

CS-CapsFPN: A Context-Augmentation and Self-Attention Capsule Feature Pyramid Network for Road Network Extraction from Remote Sensing Imagery

Yongtao Yu, Jun Wang, Haiyan Guan, Shenghua Jin, Yongjun Zhang, Changhui Yu, E. Tang, Shaozhang Xiao & Jonathan Li

To cite this article: Yongtao Yu, Jun Wang, Haiyan Guan, Shenghua Jin, Yongjun Zhang, Changhui Yu, E. Tang, Shaozhang Xiao & Jonathan Li (2021): CS-CapsFPN: A Context-Augmentation and Self-Attention Capsule Feature Pyramid Network for Road Network Extraction from Remote Sensing Imagery, Canadian Journal of Remote Sensing, DOI: [10.1080/07038992.2021.1929884](https://doi.org/10.1080/07038992.2021.1929884)

To link to this article: <https://doi.org/10.1080/07038992.2021.1929884>



Published online: 24 May 2021.



Submit your article to this journal [↗](#)



Article views: 2



View related articles [↗](#)






View Crossmark data [↗](#)



CS-CapsFPN: A Context-Augmentation and Self-Attention Capsule Feature Pyramid Network for Road Network Extraction from Remote Sensing Imagery

CS-CapsFPN: Un réseau pyramidal de fonctionnalités d'augmentation du contexte et de capsules d'auto-attention pour l'extraction du réseau routier à partir de l'imagerie de télédétection

Yongtao Yu^a , Jun Wang^a, Haiyan Guan^b , Shenghua Jin^a, Yongjun Zhang^a, Changhui Yu^a, E. Tang^a, Shaozhang Xiao^a, and Jonathan Li^c 

^aFaculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, China; ^bSchool of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China; ^cDepartment of Geography and Environmental Management, University of Waterloo, Waterloo, Canada

ABSTRACT

The information-accurate road network database is greatly significant and provides essential input to many transportation-related activities. Recently, remote sensing images have been an important data source for assisting rapid road network updating tasks. However, due to the diverse challenging scenarios of roads in remote sensing images, such as occlusions, shadows, material diversities, and topology variations, it is still difficult to realize highly accurate extraction of roads. This paper proposes a novel context-augmentation and self-attention capsule feature pyramid network (CS-CapsFPN) to extract roads from remote sensing images. By designing a capsule feature pyramid network architecture, the proposed CS-CapsFPN can extract and fuse different-level and different-scale high-order capsule features to provide a high-resolution and semantically strong feature representation for predicting the road region maps. By integrating the context-augmentation and self-attention modules, the proposed CS-CapsFPN can exploit multi-scale contextual properties at a high-resolution perspective and emphasize channel-wise informative features to further enhance the feature representation robustness. Quantitative evaluations on two test datasets show that the proposed CS-CapsFPN achieves a competitive performance with a precision, recall, intersection-over-union, and F_{score} of 0.9470, 0.9407, 0.8957, and 0.9438, respectively. Comparative studies also confirm the feasibility and superiority of the proposed CS-CapsFPN in road extraction tasks.

RÉSUMÉ

Une base de données précise du réseau routier est une information très importante qui fournit des données essentielles à de nombreuses activités liées au transport. Les images de télédétection sont maintenant une source de données essentielle pour faciliter la mise à jour rapide du réseau routier. Cependant, en raison de la complexité et de la diversité de l'apparence des routes dans les images de télédétection, causée par les occlusions, les ombres, la diversité des matériaux, et les variations de topologie, il est toujours difficile de réaliser une extraction très précise des routes. Cet article propose un nouveau réseau pyramidal de fonctionnalités d'augmentation du contexte et de capsules d'auto-attention (CS-CapsFPN) pour extraire les routes. En concevant une architecture de réseau pyramidal de type capsules, le CS-CapsFPN proposé peut extraire et fusionner des capsules de différents niveaux et à différentes échelles afin de fournir une représentation des entités en haute résolution et sémantiquement significative pour prédire les cartes routières de la région. En intégrant les modules d'augmentation du contexte et d'auto-attention, le CS-CapsFPN peut exploiter les propriétés contextuelles à plusieurs échelles dans une perspective haute résolution et mettre l'accent sur les caractéristiques informatives de chaque canal afin d'améliorer la robustesse de leur représentation. Les évaluations quantitatives sur deux ensembles de données d'essai montrent que le CS-CapsFPN obtient une performance

ARTICLE HISTORY

Received 25 October 2020
Accepted 23 April 2021

concurrentielle avec une précision, rappel, *intersection-over-union*, et F_{score} de 0.9470, 0.9407, 0.8957, et 0.9438, respectivement. Des études comparatives confirment également la faisabilité et la supériorité du CS-CapsFPN dans les tâches d'extraction routière.

Introduction

As a public infrastructure for vehicles and pedestrians, road plays an important and irreplaceable role in a variety of transportation-related activities. It provides a pathway for connecting different places, thereby effectively facilitating people's daily lives, and promoting the development of the economy and the progress of the society. The coverage rate, the network structure, the design pattern, and the condition of roads significantly affect the convenience, smoothness, and security of the transportation-related activities. Precise and complete road network information is favorable to conduct route planning and direct driving behaviors. However, due to road maintenance, road construction, and the influence of natural factors, the road network is not always changeless. The structure, the number, and the topology of roads are dynamically changing, resulting in the incompleteness and inaccuracy of the current road network database. Therefore, timely and accurately updating the road network information is of great significance and provides essential ingredients to a wide range of applications, including road planning, traffic capacity analysis, functional region partition, map navigation, etc.

Traditional means for road network updating is usually based on onsite manual investigations or accomplished using mobile mapping systems mounted with video cameras (or laser scanners) and global navigation satellite system (GNSS) antennas. However, such field surveys are time-consuming, labor-intensive, and inoperable in some extreme cases, even inaccurate caused by incomplete coverage of the roads. In recent decades, with the advances of optical remote sensing sensors in flexibility, quality, and resolution, it is quite efficient and cost-effective to collect high-quality remote sensing images covering large areas by using satellite or aerial sensors. Comparatively, satellite sensors have larger perspectives. They can periodically and easily survey an extensive area of interest, providing a time series of images with sub-meter spatial resolutions. In contrast, aerial sensors, such as airborne or unmanned aerial vehicle (UAV) systems, have the advantages of high portability and flying flexibility, and low-cost platform. They can rapidly capture the images of the surveyed areas with different levels of details. Thus,

due to the superior properties of the optical remote sensing systems and their output remote sensing images, they have been positively leveraged to assist road network updating tasks. To date, a collection of algorithms and techniques have been developed and applied to road network extraction by using remote sensing images. Consequently, a number of achievements and breakthroughs have been obtained with increasing enhanced accuracies. However, with the specific image capturing mode of bird views, roads in remote sensing images often suffer from different levels of incompleteness caused by the occlusions of roadside high-rise buildings, trees, and traffic facilities. In addition, the shadows cast on the roads, the on-road vehicles and pedestrians, and the painted road markings severely change the texture consistencies of the roads. Moreover, the material diversities, shape and width variations of the roads are also common issues. Therefore, it is still challengeable to realize accurate extraction of road networks from remote sensing images. Exploiting highly efficient and fully automated techniques to further improve road network extraction accuracy and efficiency is of great significance and also urgently demanded by a wide range of applications.

In this paper, we design a novel context-augmentation and self-attention capsule feature pyramid network architecture, named CS-CapsFPN, to extract roads from remote sensing images. The proposed CS-CapsFPN leverages vectorial capsules to encode high-order entity features. The construction of the CS-CapsFPN consists of a bottom-up pathway for extracting different-level and different-scale capsule features, and a top-down pathway, together with several lateral connections, for integrating different-level and different-scale capsule features to provide a high-resolution, semantically strong feature representation for accurate road extraction. Specifically, a context-augmentation and self-attention module is embedded into the network to recalibrate and enhance the features by taking into account both multi-scale context properties and channel-wise feature saliencies and informativeness. The proposed CS-CapsFPN performs effectively and efficiently in processing remote sensing images containing roads of varying types, diverse geometric topologies, and complicated environmental and surface conditions. The contributions of this paper

include the following: (1) a novel and deep capsule feature pyramid network architecture is designed to generate high-resolution, semantically strong capsule feature representations to predict the road region maps; (2) a context-augmentation and self-attention module is developed to exploit multi-scale context properties and emphasize channel-wise informative and salient features to enhance the feature representation capability.

Related works

Generally, the road extraction task can be accomplished through two different pipelines: road region extraction and road centerline extraction. Specifically, road region extraction methods focus on the accurate segmentation of the road regions and produce the pixel-wise labeling of the road regions in the images. In contrast, road centerline extraction methods mainly aim to delineate the architecture of the road network and generate only the skeletons of the roads in the images. The existing methods for road extraction from remote sensing images can be roughly categorized into traditional methods and deep learning based methods. In the following subsections, we will provide detailed reviews on these road extraction methods.

Traditional road extraction methods

Traditional methods for road extraction from remote sensing images usually leverage hand-crafted features and prior knowledge (e.g., geometric, spectral, topological, and contextual features) to formulate road recognition or segmentation models. Liu et al. (2020) proposed a semi-supervised high-level feature selection framework for road centerline extraction. This framework contained a processing pipeline of multiple feature selection based on adaptive sparse representation, semi-supervised road region extraction by combining feature learning with Markov random field (MRF), and road centerline extraction through Gabor filters and non-maximum suppression. To address the issues of occlusions and noise, Lv et al. (2017) integrated an adaptive multi-feature sparsity model into a semiautomatic road extraction approach. This approach leveraged the multi-feature sparse model to portray the road appearance. The extraction of roads was accomplished using a sparse constraint regularized mean-shift algorithm. Similarly, an adaptive multi-feature method combining entropy and spectral features with the digital surface model was designed

by Pan et al. (2019). To enhance both the smoothness and accuracy of the extracted road centerlines, Cheng et al. (2016) constructed a cascade framework consisting of semi-supervised segmentation, multi-scale filtering, and multi-direction non-maximum suppression. Specifically, the semi-supervised segmentation process effectively explored the intrinsic structures between the labeled and unlabeled samples, thereby improving the road extraction performance with limited labeled samples. To accurately differentiate the roads from the background, Alshehhi and Marpu (2017) presented a hierarchical graph-based road segmentation method. In this method, Gabor and morphological filtering operations were pre-applied to enhance the saliencies of the road pixels. Then, graph-based segmentation and post-processing were successively carried out to, respectively, extract road regions and remove irregularities. By integrating fuzzy logic system and ant colony optimization, Maboudi et al. (2018) proposed an object-based road extraction strategy. Through object-based image analysis, a set of spatial, spectral, and textural features were incorporated to model the roads. Similarly, an object-based method combining context-aware object feature integration and tensor voting was also suggested by Maboudi et al. (2016). A multi-stage approach integrating structural, spectral, textural, and contextual properties of objects was applied to extract the roads. To reduce computational complexity in road extraction tasks, Sghaier and Lepage (2016) proposed to use texture analysis and beamlet transform based multi-scale reasoning. Initially, mathematical morphology and Canny detector were, respectively, applied to distinguish rectilinear structures and identify road edge candidates. By combining local and global information via beamlet transform, multi-scale reasoning was performed to reconstruct the roads. In addition, normalized second derivative map (Bae et al. 2015), aperiodic directional structure measurement (Zang et al. 2016), joint enhancing filtering (Zang et al. 2017), information fusion (Miao et al. 2016), particle and extended Kalman filtering (Movaghati et al. 2010), and multi-source data integration (Li et al. 2019b; Zhang et al. 2018b), were also exploited for road extractions from remote sensing images.

