Road Extraction Using a Dual Attention Dilated-LinkNet Based on Satellite Images and Floating Vehicle Trajectory Data

Lipeng Gao, Jingyu Wang, Qixin Wang, Wenzhong Shi, Jiangbin Zheng, Hongping Gan, Zhiyong Lv, and Honghai Qiao

Abstract—Automatic extraction of road from multisource remote sensing data has always been a challenging task. Factors such as shadow occlusion and multisource data alignment errors prevent current deep learning-based road extraction methods from acquiring road features with high complementarity, redundancy, and crossover. Unlike previous works that capture contexts by multiscale feature fusion, we propose a dual attention dilated-LinkNet (DAD-LinkNet) to adaptively integrate local road features with their global dependencies by joint using satellite image and floating vehicle trajectory data. First, a joint least-squares feature matching-based floating vehicle trajectory correction model is used to correct the floating vehicle trajectory; then a convolutional network model DAD-LinkNet based on a dual-attention mechanism is proposed, and road features are extracted from the channel domain and spatial domain of the target image in turn by constructing a dual-attention module in the dilated convolutional layer and adopting a cascade connection; a weighted hyperparameter loss function is used as the loss function of the model; finally, the road extraction is completed based on the proposed DAD-LinkNet model. Experiments on three datasets show that the proposed DAD-LinkNet model outperforms the state-of-the-art methods in terms of accuracy and connectivity.

Index Terms—Dual attention, floating vehicle trajectory, road extraction, satellite image.

I. INTRODUCTION

ROADS are an important public infrastructure and play an important role in the urbanization process and smart city construction. Road facilities can be used to assess the degree of urbanization of an area, and as an important indicator of urban impervious surface, it can be used to estimate the carbon emissions of a city [1], [2]. Studies show that the length and extent of global roads will expand dramatically in the 21st century, for example: by 2050, there are expected to be at least 25 million kilometers of new roads worldwide; a 60% increase in total road length compared to 2010. 90% of road construction is taking place in developing countries, including many areas with outstanding biodiversity and important ecosystem services, and the roads that penetrate these areas are a major driver of habitat loss and fragmentation, wildfires, overhunting, and other environmental degradation [3]. These issues have received considerable attention from scientists and policy makers, and focused and sustainable efforts have been made to understand the complex environmental changes that occur in areas where roads are changing rapidly [4]. Therefore, a comprehensive understanding of the spatial distribution of global roads is important for a range of research topics related to the environmental changes caused by global urbanization processes.

Remote sensing has inherently valuable features to track large-scale land cover dynamics, such as global road mapping [40]. In recent years, remote sensing technology has shown great potential to facilitate the identification of global urban roads and to assess patterns and types of urbanization processes. A variety of satellite remote sensing images (e.g., luminescent remote sensing data, MODIS, Landsat, Quickbird, SPOT, IKONOS, Gaofen, Worldview, etc.) are used to extract road information and assess urban conditions, which are critical for both natural and socio-economic issues [5].

So far, many supervised or unsupervised machine learning based methods have been developed for road extraction from remote sensing images. In particular, the former has gained wide acceptance because of its ability to efficiently and accurately segment road areas or extract road centerlines from satellite imagery. Despite its advantages, several factors challenge the practical application of supervised methods. For example, it usually requires highly representative and complete training sets and is highly dependent on the operator's expertise. In

addition, complete road information cannot be obtained due to shadow occlusion from high-resolution images. Studies show that road extraction accuracy can be improved by multisource data fusion methods, such as fusion of high-resolution images and light detection and ranging (LiDAR) data, which improves the differentiation between ground and nonground elements and facilitates the accurate extraction of road information, but also introduces new errors, such as alignment errors between images and point clouds [6]. In addition, LiDAR point cloud cannot penetrate the lush tree crown, leading to its inability to obtain complete road information. Thus, factors such as shadow occlusion and multisource data alignment errors prevent the road extraction task from acquiring complementary, redundant, and intersection-rich road features. Therefore, to obtain accurate road information, it is necessary to solve the problem of road feature extraction in the absence of information from a single data source.

