 reports the experimental results. Section 4 discusses the superiorities and drawbacks of the proposed DA-CapsUNet. Section 5 gives the concluding remarks.

2. Methodology

2.1. Capsule Network

Generally, traditional CNNs are built based on scalar neuron representations, which function to characterize the probabilities of the existence of specific features. Thus, in order to effectively capture the variances of entities, more extra neurons are often required to be incorporated to respectively encode the different variants of the same type of entity, thereby leading to the size and parameter expansions of the entire network. Differently, capsule networks leverage vectorial capsule representations as the basic neurons to characterize entity features. Specifically, the length of a capsule encodes the probability of the presence of an entity, and the instantiation parameters of the capsule characterize the inherent properties of the entity [22]. An important property is that the vectorial formulation allows a capsule not only to detect a feature but also to learn and identify its variants, thereby constituting a powerful, but lightweight, feature representation model.

Capsule convolution operations differ greatly from traditional convolution operations. Specifically, for a capsule j in a capsule convolutional layer, the total input to the capsule is a dynamically determined weighted aggregation over all the transformed predictions from the capsules within the convolution kernel in the previous layer as follows:

$$\boldsymbol{C}_j = \Sigma_i \, a_{ij} \cdot \boldsymbol{U}_{ij} \tag{1}$$

where $\boldsymbol{C}_j$ is the total input to capsule j; a_{ij} is a dynamically determined coupling coefficient indicating the degree of contribution of the prediction from capsule i in the previous layer; $\boldsymbol{U}_{ij}$ is the prediction from capsule i, which is computed as follows:

$$\boldsymbol{U}_{ij} = \mathbf{W}_{ij} \cdot \boldsymbol{U}_i \tag{2}$$

where $\boldsymbol{U}_i$ is the output of capsule i and $\mathbf{W}_{ij}$ is a transformation matrix acting as a feature mapping function. Particularly, the coupling coefficients between capsule i and all its associated capsules in the layer above sum to 1. In this paper, the coupling coefficients are determined by using the improved dynamic routing process [38], which functions effectively and stably for constructing a deep capsule network.

As for the length-based probability encoding mechanism, capsules with long lengths should cast high probability predictions; whereas capsules with short lengths should contribute less to the probability estimations. To this end, a "squashing" function [22] is designed as the activation function to normalize the output of a capsule. The squashing function is formulated as follows:

$$\boldsymbol{U}_j = \frac{\| \boldsymbol{C}_j \|^2}{1 + \| \boldsymbol{C}_j \|^2} \cdot \frac{\boldsymbol{C}_j}{\| \boldsymbol{C}_j \|} \tag{3}$$

where $\boldsymbol{C}_j$ and $\boldsymbol{U}_j$ are, respectively, the original input and the normalized output of capsule j. Through normalization, long capsules are shrunk to a length close to one to cast high predictions; whereas short capsules are suppressed to almost a zero length to provide fewer contributions.

2.2. Dual-Attention Capsule U-Net

To take advantage of the advantageous properties of capsule networks and the powerful features of attention mechanisms, in this paper, we construct a dual-attention capsule U-Net (DA-CapsUNet) architecture in order to improve the pixel-wise road extraction accuracy. As shown in Figure 1, the proposed DA-CapsUNet takes a remote sensing image as the input and outputs an identical-spatial-size road region map totally in an end-to-end manner. The architecture of the DA-CapsUNet involves a contracting pathway, an expanding pathway, and a set of skip connections. The contracting pathway functions to extract different-level and different-scale informative capsule features with a gradual decrement of feature map resolutions. The expanding pathway takes charge of gradually recovering a high-resolution and semantically strong feature representation to generate a high-quality road region map. The skip connections concatenate the feature maps selected from the contracting pathway to the feature maps having the same resolutions in the expanding pathway for enhancing the feature representation capability. In addition, to further improve the feature robustness and to concentrate on the class-specific feature encodings, we design a kind of context-augmentation module and two types of feature attention modules and properly integrate them into the DA-CapsUNet.

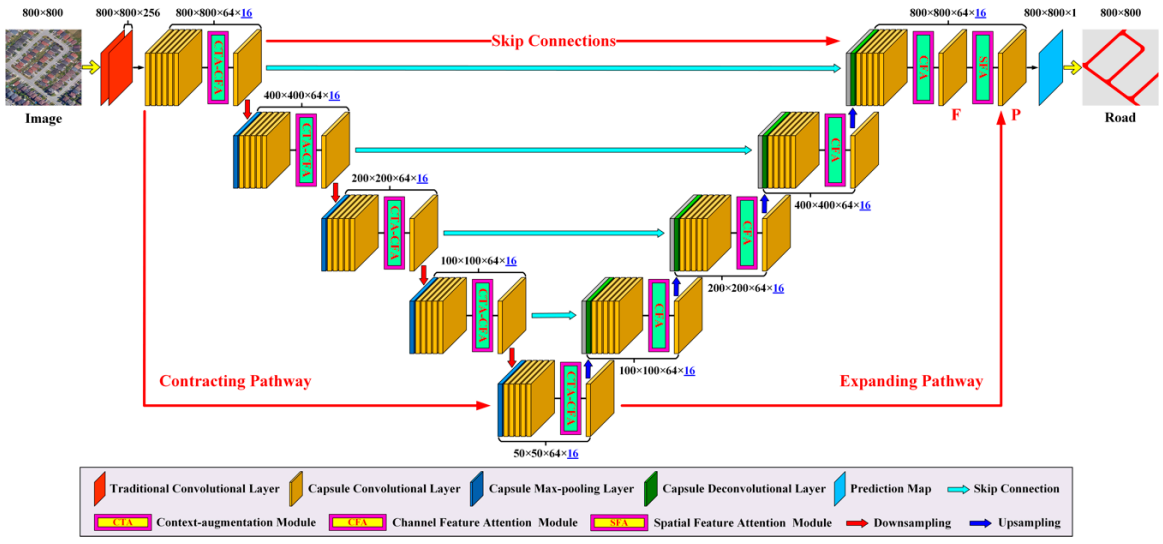


Figure 1. Architecture of the proposed dual-attention capsule U-Net (DA-CapsUNet).