Some researches convert the road extraction task into classification issues. Road recognition models are constructed using image features. Miao et al. (2015) designed a support vector machine (SVM) classifier for road extraction based on the enhanced features obtained using the object-based Frangi's filter and the object-based shape filter. The extracted road regions were further smoothed through an integration of

tensor voting, active contour, and geometrical information. Differently, Zhang et al. (2017) leveraged an SVM classifier to filter out the images containing no roads based on the histogram of oriented gradient (HOG) features. The extraction of roads was carried out by exploiting two saliency features of background differences and local linear edges. Zhou et al. (2017) learned a boosting classifier to identify road candidates from the clustered line segments. The line segments were clustered using K -means based on the stroke width transform feature map. Shi et al. (2014) extracted road centerlines by integrating spectral-spatial classification and shape features. First, an SVM classifier constructed based on morphological profiles was performed to segment road candidates. Then, road shape features were used to obtain the refined road map. In addition, Bakhtiari et al. (2017) proposed a semi-automatic road extraction approach based on an integration of edge detection, SVM classification, and mathematical morphology.

Although the traditional road extraction methods are easy to implement, the quality and effectiveness of the designed models depend greatly on the selected features and the strictness of the prior knowledge. They are usually sensitive to the type diversities, topology variations, and environmental changes of the roads.

Deep learning based road extraction methods

Recent development in deep learning techniques has burst a great number of breakthroughs on performance and accuracy in detection, classification, and segmentation tasks. Deep learning models have the superiorities of automatically abstracting high-level, representative features of entities in an end-to-end manner. Consequently, great efforts have been made to conduct road extraction from remote sensing images by using deep learning techniques. Wei et al. (2017) proposed a road structure refined convolutional neural network (CNN) for extracting roads. To obtain structured road extractions, deconvolutional and feature fusion layers were designed in the architecture of the CNN model. Dai et al. (2019) combined a multi-scale deep residual CNN (MDRCNN) with a sector descriptor-based post-processing. In this architecture, multi-scale convolution functioned to generate hierarchical features of different dimensions, whereas the residual connections and global average pooling were used to improve the efficiency of the network. Similarly, a refined deep residual CNN model was proposed by Gao et al. (2019) for road extraction. To

simultaneously deal with road detection and centerline extraction, Cheng et al. (2017) constructed a cascaded end-to-end CNN architecture (CasNet). The CasNet consisted of two subnetworks for road detection and centerline extraction tasks, respectively. Specifically, the centerline extraction subnetwork shared the features output by the road detection subnetwork. Differently, Liu et al. (2019) designed a multi-task CNN model, called RoadNet, to simultaneously segment road regions, and extract road centerlines and road boundaries. By automatically abstracting multi-scale and multi-level features, the RoadNet performed promisingly in handling the roads of diverse scales and in various scenarios. To alleviate occlusions and preserve road continuity, Tao et al. (2019) proposed a spatial information inference net (SII-Net) for road extraction. This network could learn both the local visual properties of the roads and the global spatial structure information of the roads, such as continuity and trend. Zhang et al. (2019b) leveraged a multiple feature fully convolutional network (FCN) to extract roads in mountainous areas. This network was divided into three parts: encoding, bridge, and decoding. Concretely, the encoding part functioned to extract depth feature maps, and the decoding part recovered the feature maps for road segmentation. These two parts were connected by the bridge part. To solve the imbalance between the roads and the background areas in remote sensing images, Zhang et al. (2020) presented an FCN-based ensemble approach for road extraction. In this approach, a spatial consistency based ensemble method was applied to determine the weight of the loss function to handle the imbalance. Gao et al. (2018) suggested a multiple feature pyramid network (MFPN) to extract roads. In the MFPN, an effective feature pyramid architecture and a tailored pyramid pooling module were designed to take advantage of the multi-level semantic features. Likewise, a weighted balance loss function was also used to solve the imbalance issue caused by the sparseness of roads. In addition, an atrous spatial pyramid pooling integrated encoder-decoder network was also leveraged by He et al. (2019) for road extraction.

To improve the smoothness and boundary adherence of the extracted roads, Shi et al. (2018) proposed an end-to-end generative adversarial network (GAN) for road extraction. Based on adversarial training, the GAN could discriminate between the segmentation maps from either the ground truth or the segmentation output and enforce long-range spatial label contiguity to provide consistent road extractions. Zhang et al. (2019a) designed an improved GAN

architecture, which only required a few samples for training. In this model, a content-based loss term was integrated into the original GAN's loss function to serve the road extraction task. To handle occlusions and shades, Zhang et al. (2019c) constructed a multi-supervised GAN (MsGAN), which was jointly trained using the spectral and topology features of the roads. The MsGAN was capable of identifying aberrant road cases based on the relationship between the road region and the centerline. Zhang et al. (2018a) combined the strengths of residual learning and U-Net architecture and developed a deep residual U-Net (ResUnet) for road extraction. Specifically, residual units were adopted as basic blocks to build the encoding, bridge, and decoding parts of the ResUnet. Yang et al. (2019) proposed a deep recurrent CNN U-Net (RCNN-UNet) to simultaneously perform road detection and road centerline extraction. The RCNN-UNet leveraged a U-Net architecture built based on RCNN units and contained two predictors sharing the same backbone for, respectively, segmenting road regions and generating road skeletons. Differently, inspired by the densely connected CNN and U-Net, Xin et al. (2019) designed a DenseUNet architecture with few parameters and robust characteristics. With dense connection units and skip connections, the DenseUNet effectively strengthened the feature representation capability by fusing different-scale features at various network layers. To cope with complex backgrounds and scale variations of the roads, Li et al. (2019a) developed a hybrid convolutional network (HCN) to improve road extraction accuracy. The HCN consisted of three parallel branches, including an FCN, a modified U-Net, and a VGG, to, respectively, generate a coarse-grained, a medium-grained, and a fine-grained road segmentation map. The multi-grained segmentation maps were further fused by a shallow convolutional subnetwork for final road extraction. In addition, dense refinement residual network (Eerapu et al. 2019), richer convolutional features network (Hong et al. 2018), and deep transfer learning (Senthilnath et al. 2020) were also explored for extracting roads from remote sensing images.

Despite the progress and achievements of the deep learning models made so far due to the automated end-to-end high-level feature abstraction mechanism, they still face the challenges of the requirement of large-volume annotated data and plenty of time cost for model training. The quality and the amount of the annotated data also affect significantly on the robustness and accuracy of the designed deep learning models.

Methodology

Capsule network

Traditional deep learning models are usually built based on scalar neurons, which characterize the probabilities of the existence of specific features. Thus, in order to capture the variances of entities, more extra neurons are often required to be embedded to encode the different variants of the same entity, resulting in the size expansion of the entire network. Recently, capsule networks have exhibited advantageous properties in feature extraction and representation capabilities. Particularly, capsule networks leverage vectorial capsule encodings as basic units to characterize entity features. A capsule can be viewed as a vector combination of scalar neurons, whose length encodes the certainty of the presence of an entity, and whose instantiation parameters reflect the inherent properties of the entity (Sabour et al. 2017). A ground-breaking property of such capsule formulation is that the vectorial representation allows a capsule not only to detect a feature but also to learn and identify its variants without including more capsules, thereby providing a powerful, but lightweight, feature representation model. Capsule networks have shown promising and competitive performances in many detection (Yu et al. 2019, Yu et al. 2020a), classification (Paoletti et al. 2019), and segmentation tasks (Yu et al. 2020b). Thus, in this paper, by taking advantage of the superior properties of capsule networks, we construct a CS-CapsFPN architecture aiming to exploit multi-scale context properties and channel attention mechanisms to provide high-resolution and semantically strong feature representations to improve the road extraction accuracy.

Capsule convolutions are quite different from traditional convolution operations. Concretely, for a capsule j in a capsule convolutional layer, the total input to the capsule is a dynamically determined weighted sum over all the predictions from the capsules within the convolution kernel in the previous layer as follows:

$$C_j = \sum_i a_{i,j} \times U_{i,j} \quad (1)$$

where C_j is the total input to capsule j ; $a_{i,j}$ is a dynamically determined coupling coefficient indicating the degree of contribution of the prediction from capsule i in the previous layer; $U_{i,j}$ is the prediction from capsule i to capsule j , and it has the following form:

$$U_{i,j} = \mathbf{W}_{i,j} U_i \quad (2)$$

where U_i is the output of capsule i and $\mathbf{W}_{i,j}$ is a

transformation matrix acting as a feature mapping function. Specifically, the coupling coefficients between capsule i and all its connected capsules in the layer above sum to 1 and are determined by the improved dynamic routing process (Rajasegaran et al. 2019).

Note that, we use the length of a capsule to encode the saliency of a feature. That is, short capsules should cast low probability estimations; whereas long capsules should result in high probability estimations. Thus, we adopt the nonlinear “squashing” function (Sabour et al. 2017) as the activation function to normalize the input of a capsule. The squashing function is formulated as follows:

$$U_j = \frac{\|C_j\|^2}{1 + \|C_j\|^2} \times \frac{C_j}{\|C_j\|} \quad (3)$$

where C_j and U_j are, respectively, the input and the output of capsule j . The modulus of a vector is calculated by the operator $\|\cdot\|$. 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 few contributions.