Current road feature extraction has shifted from manual feature selection to automatic feature extraction based on deep learning methods. The latter automatically extracts and learns the features of the road by convolutional networks thus obtaining semantic information to segment the road with high accuracy and robustness [7]–[9]. Despite the superiority of the deep learning approach, the method relies entirely on the features of the samples themselves and does not incorporate prior knowledge of the roads, which leads to increased sample requirements and limits the accuracy and efficiency of the model. In addition, when using deep convolutional networks for feature extraction, fixed-size convolutional kernels are used without considering the uncertainty of the inherent scale of roads, and thus it is difficult to deal with multiscale roads with complex backgrounds. Therefore, there is an urgent need to develop adaptive scale convolutional neural network models for global road mapping.

In recent years, floating vehicle trajectory data have received increasing attention in the voluntary geographic information community. The floating vehicle track data originated from urban cabs contains rich information on traffic dynamics and is a low-cost, wide coverage, and highly presentable means of GPS information collection. The coordinates (i.e., latitude and longitude) of the floating vehicle trajectory data provide an opportunity to link satellite imagery. Some studies have already shown that floating vehicle data with trajectory correction can restore part of the road information obscured by tree canopy or building shadows [10].

This study proposes a new road extraction method that uses high-resolution remote sensing imagery and floating vehicle trajectory data in a near real-time manner. The main objective of this article is to develop, test, and validate a road extraction method incorporating heterogeneous road features, which aims to extract heterogeneous road features from remote sensing images and floating vehicle trajectory data and apply them to a deep convolutional neural network to accomplish accurate road information extraction. Specifically, we introduce a spatial attention module (SAM) and a channel attention module (CAM) to capture road contextual information in the spatial and channel dimensions, respectively.

The rest of this article is organized as follows. Section II reviews the existing literature on road extraction from remotely sensed images and the application of floating vehicle trajectory data in remote sensing. In Section III, we describe the design framework of road extraction in detail. The proposed method is applied to selected test cases, and the experimental results are reported in Section IV. Finally, Section V concludes this article.

II. RELATED WORK

In this section, we present an overview of the applications of deep learning-based road extraction and floating vehicle trajectory data in remote sensing.

A. Deep Learning-Based Road Extraction

Deep learning method has developed rapidly in the past decades and attracted the attention of experts in the field of road extraction. In 2010, Minh *et al.* attempted to apply neural network technology to road extraction tasks with the city-level spatial land cover [11]. Most of the recently developed algorithms are based on FCNs with an encoder-to-decoder architecture. To simplify the training of deep neural networks, Zhang *et al.* proposed the residual U-Net, which combines the advantages of U-Net with residual blocks [12]. To extract roads of various widths, Gao *et al.* proposed a multifeature pyramid network (MFPN) [13]. Chen *et al.* proposed a reconstruction bias U-Net for road extraction from high-resolution optical remote sensing images [14]. D-Linknet [15] is an efficient method in comprehensive performance of road extraction, which won first place in the 2018 DeepGlobe Road Extraction Challenge. Despite its advantages, the down sampling process of the encoder module usually leads to a reduction in the boundary accuracy of the road extraction results. To further improve the road extraction accuracy, some scholars proposed a cascaded neural network approach based on multitask learning to obtain road surfaces, road centerlines, and road boundaries simultaneously [7], [8], [16]–[18]. These approaches use mutually constrained multiple tasks to achieve further improvement in road extraction accuracy. However, in the case, where the roads are obscured by vegetation or shadows, the contextual information modeling approach based on local receptive fields cannot establish the topological relationships between the road segments separated by trees or shadows, which leads to interrupted results of road extraction.

One of the solutions for the interruption of road extraction results is to postprocess the pixel-level segmentation results of roads. Chen *et al.* proposed a two-stage approach combining road edge features and area features, which applies connection analysis to discrete line elements with directional consistency to extract potential road objects; potential road objects are then evaluated by shape features to refine the road extraction results [19]. Gao *et al.* first used a semantic segmentation model to obtain pixel-level road segmentation results, and then used tensor voting to connect the breaks [20]. Inspired by human behavior of annotating roads, Batra *et al.* proposed a stacked multibranch convolution model and connectivity optimization method based on "orientation learning" [21]. This method can correctly predict road topology and connect blocked roads. Tao *et al.* focused

on the modeling of road context information and put forward a well-designed spatial information reasoning structure [22]. These methods improve the continuity of the road network at the pixel level. However, they do not construct the road network topology.