The contracting pathway of the DA-CapsUNet is composed of two traditional convolutional layers for extracting low-level image features, and a chain of capsule convolutional layers and capsule pooling layers for extracting different scales of high-level capsule features. The scalar output of the second traditional convolutional layer is further converted into the vectorial capsule representation to form the primary capsule layer. This can be achieved through traditional convolution operations. Specifically, for the two traditional convolutional layers, the widely used rectified linear unit (ReLU) is adopted as the activation function.

As shown in Figure 1, the contracting pathway splits the capsule convolutional layers into five network blocks for computing capsule features at different resolutions and scales. Specifically, within each block, the feature maps in all the capsule convolutional layers maintain the same spatial resolution and size. Along the contracting pathway, the spatial size of the feature maps is gradually scaled down block by block to generate lower-resolution feature maps with a scaling step of two. This is achieved by the inclusion of the capsule pooling layers at the end of each block to conduct feature down-sampling to scale down a feature map to its half size. Here, we leverage max-pooling operations to highlight the most representative and salient features. Although the spatial resolution of the feature maps in each block is gradually decreased, higher-level features with larger receptive fields are generated. Theoretically, the deepest layer in each block has the strongest and most representative feature abstractions. Thus, for the contracting pathway, we pick the feature map of the deepest capsule layer in each block as the reference feature maps for feature augmentation.

In the contracting pathway, the receptive field of a capsule in each block is enlarged slowly layer by layer through capsule convolutions. Thus, to rapidly expand the receptive field of a capsule to encapsulate a large extent of contextual properties, we append a max-pooling layer at the end of each block to scale down the feature maps to a small size. However, after max-pooling, the feature details are partially damaged in the resultant lower-resolution feature maps. In addition, capsule convolution operations behave indiscriminately on all the channels of a feature map within the local receptive fields at each layer. The interdependencies and saliences among the channels are weakly exploited, which is not helpful enough to obtain robust and informative feature encodings. Therefore, in order to rapidly enlarge the receptive field of a capsule to include more contextual information without the loss of feature map resolutions and details, and explicitly exploit the interdependencies among the feature channels to highlight informative features, we design a multiscale context-augmentation (CTA) module and a channel feature attention (CFA) module, and connect them in a cascaded way at the end of each block to perform feature augmentation and feature recalibration, respectively (see Figure 1).

In this paper, we leverage the atrous convolutions [39] to design the CTA module. Different from standard convolutions, atrous convolutions involve an extra hyperparameter, named the atrous rate, which indicates the stride of the kernel parameters. A very important property of atrous convolutions is that, by adjusting the atrous rate, different-size receptive fields can be easily accessed without increasing the number of kernel parameters. Figure 2 illustrates three examples of a 3×3 atrous convolution kernel with different atrous rates, which correspond to different-size receptive fields. Therefore, taking into account the advantageous property, we adopt the atrous convolutions to design the multiscale CTA module aiming at introducing fewer parameters. As shown in Figure 3, the CTA module consists of four parallel branches performed on the input feature map to exploit different scales of contextual information with the gradual increment of the atrous rates. Concretely, the first branch applies a 3×3 atrous convolution with an atrous rate of 1 to include a small scope of context. The second and third branches perform a 3×3 atrous convolution with atrous rates of 2 and 3, respectively, to access a medium size of context. The last branch stacks two 3×3 atrous convolutions with atrous rates of 3 and 5, respectively, to encapsulate a large area of context. For each branch, a 1×1 atrous convolutional layer is connected to smooth the extracted features and modulate the number of feature channels. Finally, the features encoding different-scale contextual properties from the four branches, along with the original input features, are concatenated and further fused through a 1×1 atrous convolution to generate the output feature map.

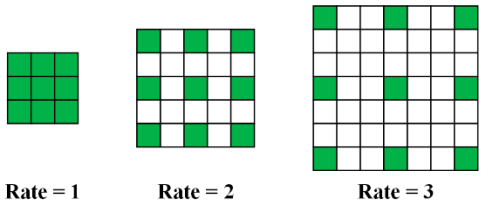


Figure 2. Illustration of the structure of a 3 × 3 atrous convolution kernel with different atrous rates of 1, 2, and 3 corresponding to different-size receptive fields.

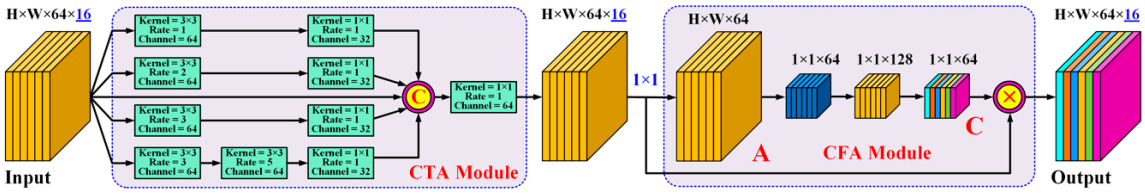


Figure 3. Architecture of the multiscale context-augmentation (CTA) module and the channel feature attention (CFA) module.