Context-augmentation and self-attention capsule feature pyramid network

As shown in Figure 1, the proposed CS-CapsFPN, which is designed as a fully convolutional feature

pyramid network architecture aiming at providing a high-resolution and semantically strong feature map by considering multi-level and multi-scale feature semantics, takes a remote sensing image as the input and outputs an equal-size road region map in an end-to-end manner. The architecture of the CS-CapsFPN involves a bottom-up pathway, a top-down pathway, and several lateral connections. The bottom-up pathway functions to extract different levels and different scales of capsule features. The top-down pathway, assisted by the lateral connections, serve to integrate the different-level and different-scale capsule features to recover high-resolution and semantically strong feature representations for generating a high-quality road map. However, different from the network architecture proposed by Yu et al. (2020a), an improved multi-scale context-augmentation module and a novel channel-wise self-attention module are integrated in the feature pyramid network architecture to effectively boost the quality of the output feature representations.

The bottom-up pathway is a feature extraction network, which is composed of two traditional convolutional layers for low-level feature extraction, and a group of capsule convolutional layers and capsule pooling layers for different scales of high-level capsule feature abstraction. The scalar features output by the second convolutional layer are further encoded into vectorial capsule representations to characterize entity features. This primary capsule layer can be constructed through traditional convolution operations.

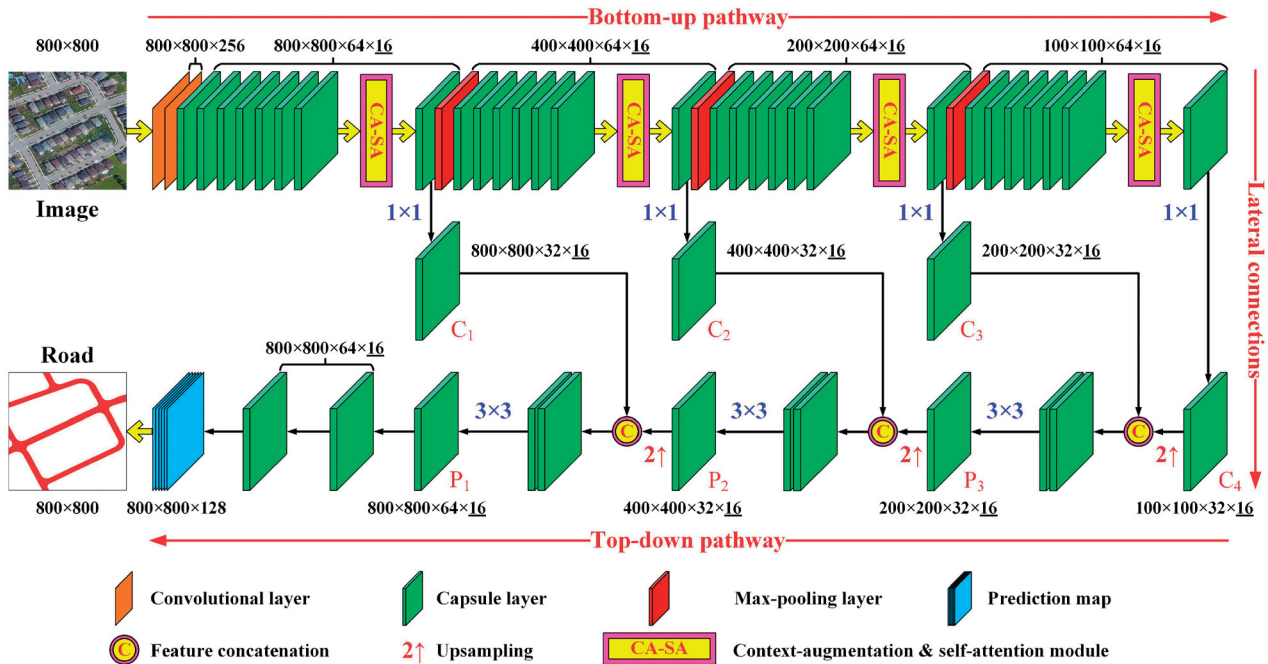


Figure 1. Architecture of the proposed context-augmentation and self-attention capsule feature pyramid network (CS-CapsFPN). The dimension of a capsule is configured as 16.

For example, denote D_p as the number of feature channels in the primary capsule layer and S_p as the dimension of a capsule. Then, a total of $D_p \times S_p$ different convolution kernels are performed on the second convolutional layer, resulting in $D_p \times S_p$ feature channels (see Figure 2). After convolutions, the generated feature channels are equally partitioned into D_p groups, each of which contains S_p feature channels. As shown in Figure 2, for each group, the S_p components at the same position are concatenated to constitute an S_p -dimensional capsule. In such a way, the low-level scalar feature representations are converted into high-level vectorial entity representations. Specifically, for these two traditional convolutional layers, the widely used rectified linear unit (ReLU) is adopted as the activation function.

As shown in Figure 1, the capsule convolutional and pooling layers in the bottom-up pathway are split into four network stages for computing a pyramid of capsule features at different scales with a scaling step of two. Within each stage, the feature maps in each capsule layer maintain the same resolution and spatial size. The spatial size of feature maps is gradually scaled down stage by stage to generate lower-resolution feature maps with a scaling step of two. Specifically, the feature maps in the first stage have the same spatial size as the input image. At the end of each stage (except the last stage), a capsule pooling layer is appended to conduct feature downsampling and salient feature selection, and to enlarge the receptive field. Here, we adopt max-pooling operations to select the most representative capsule features. Concretely, for the capsules within a max-pooling kernel in each feature channel, only the capsule having the longest length is retained and the others are discarded. After max-pooling, a feature map is scaled down to the half size. From bottom layer to top layer, the spatial resolution of the feature maps in each stage is decreased gradually, whereas higher-level features with larger receptive fields are generated.

Theoretically, the deepest layer in each stage has the strongest and most representative feature encodings. Therefore, for the bottom-up pathway, we pick the feature map of the last capsule layer in each stage as the reference set of feature maps for the subsequent feature fusion, augmentation, and refinement. To facilitate feature fusion, we apply a 1×1 capsule convolution operation (kernel size is 1×1) to each of the reference feature maps to modulate their channel numbers to the same configuration of $d = 32$ while keeping their original spatial resolutions. As shown in Figure 1, after feature map selection and modulation, we obtain the final reference set of feature maps $\{C_1, C_2, C_3, C_4\}$, which have scales of $\{1, 1/2, 1/4, 1/8\}$, respectively, with regard to the input image.

The receptive field of a capsule in each stage is enlarged slowly layer by layer through capsule convolutions. Thus, to rapidly expand the receptive field of a capsule to include more context properties, we append a max-pooling layer at the end of each stage to scale down the feature maps to a small size. This is a common design pattern in CNN models. However, after max-pooling, the feature details are partially damaged in the resultant lower-resolution feature maps, which is adverse for the feature encoding and the extraction of small-size roads. In addition, capsule convolution operations behave equally on all the channels of a feature map within the local receptive fields at each layer. The interdependencies and salencies among the channels are weakly exploited, which is not helpful to obtain informative features. Thus, in order to rapidly enlarge the receptive field of a capsule to consider more context information in each stage without the loss of feature map resolution and explicitly model the interdependencies among the feature channels to emphasize informative features and suppress the less useful ones, we embed a cascaded multi-scale context-augmentation (CA) and channel-wise self-attention (SA) module over the deepest layer in each stage. Through feature recalibration by the

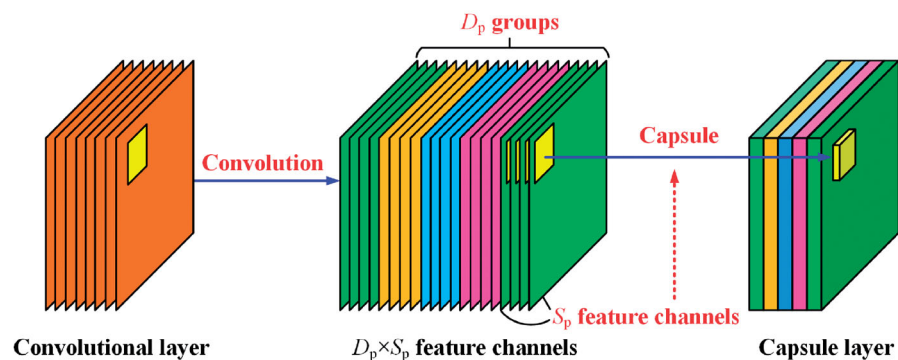


Figure 2. Illustration of the construction of the primary capsule layer.

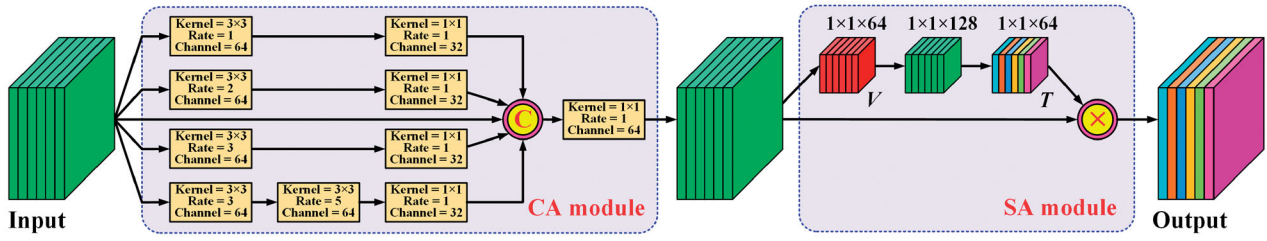


Figure 3. Architecture of the multi-scale context-augmentation (CA) and channel-wise self-attention (SA) module.

CA and SA module, the output features are more robust and more informative.