Another solution idea is to use generative adversarial networks to estimate roads that are covered by trees or shadows. To this end, Zhang *et al*. developed a novel road extraction method based on an improved generative adversarial network [23]. Yang *et al*. designed a recurrent CNN unit that exploits spatial background and low-level visual features, thereby alleviating detection problems caused by noise, occlusion, and complex road backgrounds [24]. Zhang *et al*. proposed a multisupervised generative adversarial network, which learns how to reconstruct obscured roads based on the relationship between visible road areas and road centerlines [25]. Although these methods can solve part of the road occlusion problem, it is still difficult to obtain accurate and complete road information using only a single data source in the face of complex road background.

B. Applications of Floating Vehicle Trajectory Data in Remote Sensing

Floating vehicle trajectory data from urban cabs contains rich traffic dynamic information, which is a low-cost, wide coverage and high presentational GPS information collection means. Some researchers have shown that floating vehicle trajectory data with trajectory correction can restore the road information under the shadow of tree canopy or buildings [10].

Considering the wide distribution of floating vehicle trajectories, probabilistic and statistical based methods are the main methods to extract road information from low precision GPS trajectories [26]. To ensure the extraction accuracy, the trajectory data must be preprocessed first, i.e., the trajectory correction. Existing correction methods (e.g., Hidden Markov Model-based [27], incremental route-based [28], and location sequence-based [29] methods, etc.) usually rely on existing map data and require multiple iterations, which is less efficient if long sequences of floating vehicle trajectory data are encountered. At the same time, sparse or interwoven GPS trajectories can cause the algorithm to mis-match trajectory points at intersections or multilane roads [30].

The extraction of trajectory features usually uses clustering methods, which include distance-based clustering, identifier-based clustering, and map-based clustering. Among the distance-based methods, Euclidean distance is a popular metric to measure the similarity of trajectories [31], but the method incurs a huge computational overhead because of the need to calculate the distance for each pair of trajectories [32]. Similarly, the identifier-based approach incurs additional communication and computational overhead [33]. The map-based approach achieves trajectory clustering by matching the baseline map with the intermediate representation of the trajectory, but it cannot achieve high precision clustering well due to the complexity of the road map [34].

Although the road extraction method based on GPS trajectory has many advantages, it also has the following shortcomings: 1) due to the positioning accuracy, the floating vehicle trajectory
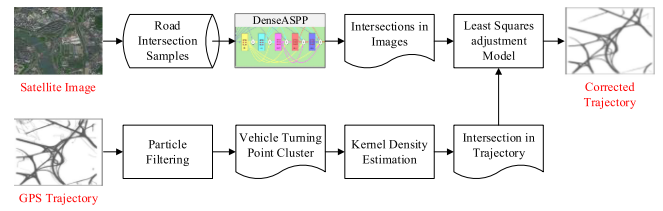

Fig. 1.   Flowchart of floating vehicles trajectory correction.

will appear to offset the road phenomenon. For such problems, existing studies focus on matching floating vehicle trajectories with maps to eliminate position offset, while there are fewer studies on matching floating vehicle trajectories with remote sensing images; 2) although massive floating vehicle data can accurately extract road information, the large data volume leads to an exponential increase in geometric complexity, computational complexity, and storage space. Meanwhile, in emergency or less economically developed areas, floating vehicle trajectory data are scarce, and the road information extracted by using floating vehicle trajectories alone is incomplete, so it needs to be combined with other data sources for joint extraction.

III. PROPOSED APPROACH

The objective of this study is to design an efficient approach to extract road information from open data, including satellite images and floating vehicle trajectory data. The proposed method consists of the following four main steps: first, a least squares-based trajectory correction method was used for floating vehicle data filtering and registering; then a dual attention module was proposed to fusion road features; and the module was used to construct a dual attention based dilated-LinkNet architecture; then, a weighted loss function was described in detail; finally, transfer learning for pretrained model is introduced.

A. Least Squares-Based Trajectory Correction for Floating Vehicles

The first step is to filter the floating vehicle trajectory data. The key is the difference of the distribution characteristics between the floating vehicle trajectory data falling into the road and the background in the "space–time–spectrum" dimension, and this article proposes to use particle filtering to eliminate the redundant floating vehicle trajectory data [35].