As shown in Figure 3, the output of the CTA module is directly fed into the CFA module. The objective of the inclusion of the CFA module is to upgrade the input features by explicitly exploiting the channel interdependencies to increase the ability of the network to emphasize informative features and weaken the contributions of the less salient ones with a global perspective. To this end, first, we apply a 1 × 1 capsule convolution operation on the input multi-dimensional capsule feature map to transform it into a one-dimensional capsule feature map **A**, which mainly encodes the feature probability properties of the input feature map. Then, to comprehensively take into consideration the channel-wise statistics from a global perspective, we perform a global average pooling operation on feature map **A** in a channel-wise manner to convert it into a channel descriptor. Formally, for each channel of feature map **A**, a scalar value is generated by averaging the lengths of the capsules in this channel via a global average pooling operation as follows:

$$v_i = \frac{1}{H \times W} \Sigma_j \| \boldsymbol{U}_j^i \| \tag{4}$$

where H and W are, respectively, the height and width of feature map **A**; $\boldsymbol{U}_j^i$ is the output of a capsule in the i-th channel of feature map **A**; v_i is the computed entry of the channel descriptor corresponding to the i-th channel of feature map **A**. Next, we further append two fully-connected layers to exploit channel-wise interdependencies in a non-mutually-exclusive manner. Specifically, the outputs of these two fully-connected layers are, respectively, activated using the ReLU and the sigmoid functions. The probability-encoded output (denoted by **C**) of the second fully-connected layer constitutes a channel-wise attention descriptor. Each entry of the attention descriptor **C** encodes the informativeness and the saliency of the corresponding feature channel of the input feature map. The attention descriptor **C** acts as a weight function to recalibrate the input feature map to highlight the contributions of the informative features. This is achieved by multiplying the attention descriptor **C** with the input feature map in a channel-wise manner as follows:

$$\overline{\boldsymbol{U}}_j^i = t_i \cdot \boldsymbol{U}_j^i \tag{5}$$

where t_i denotes the i-th entry of the attention descriptor **C**; $\boldsymbol{U}_j^i$ and $\overline{\boldsymbol{U}}_j^i$ are, respectively, the output of the original capsule in the i-th channel of the input feature map and the output of the corresponding recalibrated capsule in the i-th channel of the output feature map.

As shown in Figure 1, the expanding pathway is designed as a symmetric architecture to the contracting pathway and targets to gradually recover a high-resolution and semantically strong feature representation. Along the expanding pathway, the feature maps are gradually scaled up to a higher-resolution feature map block by block with a scaling step of two. This is achieved by adding a capsule deconvolutional layer at the end of each block to up-sample the feature map to its twice size. Particularly, after deconvolutions, to increase the feature representation robustness, the reference feature maps selected from the contracting pathway are concatenated to the up-sampled feature maps having the same spatial size in each block of the expanding pathway to conduct feature recovery through the skip connections. For the expanding pathway, we only integrate the CFA module at the end of each block to perform channel-wise feature recalibration to emphasize the informative and salient features.

In fact, as shown in Figure 1, the feature map $\mathbf{F}$ recovered by the expanding pathway encodes a general feature abstraction across the whole input image. The spatial features associated with the regions of the class of interest are not explicitly highlighted and the influences of the background features are not rationally suppressed, which is not powerful enough to obtain a high-quality road region map. In this regard, to further improve road extraction accuracy, we design a spatial feature attention (SFA) module over the feature map $\mathbf{F}$ to force the network to concentrate on the spatial features tightly associated with the road regions. The architecture of the SFA module is illustrated in Figure 4. Given the input feature map $\mathbf{F}\in\mathbf{R}^{H\times W\times 64\times 16}$ to the SFA module, first, we respectively apply two 1×1 capsule convolution operations to the input multi-dimensional capsule feature map to transform it into two identical-spatial-size one-dimensional capsule feature maps $\mathbf{B}\in\mathbf{R}^{H\times W\times 64}$ and $\mathbf{D}\in\mathbf{R}^{H\times W\times 64}$, where H and W are, respectively, the height and width of the input feature map $\mathbf{F}$. Then, we reshape feature map $\mathbf{B}$ to a feature matrix $\mathbf{G}\in\mathbf{R}^{N\times 64}$ and reshape feature map $\mathbf{D}$ to a feature matrix $\mathbf{E}\in\mathbf{R}^{64\times N}$, where $N = H\times W$ denotes the number of positions in the input feature map. Next, we perform a matrix multiplication operation between the feature map $\mathbf{G}$ and the feature map $\mathbf{E}$ (i.e., $\mathbf{GE}$) and apply a softmax function to each of the elements in the resultant matrix in a column manner to constitute a class region attention matrix $\mathbf{S}\in\mathbf{R}^{N\times N}$. Theoretically, the element $\mathbf{S}_{ij}$ at row i and column j of the class region attention matrix $\mathbf{S}$ evaluates the impact of position i on the position j in the input feature map, and it is computed as follows:

$$\mathbf{S}_{ij} = \frac{\exp\left(\Sigma_{k=1}^{64}\mathbf{G}_{ik}\mathbf{E}_{kj}\right)}{\Sigma_{m=1}^{N}\exp\left(\Sigma_{k=1}^{64}\mathbf{G}_{mk}\mathbf{E}_{kj}\right)} \tag{6}$$

where $\mathbf{G}_{ik}$ and $\mathbf{E}_{kj}$ are the elements at row i and column k of matrix $\mathbf{G}$ and the element at row k and column j of matrix $\mathbf{E}$, respectively. Afterward, we reshape the input feature map $\mathbf{F}$ to a capsule feature matrix $\mathbf{T}\in\mathbf{R}^{N\times 64\times 16}$ and multiply it by the class region attention matrix $\mathbf{S}$ (i.e., $\mathbf{ST}$) to produce a recalibrated capsule feature matrix. Finally, after reshaping the recalibrated capsule feature matrix, we obtain the class-region-highlighted feature map output by the SFA module. In the output feature map, the features at each position have been globally recalibrated by taking into account the impacts of all the other positions. As shown in Figure 1, the improved feature map $\mathbf{P}$ output by the SFA module is finally leveraged to generate the road prediction map.