As shown in Figure 3, the CA module takes a feature map as the input and outputs an augmented feature map having the same number of channels and the same size as the input feature map. In this paper, we leverage the atrous convolutions (Chen et al. 2018) to design the CA module. Different from standard convolutions, atrous convolutions involve an extra hyperparameter, i.e., atrous rate, to indicate the stride of the kernel parameters. Specifically, a standard convolution is actually a special case of an atrous convolution with the atrous rate of 1. An important property of atrous convolutions is that, by adjusting the atrous rate, different receptive fields can be accessed without increasing the number of kernel parameters. For example, a 3×3 atrous convolution kernel with an atrous rate of 3 has the same size of receptive field as a standard 7×7 convolution kernel, but fewer parameters (9 versus 49). Thus, taking into account the advantageous properties, we adopt atrous convolutions to design the multi-scale CA module. As shown in Figure 3, four parallel branches are performed on the input feature map to exploit different scales of context features with the gradual increment of the atrous rates. 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 branch perform a 3×3 atrous convolution with atrous rates of 2 and 3, respectively, to consider 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 appended to smooth the extracted features and modulate the feature channels. Afterwards, the features encoding different scales of context information from these four branches, along with the original input feature, are concatenated and further fused through a 1×1 atrous convolution to generate the output feature map. By introducing the CA module, the output feature map can rapidly include different scales of context information without the loss of feature details and

resolutions, which is quite powerful to characterize entities of varying sizes, especially those of small sizes.

As shown in Figure 3, the output of the CA module is fed into the SA module to conduct feature recalibration to emphasize informative features. The output of the SA module involves a recalibrated feature map having the same number of channels and the same size as the input feature map. We expect the input features to be enhanced by explicitly modeling the channel interdependencies in order to increase the sensitivity of the network to informative features with a global perspective. To this end, first, by collecting channel-wise statistics, we apply a global average pooling operation to the input feature map to transform it into a channel descriptor V , whose length equals to the number of channels of the input feature map. Formally, for each channel of the input feature map, a scalar value is generated by spatially squeezing the lengths of the capsules in this channel through a global average pooling operation as follows:

$$v_i = \frac{1}{H \times W} \sum_j \|U_j^i\|, \quad i = 1, 2, \dots, 64 \quad (4)$$

where H and W denote the height and width of the input feature map; U_j^i is the output of a capsule in the i -th channel; v_i denotes the squeezed value corresponding to the i -th channel. As shown in Figure 3, by concatenating the squeezed values from all the channels, we constitute the channel descriptor V . Each element of V encodes a global perspective of the feature statistics in the corresponding channel. Then, to take advantage of the globally aggregated information in the channel descriptor, we further append two fully-connected layers to exploit channel-wise interdependencies. By learning nonlinear interactions among the channels in a non-mutually-exclusive manner, these two fully-connected layers allow multiple channels to be emphasized, rather than a one-hot activation. Specifically, we adopt the ReLU and the sigmoid functions to, respectively, activate the outputs of these two layers. Finally, the probability-encoded output of the second fully-connected layer forms a channel-wise attention descriptor T , each of whose elements reflects the saliency and the informativeness of the

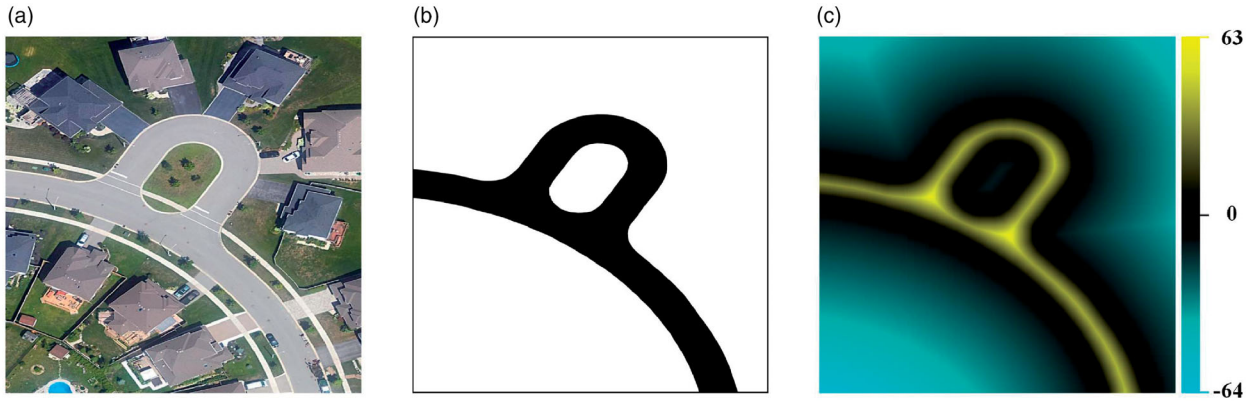


Figure 4. Illustration of the label map used for training the CS-CapsFPN. (a) Remote sensing image, (b) road region map, and (c) signed distance map.

corresponding channel of the input feature map. The attention descriptor T functions as a weight descriptor to recalibrate the input feature map. As shown in Figure 3, this is achieved by multiplying the attention descriptor with the input feature map in a channel-wise way as follows:

$$U_j^{i*} = t_i \times U_j^i, \quad i = 1, 2, \dots, 64 \quad (5)$$

where t_i denotes the i -th element of the attention descriptor T ; U_j^i and U_j^{i*} are the outputs of the original capsule and the recalibrated capsule in the i -th channel, respectively. After channel-wise recalibration, the informative features are effectively highlighted, whereas the less salient features are rationally suppressed, thereby further enhancing the feature representation robustness.

Generally, with the layers going deeper, the extracted features become higher-level and semantically stronger. That is, feature maps at high stages have semantically strong features with large receptive fields, but low resolutions. In contrast, feature maps at low stages have high resolutions, but semantically weak features with small receptive fields. Thus, through the top-down pathway and lateral connections, we aim to fuse the different-level and different-scale capsule features in $\{C_1, C_2, C_3, C_4\}$ selected from the bottom-up pathway to provide a high-resolution and semantically strong feature representation to improve the road extraction accuracy. To this end, as shown in Figure 1, first, the top-down pathway recovers higher-resolution features by upsampling spatially a feature map to its twice size to facilitate feature concatenation and fusion. In this paper, we adopt capsule deconvolution operations to perform feature upsampling. Then, the upsampled features are concatenated with the modulated features from the previous stage through the lateral connection. Finally, we apply a 3×3 capsule convolutional layer to the concatenated

features to conduct feature fusion and modulate the channel number to the configuration of d . In this way, high-resolution features (with accurate localization details) are fused with and augmented by semantically strong features (with large scopes of context information) to provide a high-resolution, semantically strong feature representation. As shown in Figure 1, the above feature fusion process starts from feature map C_4 , and repeats downward till all the features in $\{C_1, C_2, C_3, C_4\}$ are fused, resulting in a set of fused feature maps $\{P_1, P_2, P_3\}$. The feature map P_1 combining all the scales of features and with the highest resolution is used to generate the final road prediction map.

To train the CS-CapsFPN, an input image should be coupled with a road label map as the ground truth. A common way is to assign a binary road region map with the input image (see Figure 4(a and b)). However, a deficiency of using binary road region map is that the extracted roads might not be perfectly solid probably with the presence of tiny holes. In this paper, we leverage the signed distance map (Yuan 2018) as the road label map, which is capable of encoding both the interior and exterior information of roads simultaneously. As shown in Figure 4(b), based on the binary road region map, we transform it into the signed distance map shown in Figure 4(c). The value at a position in the signed distance map is calculated as the distance from the position to its nearest point on the road boundary with a positive value indicating the road interior and a negative value indicating the background. Specifically, rather than directly using the signed values as the ground truths, we bin and quantify the signed values into 128 integers ranging from -64 to 63 as the ground truths to direct the training process.

As shown in Figure 1, based on the quantified signed distance map, the road prediction layer

contains a stack of 128 prediction maps, each of which corresponds to one of the 128 binned distance ranges (i.e., -64 to 63), to convert the road extraction task into a multi-class classification issue. At each position of the prediction maps, the outputs indicate the probability of the pixel in the input image belonging to one of the 128 classes (i.e., a certain distance range from the interior or the exterior of a road). To this end, the outputs at each position of the prediction maps are normalized by the softmax function to provide a one-hot activation. At the training stage, the loss function used for training the CS-CapsFPN is formulated as the Focal Loss (Lin et al. 2020) between the prediction results and the quantified signed distance map as follows:

$$L = \sum_i - (1 - p_i)^2 \log(p_i) \quad (6)$$

where p_i denotes the softmax output at a position on the prediction map corresponding to the ground truth.

Road extraction

Once the CS-CapsFPN is constructed, we apply it to the remote sensing images to conduct road extraction. For a test image fed into the CS-CapsFPN, the output of the road prediction layer contains a stack of 128 prediction maps. Then, we convert these prediction maps into a signed distance map having the same spatial size as the prediction maps for road extraction. To this end, for each position of the prediction maps, the resultant value of the corresponding position in the signed distance map is assigned as the class label (i.e., -64 to 63) of the prediction maps with the maximum softmax output. Finally, based on the generated signed distance map, the positions with values greater than or equal to 0 are marked as the road regions (see Figure 1).

Results and discussion

Datasets

In this paper, we built two big remote sensing image datasets to evaluate the performance of the proposed CS-CapsFPN on road extraction. The first dataset was collected from the Google Earth service using the BIGEMAP software (<http://www.bigemap.com>). We named it as the GE-Road dataset. 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 (e.g., occlusions and shadows) in urban, rural, and mountainous areas. All the images in the GE-Road dataset

have the same image size of 800×800 pixels. The second dataset was collected using a DJI Phantom 4 Pro UAV system flying in the urban and suburban areas in China. We named it as the UAV-Road dataset. The UAV-Road dataset consists of 15,000 images, each of which has an image size of 800×800 pixels. The roads in the images exhibit with different widths, types, and shapes, as well as diverse scenarios and surface conditions. To facilitate network training and performance assessment, each of the images in these two datasets has been annotated with a binary road region map as the ground truth (see Figure 4(b)). In our experiments, we randomly divided the two datasets into a training set, a validation set, and a test set. Concretely, 60% of the images in each dataset were randomly selected to construct the training set, 5% of the images in each dataset were randomly selected as the validation set, and the rest 35% of the images in each dataset were used as the test set for performance evaluation. For the training set, a signed distance map (see Figure 4(c)) was generated for each training image based on the associated binary road region map as the ground truth for constructing the proposed CS-CapsFPN.

Network training

The proposed CS-CapsFPN 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 CS-CapsFPN 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 rest 200 epochs. To trade off the computational efficiency and the feature representation capability, as well as the road extraction accuracy, we configured the value of d as 32 and the dimension of a capsule as 16 for all capsule layers.

At the training stage, to effectively train the CS-CapsFPN toward high-performance road extraction, data augmentation was also conducted on the training images to enlarge the training set to cover roads of different orientations and illumination conditions. Concretely, first, each training image was flipped in the horizontal direction to generate a horizontal mirror image. Then, these two images were rotated clockwise in four directions with an angle interval of 90

degrees. Finally, we increased and decreased the image brightness of each of the eight images to generate two other images. The ground-truth binary road region map of a training image was also transformed in the same way. As a result, after data augmentation, a training image was transformed into 24 images covering roads of different orientations and illumination conditions. The data-augmented training set, which was 24 times in size of the original training set, was finally leveraged to train the CS-CapsFPN.

Road extraction

At the test stage, we applied the constructed CS-CapsFPN to the test set containing 12,250 images to examine the road extraction performance. For a test image, first, it was passed to the bottom-up pathway to extract different-level and different-scale capsule features, which were contextually augmented and channel-wisely recalibrated by the CA and SA module in each stage. Then, the top-down pathway and lateral connections were executed to integrate these multi-level and multi-scale capsule features to provide a high-resolution, semantically strong feature representation for generating the road prediction maps. Finally, the road prediction maps were converted into a signed distance map to generate the road extraction result by labeling the positions with values greater than or equal to zero on the signed distance map.

To quantitatively evaluate the road extraction accuracy, we adopted the following four commonly used evaluation metrics: precision (P), recall (R), intersection-over-union (IoU), and F_{score} . Precision is defined as the proportion of the correctly identified road pixels with regard to the road extraction result. Recall is defined as the proportion of the correctly extracted road pixels with regard to the annotated ground truth. IoU is defined as the proportion of the correctly recognized road pixels with regard to the union of the annotated ground truth and the extraction result. F_{score} evaluates the overall road extraction performance by taking into account both the precision and the recall measures. Let denote TP, FP, and FN as the numbers of true positives, false positives, and false negatives, respectively. Then, these four metrics can be defined as follows:

$$P = \frac{(TP)}{(TP) + (FP)} \quad (7)$$

$$R = \frac{(TP)}{(TP) + (FN)} \quad (8)$$

$$IoU = \frac{(TP)}{(TP) + (FN) + (FP)} \quad (9)$$

$$F_{score} = 2 \times \frac{P \times R}{P + R} \quad (10)$$

The road extraction results on the two test datasets are reported in Table 1 by using the above four evaluation metrics. In Table 1, “Overall” denotes the overall performance obtained on the two test datasets. It includes the average precision, average recall, average IoU , and average F_{score} . In addition, a precision-recall curve is also provided in Figure 5.

As reflected in Table 1, the proposed CS-CapsFPN achieved quite promising road extraction performances on the two test datasets. Specifically, a road extraction performance with a precision of 0.9562, a recall of 0.9493, an IoU of 0.9056, and an F_{score} of 0.9527 was obtained on the GE-Road dataset. For the UAV-Road dataset, an extraction performance with a precision, a recall, an IoU , and an F_{score} of 0.9378, 0.9321, 0.8857, and 0.9349, respectively, was achieved on road extraction. The GE-Road dataset is a very challenging dataset due to the following aspects: (1) Surface material differences of the roads, such as asphalt-paved, concrete-paved, and unsurfaced roads. The surface material differences of the roads result in different spectral and textural properties of the roads in the remote sensing images. Thus, it requires that the road extraction model should be accurately enough to identify all kinds of roads with low false detection and misdetection rates. (2) Width and shape diversities of the roads. The roads in the images exhibit with diverse widths, shapes, and distributions. Such diversities require that the road extraction model should have high adaptabilities to correctly locate small-size roads and adhere tightly to the edges of large-size and varying-shape roads. (3) Complicated environmental conditions of the roads. Due to the existence of spectral and textural similarities between the roads and the scene objects, the road extraction model should perform efficiently to verify the presence of the roads and correctly differentiate the roads from the surrounding environments. (4) Occlusions and shadow covers of the roads. Due to the bird-view image capture mode of remote sensing images, the roadside high-rise buildings and trees often cause different levels of occlusions to the roads, thereby resulting in the incompleteness and topology changes of the roads in the images. In addition, the shadows cast on the roads also change the spectral and texture consistency of the roads. Thus, it requires that the road extraction model should be highly capable of guaranteeing the continuities and the topology completeness of the extracted

Table 1. Road extraction results obtained by different methods.

Method	Dataset	Quantitative evaluation				Speed (images s ⁻¹)
		Precision	Recall	<i>IoU</i>	F_{score}	
CS-CapsFPN	GE-Road	0.9562	0.9493	0.9056	0.9527	7
	UAV-Road	0.9378	0.9321	0.8857	0.9349	
	Overall	0.9470	0.9407	0.8957	0.9438	
C-CapsFPN	GE-Road	0.9407	0.9356	0.8924	0.9381	7.3
	UAV-Road	0.9256	0.9231	0.8719	0.9243	
	Overall	0.9332	0.9294	0.8822	0.9312	
S-CapsFPN	GE-Road	0.9382	0.9335	0.8879	0.9358	7.7
	UAV-Road	0.9234	0.9217	0.8613	0.9225	
	Overall	0.9308	0.9276	0.8746	0.9292	
CapsFPN	GE-Road	0.9289	0.9235	0.8765	0.9262	8
	UAV-Road	0.9168	0.9146	0.8469	0.9157	
	Overall	0.9229	0.9191	0.8617	0.9210	
MDRCNN	GE-Road	0.8968	0.8872	0.8053	0.8920	11
	UAV-Road	0.8734	0.8711	0.7853	0.8722	
	Overall	0.8851	0.8792	0.7953	0.8821	
SII-Net	GE-Road	0.9392	0.9381	0.8943	0.9386	6
	UAV-Road	0.9209	0.9237	0.8598	0.9223	
	Overall	0.9301	0.9309	0.8771	0.9305	
FCN	GE-Road	0.9261	0.9214	0.8671	0.9237	9
	UAV-Road	0.9089	0.9076	0.8315	0.9082	
	Overall	0.9175	0.9145	0.8493	0.9160	
MFPN	GE-Road	0.9274	0.9255	0.8778	0.9264	8
	UAV-Road	0.9115	0.9103	0.8374	0.9109	
	Overall	0.9195	0.9179	0.8576	0.9187	
GAN	GE-Road	0.9251	0.9203	0.8632	0.9227	7
	UAV-Road	0.9068	0.9057	0.8287	0.9062	
	Overall	0.9160	0.9130	0.8460	0.9145	
ResUnet	GE-Road	0.9049	0.8977	0.8212	0.9013	7
	UAV-Road	0.8864	0.8767	0.7945	0.8815	
	Overall	0.8957	0.8872	0.8079	0.8914	
HCN	GE-Road	0.9373	0.9366	0.8892	0.9369	2
	UAV-Road	0.9194	0.9222	0.8537	0.9208	
	Overall	0.9284	0.9294	0.8715	0.9289	

roads. The UAV-Road dataset mainly covers asphalt-paved roads. Except for the challenges caused by width and shape diversities, complicated scenarios, occlusions, and shadow covers, the roads also suffer from the contaminations of on-road objects (e.g., vehicles and pedestrians) due to the high-resolution properties of the UAV images. The influence of on-road objects alters the solid and continuity properties of the roads. Thus, it requires that the road extraction model should be robust enough to suppress the influences of on-road objects and extract solid and continuous roads. Despite the challenging scenarios of the GE-Road and UAV-Road datasets, our proposed CS-CapsFPN still performed effectively with high road extraction accuracies. On the whole, the CS-CapsFPN showed superior performance in processing rural roads and mountainous roads since the road conditions were less complicated and the roads usually exhibited with stronger contrasts with their surroundings. In contrast, a relatively lower performance was obtained in extracting urban roads and suburban roads because of the complicated scenarios. However, the performance was still promising and competitive. The advantageous performance is benefited from the following two aspects. First, by integrating

different-level and different-scale capsule features, the deep capsule feature pyramid network architecture contributes to enhance both the localization accuracy and the feature representation capability. Second, by performing feature augmentation and recalibration, the embedded CA and SA module is capable of enhancing informative features and suppressing the less useful ones to further improve the feature representation robustness. In the whole, the proposed CS-CapsFPN achieved a high overall performance (0.9470 for precision, 0.9407 for recall, 0.8957 for *IoU*, and 0.9438 for F_{score}) on the two test datasets in extracting roads of different types, diverse topologies, and varying environmental and surface conditions.

To visually inspect the road extraction accuracy, Figures 6 and 7 present two subsets of the road extraction results obtained on the GE-Road dataset and the UAV-Road dataset, respectively. As observed from the road extraction results, the majority of the roads were correctly detected and segmented. Despite the diversities in road surface materials, road topologies, and complex scenarios, the road regions were accurately differentiated from the surroundings with a very small proportion of false detections and misdetections. In addition, the extracted roads were well

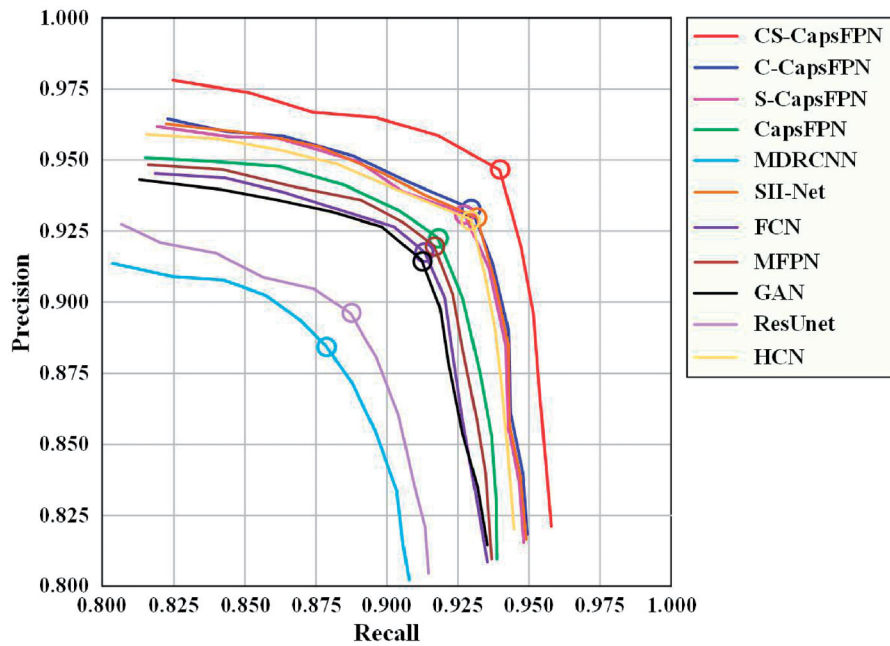


Figure 5. Illustration of the overall precision-recall curves of different models on the test datasets.