Each position on the prediction map encodes the certainty of the corresponding pixel in the input image belonging to the road region. To this end, a sigmoid function is performed on the last layer of the DA-CapsUNet to normalize the outputs to provide a probability-form feature encoding.

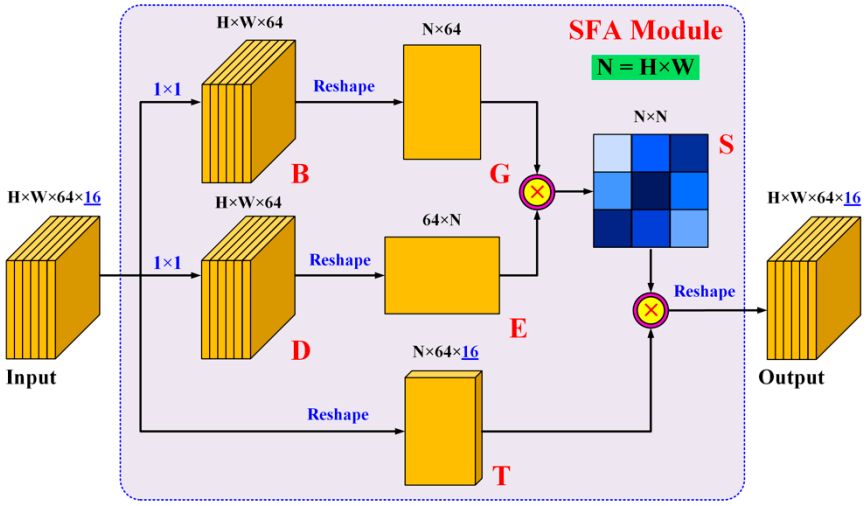


Figure 4. Architecture of the spatial feature attention (SFA) module.

To train the proposed DA-CapsUNet, an input image should be coupled with a binary label map as the ground truth for directing the backpropagation process and the minimization of the loss function. In the binary label map, the values of 1 indicate the road regions and the values of 0 denote the background. Figure 5 presents a sample of the binary label map. The loss function used to train the DA-CapsUNet is formulated as the focal loss [40] between the prediction map output by the network and the ground-truth binary label map as follows:

$$L = \sum_i \left[-p_i^*(1-p_i)^2 \log(p_i) - \left(1-p_i^*\right)p_i^2 \log(1-p_i)\right] \quad (7)$$

where p_i^* is the ground-truth label with regard to position i, whose value is 1 or 0 for, respectively, indicating the corresponding pixel at position i belonging to the road region or the background. p_i is the predicted value on the output prediction map, forecasting the certainty of the corresponding pixel at position i being within the road region.



Figure 5. Illustration of the binary label map used for training the proposed DA-CapsUNet. (**a**) Remote sensing image, and (**b**) binary label map.

3. Results

3.1. Dataset

In this paper, we built a large remote sensing image dataset, named GE-Road, for road extraction tasks and applied it to construct and evaluate the proposed DA-CapsUNet. The remote sensing images in this dataset were downloaded from the Google Earth service using the BIGEMAP software (http://www.bigemap.com). The GE-Road dataset contains 20,000 images covering roads of different materials and types, varying widths and shapes, and diverse environmental and surface conditions in

urban, rural, and mountainous areas all around the world. In the GE-Road dataset, each image has an image size of 800 × 800 pixels and a spatial resolution of about 0.3 to 0.6 m. This is a remarkably complex and challenging road extraction dataset. The road segments in the GE-Road dataset exhibit with different geometric topologies and distributions, and different levels of occlusions and shadow covers, as well as different image qualities. To facilitate network training and performance evaluation, each image in the GE-Road dataset has been annotated with a binary road region map as the ground truth (see Figure 5b). In our experiments, the GE-Road dataset was randomly divided into a training set, a validation set, and a test set. Concretely, 60% of the images were randomly selected from the GE-Road dataset to constitute the training set, 5% of the images were randomly selected as the validation set, and the remaining 35% of the images were used as the test set for performance assessment.

3.2. Network Training

The proposed DA-CapsUNet was trained in an end-to-end manner by backpropagation and stochastic gradient descent on a cloud computing platform with ten 16-GB GPU, one 16-core CPU, and a memory size of 64 GB. Before training, we randomly initialized all layers of the DA-CapsUNet by drawing parameters from a zero-mean Gaussian distribution with a standard deviation of 0.01. Each training batch contained two images per GPU and was trained for 1000 epochs. During training, we configured the initial learning rate as 0.001 for the first 800 epochs, and decreased it to 0.0001 for the other 200 epochs.

At the training stage, taking into consideration the orientation variation properties of the roads in the bird-view remote sensing images and the illumination condition variations of the remote sensing images, we also carried out data augmentation on the training images to enlarge the training set to cover roads with more conditions. Concretely, first, we flipped each training image in the horizontal direction to generate a horizontal mirror image. Then, we clockwise rotated the training image and its associated horizontal mirror image, respectively, in four directions with an angle interval of 90 degrees. As a result, a training image was transformed into eight images covering different orientations. The associated binary road region map of the training image was also transformed in the same way. Finally, we, respectively, increased and decreased the image brightness of each of the eight images to generate two other images aiming at simulating different illumination conditions. Therefore, after data augmentation, a training image was transformed into 24 images. The data-augmented training set was finally used to train the proposed DA-CapsUNet.

3.3. Road Extraction

To quantitatively evaluate the road extraction accuracy, we adopted the following three commonly used evaluation metrics: precision, recall, and F-score. The road extraction results obtained by the proposed DA-CapsUNet on the test set are quantitatively reported in detail in Table 1.