adhered to the road edges and solid in spite of the existence of shadow covers and on-road objects. Specifically, as shown in Figures 6 and 7, due to different road surface materials and illumination condition variations, the roads exhibited with different spectral and texture properties in the remote sensing images. The topologies of the roads varied greatly in width, length, and shape. In addition, some roads were covered with shadows cast by roadside objects, leading to the spectral and texture inconsistencies of the roads. Some roads were partially occluded by roadside high-rise buildings, trees, and vehicles, thereby changing the topologies of the roads. Moreover, the on-road objects in the UAV images replaced the presence of the corresponding road regions and resulted in a hole-like structure on the road surfaces, thereby affecting the consistencies of the roads. All of the above phenomena bring challenges to accurate extraction of the roads from the remote sensing images. Fortunately, owing to the advantages of the deep capsule network formulation and the SA module in abstracting high-level, distinctive, and informative feature representations and the superiorities of the feature pyramid architecture and the CA module in effectively exploiting and fusing multi-scale contextual properties, the proposed CS-CapsFPN still performed promisingly in dealing with such road images toward road extraction. However, as shown in Figures 6 and 7, some roads were severely occluded by the roadside buildings, trees, and vehicles. Thus, the proposed CS-CapsFPN failed to accurately

locate the actual road edges. In addition, some courtyards were directly connected with the roads and exhibited similar spectral and texture properties with the roads. As a result, such courtyards were falsely recognized as the road regions. Moreover, the proposed CS-CapsFPN performed less effectively on the busy road segments with many on-road vehicles due to the severe damage of the road consistencies. On the whole, designed with a novel and high-performance capsule network architecture, the proposed CS-CapsFPN was promising and feasible in extracting roads of varying surface and environmental conditions from remote sensing images.

At the test stage, the proposed CS-CapsFPN was run on the aforementioned cloud computing platform. The processing time was also recorded and analyzed with the means of processing speed to evaluate the computational performance of the proposed CS-CapsFPN. The processing speed was computed as the number of images processed per second on a GPU. On average, the proposed CS-CapsFPN achieved a processing speed of 7 images per second on a GPU. The processing speed was quite acceptable. Thus, through computational performance analysis, we concluded that the proposed CS-CapsFPN provided an efficient and promising solution to remote sensing image based road extraction tasks.

Ablation studies

As ablation studies, we demonstrated the advantageous performance of embedding the CA and SA

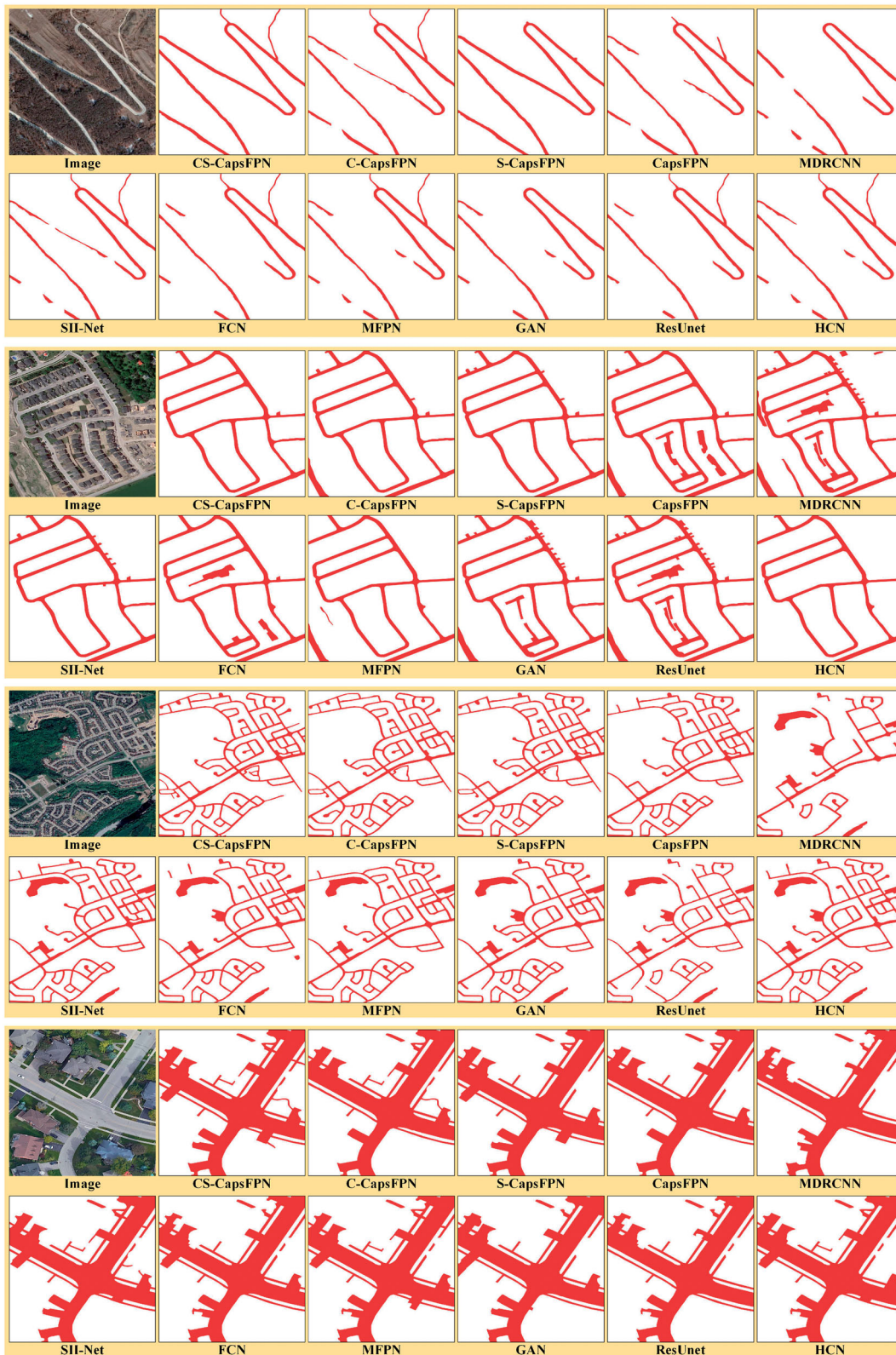


Figure 6. Illustration of a subset of the road extraction results obtained by different models on the GE-Road dataset.

modules into the capsule feature pyramid network architecture. The CA and SA modules functioned to exploit multi-scale context properties and emphasize informative features to enhance the feature

representation capabilities. To this end, we constructed three modified networks based on the CS-CapsFPN. First, we removed all the SA modules from the CS-CapsFPN (leaving only the CA modules)

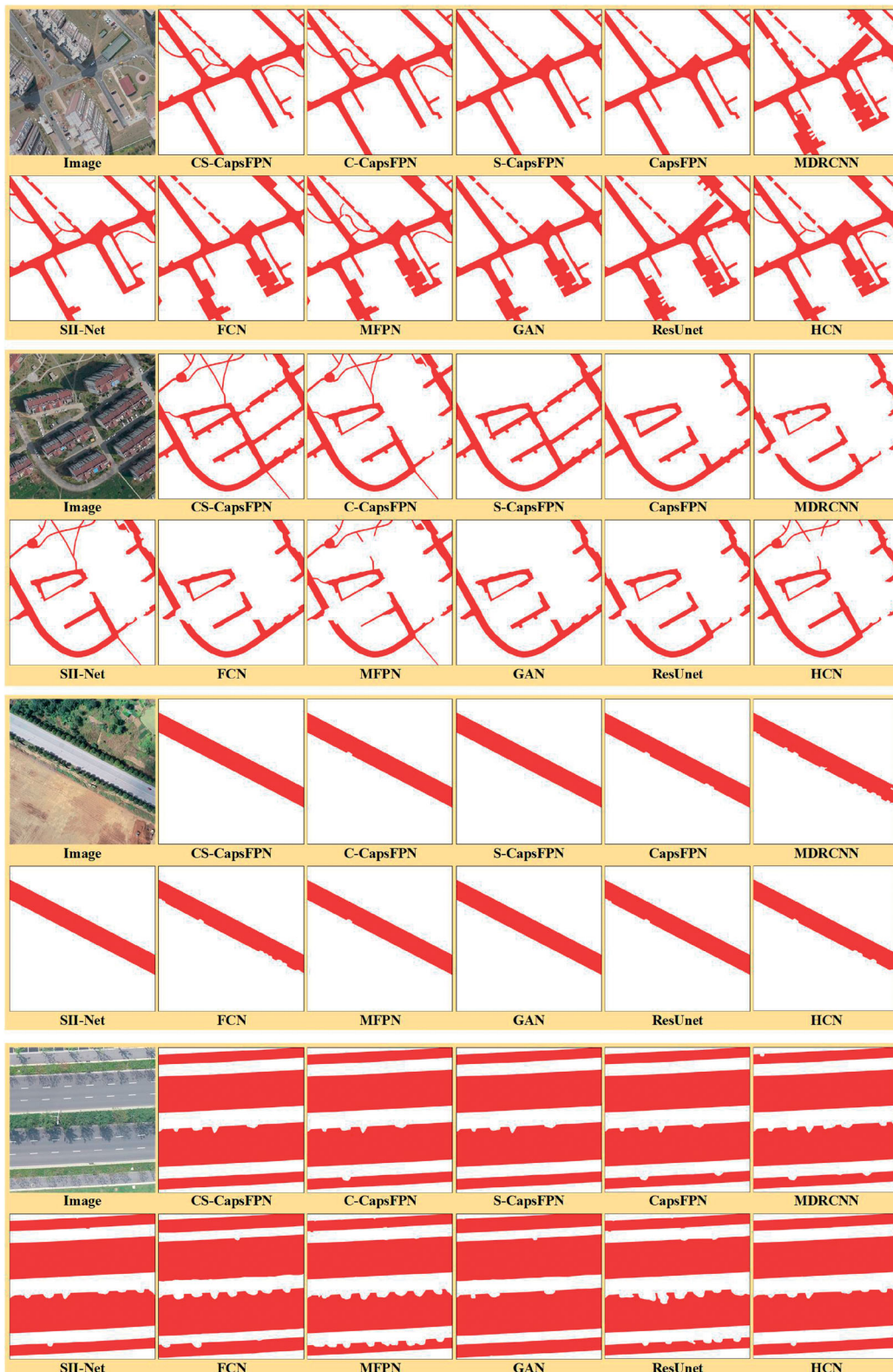


Figure 7. Illustration of a subset of the road extraction results obtained by different models on the UAV-Road dataset.

and named the modified network architecture as C-CapsFPN. Then, we removed all the CA modules from the CS-CapsFPN (leaving only the SA modules) and named the modified network architecture as

S-CapsFPN. Finally, we removed all the CA and SA modules from the CS-CapsFPN and named the modified network architecture as CapsFPN. For fair comparisons, the same training and validation sets and the

same data-augmentation and training strategies were applied to construct and optimize these three modified models. Once these models were optimized, we applied them to the same test set to analyze their road extraction performances. The quantitative evaluations of the road extraction results on the test set are reported in Table 1. In addition, the precision-recall curves of these three modified models are also provided in Figure 5. Apparently, without the CA and SA modules, the road extraction performance of the CapsFPN was degraded on both of the GE-Road and the UAV-Road datasets. The performance degradation was mainly caused by the images with the roads occluded by roadside objects, the roads covered with shadows, and the roads containing on-road objects. Comparatively, the performance degradation was more obvious on the UAV-Road dataset because of the influences of the on-road vehicles. In contrast, with the integration of the CA modules to comprehensively consider multi-scale context properties with a high-resolution perspective, the C-CapsFPN performed superiorly in processing small-scale roads, thereby effectively improving the road extraction accuracy. In addition, with the integration of the SA modules to distinguish and highlight informative features, the roads covered with shadows or containing on-road objects were accurately recognized by the S-CapsFPN, resulting in the enhancement of the road extraction accuracy. Therefore, by integrating both the CA and SA modules, the proposed CS-CapsFPN achieved a dramatic performance enhancement compared with the CapsFPN. However, the CapsFPN still performed effectively on road extraction and the road extraction accuracy was still very promising. This is actually benefited from the superior properties of the capsule formulations in abstracting high-level, distinctive, and representative features of the roads and the design of the feature pyramid architecture of the network by integrating different-level and different-scale capsule features. For visual comparisons, Figures 6 and 7 also present a subset of the road extraction results obtained by these three models.

Furthermore, the computational performances of these three modified models were also analyzed. Specifically, without the CA and SA modules embedded in each stage, the CapsFPN became less light-weight and ran faster than the C-CapsFPN and the S-CapsFPN at the test stage. On average, the CapsFPN achieved a processing speed of 8 images per second on a GPU, the C-CapsFPN achieved a processing speed of 7.3 images per second on a GPU, and the S-CapsFPN achieved a processing speed of 7.7

images per second on a GPU. Through ablation analysis, we confirmed that the road extraction performance can be effectively upgraded by embedding the CA and SA modules to take into consideration the multi-scale context properties and the channel-wise feature informativeness and saliencies.

Comparative studies

To further evaluate the accuracy and robustness of the proposed CS-CapsFPN, we conducted a group of comparative studies with some recently developed deep learning based road extraction methods. The following seven methods were selected for performance comparisons: MDRCNN (Dai et al. 2019), SII-Net (Tao et al. 2019), FCN (Zhang et al. 2020), MFPN (Gao et al. 2018), GAN (Zhang et al. 2019a), ResUnet (Zhang et al. 2018a), and HCN (Li et al. 2019a). Specifically, the MDRCNN adopted multi-size kernels with different receptive fields to obtain hierarchical features and concatenated these features to predict road maps. The HCN consisted of multiple parallel backbone branches to extract different-grained features and fused these features to conduct road extraction. The other models focused on extracting and fusing different-level and different-scale features to enhance the feature representation capabilities. For fair comparisons, the same training and validation sets and the same data-augmentation strategy were leveraged to train and optimize these road extraction models. Once these models were optimized, we applied them to the same test set to conduct performance evaluations. As shown in Table 1, the road extraction performances obtained on the test set by these models were quantitatively measured using precision, recall, IoU , and F_{score} . In addition, the precision-recall curves of these models are provided in Figure 5. For visual comparisons, Figures 6 and 7 also present some road extraction results obtained by these models. As observed in Table 1, the SII-Net and the HCN performed better than the other models with respect to the overall accuracy. The FCN, the MFPN, and the GAN obtained similar performances. In contrast, the MDRCNN and the ResUnet behaved less effectively than the other models. By developing a spatial information inference structure, the SII-Net was capable of learning both the local visual characteristics and the global spatial structure information of the roads, thereby effectively solving less severe occlusions and preserving the continuity of the extracted roads. The competitive accuracy obtained by the HCN owed to the integration of the multi-grained features

extracted by different parallel backbones, resulting in the dramatic enhancement of the output features. In addition, by fusing multi-level and multi-scale features and improving the loss functions to alleviate the imbalance issues, the FCN, the MFPN, and the GAN obtained superior performances over the MDRCNN and the ResUnet models. Comparatively, by designing the capsule feature pyramid architecture to extract and integrate different-level and different-scale high-order capsule features and embedding the CA and SA modules to effectively exploit multi-scale context properties and emphasize channel-wise informative and salient features, our proposed CS-CapsFPN exhibited advantageous performance over the seven compared models in extracting roads of varying scales.

In addition, to compare the computational efficiencies of these models, the processing time was also recorded and analyzed with the means of processing speed for each model and reported in Table 1. The processing speed was measured by the number of remote sensing images processed each second on a GPU. As shown in Table 1, the MDRCNN executed faster than the other models because of the single forward feature abstraction architecture. In contrast, the HCN was less efficient than the other models due to the hybrid network architecture with three parallel multi-grained feature extraction backbones.

Conclusion

This paper has presented a novel and high-performance deep capsule network, named CS-CapsFPN, for extracting roads from remote sensing images. Taking advantage of the superior properties of vectorial capsule representations, the CS-CapsFPN can extract different levels and different scales of inherent, distinctive, and salient features of the roads. By integrating the multi-level and multi-scale capsule features, the CS-CapsFPN provided a high-resolution and semantically strong feature representation for improving the road extraction accuracies. Benefited from the embedding of the CA and SA modules, the CS-CapsFPN is capable of considering multi-scale context properties at a high-resolution perspective and emphasizing channel-wise informative and salient features, thereby further enhancing the feature representation capabilities and improving the road extraction accuracies. The proposed CS-CapsFPN performed effectively and efficiently in processing roads of different types, diverse topologies, and complicated environmental and surface conditions. Quantitative evaluations on two large remote sensing image

datasets showed that an overall road extraction performance with a precision, a recall, an *IoU*, and an F_{score} of 0.9470, 0.9407, 0.8957, and 0.9438, respectively, was achieved. Comparative studies and detailed analysis with seven recently developed deep learning based road extraction methods also demonstrated the robust applicability and the superior performance of the proposed CS-CapsFPN in road extraction tasks.




Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the [National Natural Science Foundation of China] under Grants [62076107], [51975239], [41971414], and [41671454]; by the [Six Talent Peaks Project in Jiangsu Province] under Grant [XYDXX-098]; and by the [Natural Science Foundation of Jiangsu Province] under Grant [BK20191214].

ORCID

Yongtao Yu  <http://orcid.org/0000-0001-7204-9346>
 Haiyan Guan  <http://orcid.org/0000-0003-3691-8721>
 Jonathan Li  <http://orcid.org/0000-0001-7899-0049>

References

- Alshehhi, R., and Marpu, P.R. 2017. "Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 126: pp. 245–260. doi:10.1016/j.isprsjprs.2017.02.008.
- Bae, Y., Lee, W.H., Choi, Y., Jeon, Y.W., and Ra, J.B. 2015. "Automatic road extraction from remote sensing images based on a normalized second derivative map." *IEEE Geoscience and Remote Sensing Letters*, Vol. 12(No. 9): pp. 1858–1862. doi:10.1109/LGRS.2015.2431268.
- Bakhtiari, H.R.R., Abdollahi, A., and Rezaeian, H. 2017. "Semi automatic road extraction from digital images." *The Egyptian Journal of Remote Sensing and Space Science*, Vol. 20(No. 1): pp. 117–123. doi:10.1016/j.ejrs.2017.03.001.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A.L. 2018. "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40(No. 4): pp. 834–848. doi:10.1109/TPAMI.2017.2699184.
- Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., and Pan, C. 2017. "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55(No. 6): pp. 3322–3337. doi:10.1109/TGRS.2017.2669341.