Table 1. Road extraction results and statistical tests obtained by different methods.

Method	Precision	Recall	F-Score	McNemar Test ($\alpha = 0.05$)
DA-CapsUNet	0.9523	0.9486	0.9504	-
MDRCNN [4]	0.8968	0.8872	0.8920	0.0015
HCN [6]	0.9373	0.9366	0.9369	0.0441
SII-Net [8]	0.9392	0.9381	0.9386	0.0474
MFPN [10]	0.9274	0.9255	0.9264	0.0227
SC-FCN [11]	0.9261	0.9214	0.9237	0.0159
SS-CNN [13]	0.9335	0.9286	0.9310	0.0351
DenseUNet [15]	0.9124	0.9012	0.9068	0.0032
GAN [18]	0.9251	0.9203	0.9227	0.0118
DA-CapsUNet-CTA	0.9346	0.9319	0.9332	0.0382
DA-CapsUNet-CFA	0.9223	0.9186	0.9204	0.0096
DA-CapsUNet-SFA	0.9388	0.9375	0.9381	0.0462
CapsUNet	0.9182	0.9079	0.9130	0.0041

As shown in Table 1, the proposed DA-CapsUNet achieved quite promising road extraction results on the test set. A high road extraction performance with a precision of 0.9523, a recall of 0.9486, and an F-score of 0.9504, respectively, were obtained. Specifically, the precision metric was slightly higher than the recall metric. It means that the proposed DA-CapsUNet performed effectively in differentiating the road regions from the background, thereby resulting in less false detections. The road extraction performance was quite competitive when processing the complex and challenging GE-Road dataset. The challenging scenarios of the GE-Road dataset require that the road extraction model should be robust enough to correctly identify any kinds of roads with low rates of false detections and misdetections, behave promisingly to determine the presence and the location of the roads and accurately differentiate them from the complicated environments, and be highly effective in maintaining the continuity and the topology completeness of the extracted roads. Fortunately, despite the remarkably challenging scenarios of the GE-Road dataset, the proposed DA-CapsUNet still performed effectively with a high road extraction accuracy. The advantageous performance obtained benefits from the following aspects. First, adopting a U-Net architecture with the fusion of multiscale and multilevel capsule features, the proposed DA-CapsUNet performed promisingly in enhancing both the localization accuracy and the feature representation capability. Second, integrated with the multiscale CTA modules, the proposed DA-CapsUNet was capable of taking advantage of different scales of contextual properties simultaneously at a high-resolution perspective without the loss of feature details. Third, integrated with the CFA modules, the proposed DA-CapsUNet contributed to emphasize the informative and salient features, which helped to further improve the feature representation robustness. Last but not least, designed with the SFA module, the proposed DA-CapsUNet was forced to concentrate on the spatial features tightly associated with the road regions and effectively suppress the influences of the background features, thereby providing a powerful class-specific feature encoding. On the whole, the proposed DA-CapsUNet achieved a competitive road extraction performance in extracting roads of different types and spectral appearances, varying topologies and distributions, and diverse surface and environmental conditions.

To visually inspect the road extraction performance, Figure 6 presents a subset of road extraction results obtained by the proposed DA-CapsUNet on the GE-Road dataset. As shown by the road extraction results in Figure 6, despite the considerable challenging scenarios, the road regions were well differentiated from the surrounding environments and the road boundaries were well fitted with a very small amount of false detections and misdetections. In addition, the extracted road regions were quite solid and continuous in spite of the existence of challenging scenarios. However, as shown in Figure 6a, some road segments were severely occluded by the roadside trees and buildings or covered with heavy dark shadows, especially for the cases of urban areas. Thus, the proposed DA-CapsUNet failed to maintain the completeness of these roads, thereby degrading the true positive

rate. In addition, some courtyards were directly connected with the roads and exhibited similar spectral and textural properties to the roads. As a result, such courtyards were falsely identified as the road regions, thereby increasing the false positive rate. Overall, designed with a novel and high-performance capsule U-Net architecture integrated with powerful context-augmentation and feature attention mechanisms, the proposed DA-CapsUNet provided a feasible and competitive model for road extraction applications.



Figure 6. Illustration of a subset of road extraction results. (**a**) Test images and (**b**) road extraction results.

3.4. Comparative Studies

To further demonstrate the applicability and the robustness of the proposed DA-CapsUNet in road extraction tasks, we conducted a set of comparative studies with some recently developed deep learning-based road extraction methods. The following eight methods were selected for performance comparisons: MDRCNN [4], HCN [6], SII-Net [8], MFPN [10], SC-FCN [11], scale-sensitive CNN (SS-CNN) [13], DenseUNet [15], and GAN [18]. Specifically, the MDRCNN leveraged multi-size kernels with different receptive fields to obtain hierarchical features and concatenated these features to predict the road region maps. The integrity of the extracted roads was further enhanced via a sector descriptor-based tracking process. The HCN consisted of three parallel backbones including a fully convolutional subnetwork, a modified U-Net subnetwork, and a VGG subnetwork to extract different-grained features and fused these features through a shallow convolutional subnetwork to conduct road extraction. Differently, the other models extracted and fused multilevel and multiscale features to enhance the feature representation capabilities. For the SII-Net, designed with a spatial information inference structure, both the local visual properties and the global spatial structure

features of the roads were abstracted and considered for road identification. For the MFPN, multilevel semantic features were exploited through a feature pyramid architecture, and a weighted balance loss function was designed to alleviate the class imbalance issue. For the SC-FCN, multiscale features extracted by the VGG subnetwork were integrated through deconvolution operations. Similarly, a spatial consistency-based ensemble method was proposed to optimize the loss function for solving the class imbalance problem. For the SS-CNN, an encoder-decoder architecture was constructed to extract and fuse multiscale features to improve the localization accuracy. The decoder involved a scale fusion module and a scale sensitive module for, respectively, conducting feature fusion and evaluating the feature importance. For the DenseUNet, a U-Net architecture with dense connection units was proposed to strengthen the quality of the fused features. For the GAN, deep convolutional GAN and conditional GAN were combined to improve the robustness of the network in the case of small-size samples.

For fair comparisons, at the training stage, the same training and validation sets and the same training data augmentation strategy were, respectively, applied to train these models. At the test stage, the same test set was used to evaluate the road extraction performances of these methods. The road extraction results obtained by these methods are quantitatively reported in detail in Table 1 by using the precision, recall, and F-score metrics.

As shown in Table 1, in terms of the road extraction accuracy, the HCN, SII-Net, and SS-CNN outperformed the other methods with an overall road extraction accuracy of over 0.93. The MFPN, SC-FCN, and GAN performed moderately and obtained similar road extraction performances. In contrast, the MDRCNN and DenseUNet behaved relatively less effectively than the other methods with an overall road extraction accuracy of around 0.90. In particular, designed with a spatial information inference structure, the SII-Net performed effectively in preserving the continuity of the extracted roads at the presence of less severe occlusions. The advantageous performance obtained by the HCN benefitted from the aggregation of the multi-grained features, which contributed positively to the enhancement of the feature representation capability. For the SS-CNN, the integration of the scale fusion and scale sensitive modules can direct the network to fuse multiscale features and evaluate the importance of the fused features, thereby upgrading the quality and robustness of the output features used for predicting road regions. In addition, by fusing multiscale and multilevel features and improving the loss functions for road extraction purposes, the MFPN, SC-FCN, and GAN obtained better performances than the MDRCNN and DenseUNet. However, these methods performed less effectively with a degradation of the road extraction accuracy when handling the roads covered with large areas of shadows or the roads with quite low contrasts with their surrounding scenes. For further comparisons, we also conducted a set of McNemar tests between the proposed DA-CapsUNet and each of the eight methods to assess their significant differences in extracting roads on the test set. The threshold in the McNemar tests was configured as $\alpha = 0.05$. That is, a concomitant probability with a value less than or equal to α means a significant difference; otherwise, a concomitant probability with a value greater than α means no significant difference. The computed concomitant probabilities are reported in Table 1. As shown in Table 1, the proposed DA-CapsUNet showed significant differences with the eight methods.

Comparatively, by designing the capsule U-Net architecture to extract and aggregate different-level and different-scale high-order capsule features and integrating the CTA, CFA, and SFA modules for, respectively, exploiting multiscale contextual information at a high-resolution perspective, emphasizing channel-wise informative and salient features, and focusing on the spatial features tightly associated with the road regions, our proposed DA-CapsUNet showed significantly competitive and superior or compatible performance over the eight compared methods. Therefore, through comparative analysis, we concluded that the proposed DA-CapsUNet provided a feasible, effective, and high-performance solution to the road extraction tasks based on remote sensing images.

3.5. Ablation Experiments

As ablation experiments, we further verified the effectiveness of the integrations of the CTA, CFA, and SFA modules on the upgrading of the road extraction accuracy. To this end, we modified the proposed DA-CapsUNet to construct four networks. First, we removed the CTA modules from the DA-CapsUNet and named the resultant network as DA-CapsUNet-CTA. Then, we removed the CFA modules from the DA-CapsUNet and named the resultant network as DA-CapsUNet-CFA. Next, we removed the SFA module from the DA-CapsUNet and named the resultant network as DA-CapsUNet-SFA. Finally, we removed all the CTA, CFA, and SFA modules from the DA-CapsUNet and named the resultant network as CapsUNet. For these four modified networks, the same training and validation sets, and the same training strategy were applied to construct these networks. The road extraction results obtained by these networks on the test set are quantitatively reported in Table 1. Obviously, in terms of the road extraction accuracy, without the integration of any of the three types of modules, an accuracy degradation was generated for all the four networks. In particular, a relatively large-margin performance degradation of about 3.74% was generated by the CapsUNet. The performance degradation was mainly caused by the existence of narrow roads and the road segments exhibiting low contrasts with their surrounding environments. Therefore, without the CTA modules for comprehensively exploiting multiscale contextual properties at a high-resolution perspective and the CFA and SFA modules for emphasizing the channel-wise informative features and highlighting the spatial features tightly associated with the road regions, the quality and the representativeness of the output features used for predicting the road regions were weakened, thereby leading to the performance degradation when dealing with these challenging road scenarios. In addition, McNemar tests were also conducted between the proposed DA-CapsUNet and each of the four modified networks to assess their significant differences in extracting roads on the test set. As shown in Table 1, the proposed DA-CapsUNet showed significant differences with the four modified networks. Through ablation experiments, we confirmed that the integrations of the CTA, CFA, and SFA modules performed positively and effectively in enhancing the feature representation capability of the proposed DA-CapsUNet and in improving the quality of the road extraction results.