- Cheng, G., Zhu, F., Xiang, S., and Pan, C. 2016. "Road centerline extraction via semisupervised segmentation and multidirection nonmaximum suppression." *IEEE Geoscience and Remote Sensing Letters*, Vol. 13(No. 4): pp. 545–549. doi:10.1109/LGRS.2016.2524025.
- Dai, J., Du, Y., Zhu, T., Wang, Y., and Gao, L. 2019. "Multiscale residual convolution neural network and sector descriptor-based road detection method." *IEEE Access.*, Vol. 7 pp. 173377–173392. doi:10.1109/ACCESS.2019.2956725.
- Erapu, K.K., Ashwath, B., Lal, S., Dell'Acqua, F., and Narasimha Dhan, A.V. 2019. "Dense refinement residual network for road extraction from aerial imagery data." *IEEE Access.*, Vol. 7 pp. 151764–151782. doi:10.1109/ACCESS.2019.2928882.
- Gao, L., Song, W., Dai, J., and Chen, Y. 2019. "Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network." *Remote Sensing*, Vol. 11(No. 5): pp. 552. doi:10.3390/rs11050552.
- Gao, X., Sun, X., Zhang, Y., Yan, M., Xu, G., Sun, H., Jiao, J., and Fu, K. 2018. "An end-to-end neural network for road extraction from remote sensing imagery by multiple feature pyramid network." *IEEE Access.*, Vol. 6: pp. 39401–39414. doi:10.1109/ACCESS.2018.2856088.
- He, H., Yang, D., Wang, S., Wang, S., and Li, Y. 2019. "Road extraction by using atrous spatial pyramid pooling integrated encoder-decoder network and structural similarity loss." *Remote Sensing*, Vol. 11(No. 9): pp. 1015. doi:10.3390/rs11091015.
- Hong, Z., Ming, D., Zhou, K., Guo, Y., and Lu, T. 2018. "Road extraction from a high spatial resolution remote sensing image based on richer convolutional features." *IEEE Access.*, Vol. 6: pp. 46988–47000. doi:10.1109/ACCESS.2018.2867210.
- Li, Y., Guo, L., Rao, J., Xu, L., and Jin, S. 2019a. "Road segmentation based on hybrid convolutional network for high-resolution visible remote sensing image." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 4): pp. 613–617. doi:10.1109/LGRS.2018.2878771.
- Li, Y., Xiang, L., Zhang, C., and Wu, H. 2019b. "Fusing taxi trajectories and RS images to build road map via DCNN." *IEEE Access.*, Vol. 7: pp. 161487–161498. doi:10.1109/ACCESS.2019.2951730.
- Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P. 2020. "Focal loss for dense object detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42(No. 2): pp. 318–327. doi:10.1109/TPAMI.2018.2858826.
- Liu, R., Miao, Q., Zhang, Y., Gong, M., and Xu, P. 2020. "A semi-supervised high-level feature selection framework for road centerline extraction." *IEEE Geoscience and Remote Sensing Letters*, Vol. 17(No. 5): pp. 894–898. doi:10.1109/LGRS.2019.2931928.
- Liu, Y., Yao, J., Lu, X., Xia, M., Wang, X., and Liu, Y. 2019. "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57(No. 4): pp. 2043–2056. doi:10.1109/TGRS.2018.2870871.
- Lv, Z., Jia, Y., Zhang, Q., and Chen, Y. 2017. "An adaptive multifeature sparsity-based model for semiautomatic road extraction from high-resolution satellite images in urban areas." *IEEE Geoscience and Remote Sensing Letters*, Vol. 14(No. 8): pp. 1238–1242. doi:10.1109/LGRS.2017.2704120.
- Maboudi, M., Amini, J., Hahn, M., and Saati, M. 2016. "Road network extraction from VHR satellite images using context aware object feature integration and tensor voting." *Remote Sensing*, Vol. 8(No. 8): pp. 637. doi:10.3390/rs8080637.
- Maboudi, M., Amini, J., Malihi, S., and Hahn, M. 2018. "Integrating fuzzy object based image analysis and ant colony optimization for road extraction from remote sensing images." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 138: pp. 151–163. doi:10.1016/j.isprsjprs.2017.11.014.
- Miao, Z., Shi, W., Gamba, P., and Li, Z. 2015. "An object-based method for road network extraction in VHR satellite images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 8(No. 10): pp. 4853–4862. doi:10.1109/JSTARS.2015.2443552.
- Miao, Z., Shi, W., Samat, A., Lisini, G., and Gamba, P. 2016. "Information fusion for urban road extraction from VHR optical satellite images." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9(No. 5): pp. 1817–1829. doi:10.1109/JSTARS.2015.2498663.
- Movaghati, S., Moghaddamjoo, A., and Tavakoli, A. 2010. "Road extraction from satellite images using particle filtering and extended Kalman filtering." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 48(No. 7): pp. 2807–2817. doi:10.1109/TGRS.2010.2041783.
- Pan, H., Jia, Y., and Lv, Z. 2019. "An adaptive multifeature method for semiautomatic road extraction from high-resolution stereo mapping satellite images." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 2): pp. 201–205. doi:10.1109/LGRS.2018.2870488.
- Paoletti, M.E., Haut, J.M., Fernandez-Beltran, R., Plaza, J., Plaza, A., Li, J., and Pla, F. 2019. "Capsule networks for hyperspectral image classification." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57(No. 4): pp. 2145–2160. doi:10.1109/TGRS.2018.2871782.
- Rajasegaran, J., Jayasundara, V., Jayasekara, S., Jayasekara, H., Seneviratne, S., and Rodrigo, R. 2019. "DeepCaps: Going deeper with capsule networks." Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, June 2019.
- Sabour, S., Frosst, N., and Hinton, G. E. 2017. "Dynamic routing between capsules." Paper presented at the 31st Conference on Neural Information Processing Systems, Long Beach, USA, December 2017.
- Senthilnath, J., Varia, N., Dokania, A., Anand, G., and Benediktsson, J.A. 2020. "Deep TEC: Deep transfer learning with ensemble classifier for road extraction from UAV imagery." *Remote Sensing*, Vol. 12(No. 2): pp. 245. doi:10.3390/rs12020245.
- Sghaier, M.O., and Lepage, R. 2016. "Road extraction from very high resolution remote sensing optical images based on texture analysis and beamlet transform." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 9(No. 5): pp. 1946–1958. doi:10.1109/JSTARS.2015.2449296.

- Shi, Q., Liu, X., and Li, X. 2018. "Road detection from remote sensing images by generative adversarial networks." *IEEE Access.*, Vol. 6: pp. 25486–25494. doi:10.1109/ACCESS.2017.2773142.
- Shi, W., Miao, Z., Wang, Q., and Zhang, H. 2014. "Spectral-spatial classification and shape features for urban road centerline extraction." *IEEE Geoscience and Remote Sensing Letters*, Vol. 11(No. 4): pp. 788–792. doi:10.1109/LGRS.2013.2279034.
- Tao, C., Qi, J., Li, Y., Wang, H., and Li, H. 2019. "Spatial information inference net: Road extraction using road-specific contextual information." *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 158: pp. 155–166. doi:10.1016/j.isprsjprs.2019.10.001.
- Wei, Y., Wang, Z., and Xu, M. 2017. "Road structure refined CNN for road extraction in aerial image." *IEEE Geoscience and Remote Sensing Letters*, Vol. 14(No. 5): pp. 709–713. doi:10.1109/LGRS.2017.2672734.
- Xin, J., Zhang, X., Zhang, Z., and Fang, W. 2019. "Road extraction of high-resolution remote sensing images derived from DenseUNet." *Remote Sensing*, Vol. 11(No. 21): pp. 2499. doi:10.3390/rs11212499.
- Yang, X., Li, X., Ye, Y., Lau, R.Y.K., Zhang, X., and Huang, X. 2019. "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 57(No. 9): pp. 7209–7220. doi:10.1109/TGRS.2019.2912301.
- Yu, Y., Gu, T., Guan, H., Li, D., and Jin, S. 2019. "Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 12): pp. 1894–1898. doi:10.1109/LGRS.2019.2912582.
- Yu, Y., Guan, H., Li, D., Zhang, Y., Jin, S., and Yu, C. 2020a. "CCapFPN: A context-augmented capsule feature pyramid network for pavement crack detection." *IEEE Transactions on Intelligent Transportation Systems*. Advance online publication. doi:10.1109/TITS.2020.3035663.
- Yu, Y., Ren, Y., Guan, H., Li, D., Yu, C., Jin, S., and Wang, L. 2020b. "Capsule feature pyramid network for building footprint extraction from high-resolution aerial imagery." *IEEE Geoscience and Remote Sensing Letters*, Vol. 18(No. 5): pp. 895–899. doi:10.1109/LGRS.2020.2986380.
- Yuan, J. 2018. "Learning building extraction in aerial scenes with convolutional networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40(No. 11): pp. 2793–2798. doi:10.1109/TPAMI.2017.2750680.
- Zang, Y., Wang, C., Cao, L., Yu, Y., and Li, J. 2016. "Road network extraction via aperiodic directional structure measurement." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 54(No. 6): pp. 3322–3335. doi:10.1109/TGRS.2016.2514602.
- Zang, Y., Wang, C., Yu, Y., Luo, L., Yang, K., and Li, J. 2017. "Joint enhancing filtering for road network extraction." *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 55(No. 3): pp. 1511–1525. doi:10.1109/TGRS.2016.2626378.
- Zhang, J., Chen, L., Wang, C., Zhuo, L., Tian, Q., and Liang, X. 2017. "Road recognition from remote sensing imagery using incremental learning." *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18(No. 11): pp. 2993–3005. doi:10.1109/TITS.2017.2665658.
- Zhang, X., Han, X., Li, C., Tang, X., Zhou, H., and Jiao, L. 2019a. "Aerial image road extraction based on an improved generative adversarial network." *Remote Sensing*, Vol. 11(No. 8): pp. 930. doi:10.3390/rs11080930.
- Zhang, Z., Liu, Q., and Wang, Y. 2018a. "Road extraction by deep residual U-Net." *IEEE Geoscience and Remote Sensing Letters*, Vol. 15(No. 5): pp. 749–753. doi:10.1109/LGRS.2018.2802944.
- Zhang, X., Ma, W., Li, C., Wu, J., Tang, X., and Jiao, L. 2020. "Fully convolutional network-based ensemble method for road extraction from aerial images." *IEEE Geoscience and Remote Sensing Letters*, Vol. 17(No. 10): pp. 1777–1781. doi:10.1109/LGRS.2019.2953523.
- Zhang, Y., Xia, G., Wang, J., and Lha, D. 2019b. "A multiple feature fully convolutional network for road extraction from high-resolution remote sensing image over mountainous areas." *IEEE Geoscience and Remote Sensing Letters*, Vol. 16(No. 10): pp. 1600–1604. doi:10.1109/LGRS.2019.2905350.
- Zhang, Y., Xiong, Z., Zang, Y., Wang, C., Li, J., and Li, X. 2019c. "Topology-aware road network extraction via multi-supervised generative adversarial networks." *Remote Sensing*, Vol. 11(No. 9): pp. 1017. doi:10.3390/rs11091017.
- Zhang, Z., Zhang, X., Sun, Y., and Zhang, P. 2018b. "Road centerline extraction from very-high-resolution aerial image and LiDAR data based on road connectivity." *Remote Sensing*, Vol. 10(No. 8): pp. 1284. doi:10.3390/rs10081284.
- Zhou, H., Kong, H., Wei, L., Creighton, D., and Nahavandi, S. 2017. "On detecting road regions in a single UAV image." *IEEE Transactions on Intelligent Transportation Systems*, Vol. 18(No. 7): pp. 1713–1722. doi:10.1109/TITS.2016.2622280.