4. Discussion

Compared to traditional scalar representation-based neural networks, capsule networks leverage vectorial representations to characterize more inherent and information-rich entity features. The capsule formulation can not only identify the existence of a specific feature as is done by the traditional neural network through the probability-based length but also depict how it actually looks using the instantiation parameters. Therefore, the feature abstraction level with respect to the feature distinctions and informativeness are dramatically enhanced to a large extent. In addition, contextual information provides important cues to recognize the presence and the appearance of an entity, especially for the objects with large topology variations and specific environments. Therefore, by introducing the multiscale CTA module into each block of the contracting pathway, multiscale contextual properties are intensively exploited and comprehensively aggregated to upgrade the quality of the extracted features at a high-resolution perspective without the loss of feature resolutions and details, which effectively alleviates the drawbacks caused by the gradual pooling operations. Moreover, the low-information feature channels usually contribute quite little to the prediction of distinctive and representative features, sometimes even degrade the feature abstraction capability. The influences of the background features also suppress the saliency of the features of the class of interest. Therefore, by introducing the CFA and SFA modules into the network architecture, the informative and salient features are effectively emphasized to cast more contributions and the class-specific features are positively highlighted to suppress the impacts of the background, which perform promisingly to enhance the feature representation robustness and directionality. As demonstrated in this paper, by taking advantage of these powerful techniques, the proposed DA-CapsUNet achieved a competitive performance and a superior accuracy over the compared deep learning methods in handling roads of varying challenging

scenarios, particularly in the conditions of low-contrast road boundaries, shadow covers, and large topology variations.

However, we also found that the proposed DA-CapsUNet still performed less effectively on the road segments with heavy dark shadows and severe occlusions. This is because the proposed DA-CapsUNet is actually a segmentation model. The prior knowledge regarding the continuity property of the roads and the characteristic of small width perturbations along a short distance of the roads was not reasonably considered and adopted to track the roads. As a result, there were still false detections and misidentifications in the road extraction results. In fact, this is an extremely challenging issue to maintain the completeness and accuracy of the extracted roads with the existence of varying challenging scenarios. It is still an open problem to be solved. In our future work, we will try to design a more powerful network architecture to further upgrade the robustness, representativeness, and distinctiveness of the output features, and exploit and formulate favorable prior knowledge of the roads to further improve the road extraction completeness and accuracy.

5. Conclusions

In this paper, to improve pixel-wise road extraction accuracy, we have proposed a novel dual-attention capsule U-Net, named DA-CapsUNet, by taking advantage of the superior properties of capsule networks and the U-Net architecture, and the powerful features of the feature attention mechanisms. The DA-CapsUNet employed a deep capsule U-Net architecture, which functioned promisingly to extract multiscale and multilevel capsule features and recover a high-quality, high-resolution, and semantically strong feature representation for predicting the road regions. By integrating the multiscale CTA module into each block of the contracting pathway, the proposed DA-CapsUNet can rapidly exploit and comprehensively consider multiscale contextual properties at a high-resolution perspective, thereby improving the feature representation capability and robustness. In addition, with the design of the CFA and SFA modules, the proposed DA-CapsUNet was sensitive to highlight informative and salient features and focus on the spatial features tightly associated with the road regions, which effectively suppressed the impacts of the background and provided a high-quality and class-specific feature representation. The proposed DA-CapsUNet has been evaluated on a large remote sensing image dataset and achieved a competitive performance in extracting roads of different types and spectral appearances, varying topologies and distributions, and diverse surface and environmental conditions. Quantitative assessments showed that an advantageous road extraction accuracy with a precision, a recall, and an F-score of 0.9523, 0.9486, and 0.9504, respectively, was obtained. Moreover, comparative experiments with a set of existing methods also confirmed the robust applicability and superior or compatible performance of the proposed DA-CapsUNet in road extraction tasks.

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

Funding: This research was funded by the Six Talent Peaks Project in Jiangsu Province, grant number XYDXX-098, and by the National Natural Science Foundation of China, grant numbers 41971414 and 41671454.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments in improving the paper, and the volunteers for their extensive efforts in building the GE-Road dataset.

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

References

1. Ma, L.; Liu, Y.; Zhang, X.; Ye, Y.; Yin, G.; Johnson, B.A. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS J. Photogramm. Remote Sens.* **2019**, *152*, 166–177. [CrossRef]
2. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
3. Abdollahi, A.; Pradhan, B.; Shukla, N.; Chakraborty, S.; Alamri, A. Deep learning approaches applied to remote sensing datasets for road extraction: A state-of-the-art review. *Remote Sens.* **2020**, *12*, 1444. [CrossRef]
4. Dai, J.; Du, Y.; Zhu, T.; Wang, Y.; Gao, L. Multiscale residual convolution neural network and sector descriptor-based road detection method. *IEEE Access* **2019**, *7*, 173377–173392. [CrossRef]
5. Eerapu, K.K.; Ashwath, B.; Lal, S.; Dell'Acqua, F.; Dhan, A.V.N. Dense refinement residual network for road extraction from aerial imagery data. *IEEE Access* **2019**, *7*, 151764–151782. [CrossRef]
6. Li, Y.; Guo, L.; Rao, J.; Xu, L.; Jin, S. Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 613–617. [CrossRef]
7. Liu, Y.; Yao, J.; Lu, X.; Xia, M.; Wang, X.; Liu, Y. RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2043–2056. [CrossRef]
8. Tao, C.; Qi, J.; Li, Y.; Wang, H.; Li, H. Spatial information inference net: Road extraction using road-specific contextual information. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 155–166. [CrossRef]
9. Senthilnath, J.; Varia, N.; Dokania, A.; Anand, G.; Benediktsson, J.A. Deep TEC: Deep transfer learning with ensemble classifier for road extraction from UAV imagery. *Remote Sens.* **2020**, *12*, 245. [CrossRef]
10. Gao, X.; Sun, X.; Zhang, Y.; Yan, M.; Xu, G.; Sun, H.; Jiao, J.; Fu, K. An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network. *IEEE Access* **2018**, *6*, 39401–39414. [CrossRef]
11. Zhang, X.; Ma, W.; Li, C.; Wu, J.; Tang, X.; Jiao, L. Fully convolutional network-based ensemble method for road extraction from aerial images. *IEEE Geosci. Remote Sens. Lett.* **2019**. [CrossRef]
12. He, H.; Yang, D.; Wang, S.; Wang, S.; Li, Y. Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss. *Remote Sens.* **2019**, *11*, 1015. [CrossRef]
13. Tan, X.; Xiao, Z.; Wan, Q.; Shao, W. Scale sensitive neural network for road segmentation in high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**. [CrossRef]
14. Zhang, Z.; Wang, Y. JointNet: A common neural network for road and building extraction. *Remote Sens.* **2019**, *11*, 696. [CrossRef]
15. Xin, J.; Zhang, X.; Zhang, Z.; Fang, W. Road extraction of high-resolution remote sensing images derived from DenseUNet. *Remote Sens.* **2019**, *11*, 2499. [CrossRef]
16. Yang, X.; Li, X.; Ye, Y.; Lau, R.Y.K.; Zhang, X.; Huang, X. Road detection and centerline extraction via deep recurrent convolutional neural network U-Net. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7209–7220. [CrossRef]
17. Zhang, Y.; Xiong, Z.; Zang, Y.; Wang, C.; Li, J.; Li, X. Topology-aware road network extraction via multi-supervised generative adversarial networks. *Remote Sens.* **2019**, *11*, 1017. [CrossRef]
18. Zhang, X.; Han, X.; Li, C.; Tang, X.; Zhou, H.; Jiao, L. Aerial image road extraction based on an improved generative adversarial network. *Remote Sens.* **2019**, *11*, 930. [CrossRef]
19. Lian, R.; Huang, L. DeepWindow: Sliding window based on deep learning for road extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 1905–1916. [CrossRef]
20. Wang, S.; Yang, H.; Wu, Q.; Zheng, Z.; Wu, Y.; Li, J. An improved method for road extraction from high-resolution remote-sensing images that enhances boundary information. *Sensors* **2020**, *20*, 2064. [CrossRef]
21. Hinton, G.E.; Krizhevsky, A.; Wang, S.D. Transforming auto-encoders. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Artificial Neural Networks – ICANN 2011, Espoo, Finland, 14–17 June 2011*; Springer: Berlin/Heidelberg, German, 2011; pp. 44–51.
22. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–10 December 2017; pp. 1–11.
23. Yu, Y.; Gu, T.; Guan, H.; Li, D.; Jin, S. Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1894–1898. [CrossRef]

24. Gao, Y.; Gao, F.; Dong, J.; Li, H.C. SAR image change detection based on multiscale capsule network. *IEEE Geosci. Remote Sens. Lett.* **2020**. [CrossRef]
25. Ma, L.; Li, Y.; Li, J.; Yu, Y.; Junior, J.M.; Gonçalves, W.N.; Chapman, M.A. Capsule-based networks for road marking extraction and classification from mobile LiDAR point clouds. *IEEE Trans. Intell. Transp. Syst.* **2020**. [CrossRef]
26. Guan, H.; Yu, Y.; Peng, D.; Zang, Y.; Lu, J.; Li, A.; Li, J. A convolutional capsule network for traffic-sign recognition using mobile LiDAR data with digital images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1067–1071. [CrossRef]
27. Paoletti, M.E.; Haut, J.M.; Fernandez-Beltran, R.; Plaza, J.; Plaza, A.; Li, J.; Pla, F. Capsule networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2145–2160. [CrossRef]
28. Xu, Q.; Wang, D.Y.; Luo, B. Faster multiscale capsule network with octave convolution for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* **2020**. [CrossRef]
29. Yu, Y.; Guan, H.; Li, D.; Gu, T.; Wang, L.; Ma, L.; Li, J. A hybrid capsule network for land cover classification using multispectral LiDAR data. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1263–1267. [CrossRef]
30. LaLonde, R.; Bagci, U. Capsules for object segmentation. In Proceedings of the 1st Conference on Medical Imaging with Deep Learning, Amsterdam, The Netherlands, 4–6 July 2018; pp. 1–9.
31. Yu, Y.; Ren, Y.; Guan, H.; Li, D.; Yu, C.; Jin, S.; Wang, L. Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery. *IEEE Geosci. Remote Sens. Lett.* **2020**. [CrossRef]
32. Li, S.; Li, M.; Xu, Y.; Bao, Z.; Fu, L.; Zhu, Y. Capsules based Chinese word segmentation for ancient Chinese medical books. *IEEE Access* **2018**, *6*, 70874–70883. [CrossRef]
33. Ma, X.; Zhong, H.; Li, Y.; Ma, J.; Cui, Z.; Wang, Y. Forecasting transportation network speed using deep capsule networks with nested LSTM models. *IEEE Trans. Intell. Transp. Syst.* **2020**. [CrossRef]
34. Liu, J.; Lin, H.; Yang, L.; Xu, B.; Wen, D. Multi-element hierarchical attention capsule network for stock prediction. *IEEE Access* **2020**, *8*, 143114–143123. [CrossRef]
35. Hsu, J.T.; Kuo, C.H.; Chen, D.W. Image super-resolution using capsule neural networks. *IEEE Access* **2020**, *8*, 9751–9759. [CrossRef]
36. Dong, Y.; Fu, Y.; Wang, L.; Chen, Y.; Dong, Y.; Li, J. A sentiment analysis method of capsule network based on BiLSTM. *IEEE Access* **2020**, *8*, 37014–37020. [CrossRef]
37. Huang, R.; Li, J.; Wang, S.; Li, G.; Li, W. A robust weight-shared capsule network for intelligent machinery fault diagnosis. *IEEE Trans. Indust. Inform.* **2020**, *16*, 6466–6475. [CrossRef]
38. Rajasegaran, J.; Jayasundara, V.; Jayasekara, S.; Jayasekara, H.; Seneviratne, S.; Rodrigo, R. DeepCaps: Going deeper with capsule networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10725–10733.
39. Chen, L.C.; Papandreou, G.; Kollinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
40. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).