Then, the remaining trajectory data are corrected using satellite images. Fig. 1 gives the flowchart of joint correction of floating vehicle trajectory based on least squares feature matching. First, the vehicle turning point class clusters are extracted from the filtered floating vehicle trajectory data, then the road intersections are extracted from satellite images and vehicle turning point data using deep learning and similarity clustering method, respectively [36], [37]. Finally, the least square method is used to match and fit the extracted intersection centroid data, and an adjustment model is constructed to correct the floating vehicle trajectory data.

Assuming the road intersections extracted from the images as an independent point set P and the road intersections extracted

from the floating vehicle trajectories as another independent point set Q, find a set of transformations (R, T) consisting of a rotation matrix R, and a translation matrix T. The alignment matching transformation of the two point-sets minimizes the following objective function:

$$E(R, T) = \frac{1}{N} \sum_{i=1}^{N} \| p_i - (Rq_i + T) \|^2 \qquad (1)$$

where the rotation matrix R and translation matrix T are the rotation parameters and translation parameters between the floating vehicle trajectory data and images, such that the optimal matching between two intersection point sets satisfies the Euclidean distance minimization criterion.

The algorithm is an iterative process with the following steps.
1) Search the point set P and Q with initial value Q_0, define the maximum number of iterations k_{max}, initialize the rotation matrix R and translation matrix T.
2) For each point q_i in the target point set Q, find the point p_i with the smallest Euclidean distance from the point set P to form a point pair (p_i, q_i).
3) Use the point pairs in 2) to calculate the rotation matrix R and the translation matrix T and substitute the calculated R and T into (1) to minimize its value.
4) The transformation matrix after k iterations is expressed as R_k and T_k, and then Q is calculated:

$$Q = R_k Q_0 + T_k \qquad (2)$$

5) If the value of $E(R, T)$ is greater than the set threshold and at the same time the number of iterations has not reached k_{max}, then restart the iteration, and vice versa stop the iteration and exit.

B. Dual Attention Module

In the field of image recognition, feature extraction based on attention mechanism belongs to an emerging theoretical system. The core applied idea of the attention mechanism lies in how to make the system model learn to focus only on the information that needs to be studied in the research and choose to ignore the information that is not related to the research project.

The introduction of the attention mechanism allows flexible weighting of road features from different data sources during the learning process, thus enhancing the accuracy of the valid road features eventually acquired and thus extracting road contours more efficiently from the background interference information (e.g., vehicles, pedestrians, standing water, sunlight reflections, etc.) of complex data images.

This study adopts the image numerical mask to construct the attention mechanism, which is implemented by designing a new weight distribution layer to extract and mark the core road feature information from the target image, and then extract the road features from the spatial domain and channel domain of the target image in turn to form dual attention.

1) Spatial Attention Module: As shown in Fig. 2, a local feature map $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ is first passed through three convolutional layers to obtain three feature maps A, B, and C, respectively, where $\{\mathbf{A}, \mathbf{B}, \mathbf{C}\} \in \mathbb{R}^{C \times H \times W}$. Then they are reshaped
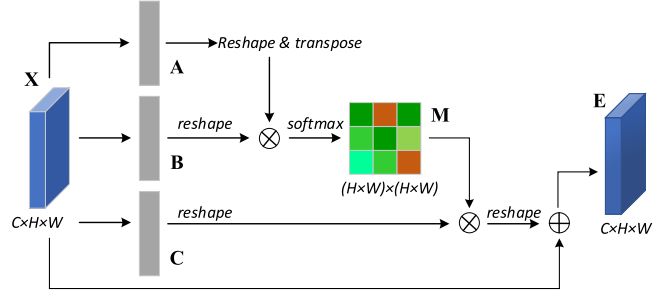

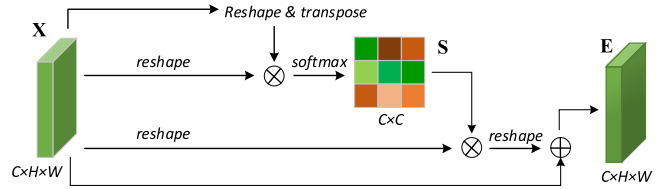Fig. 2. Spatial attention module.

Fig. 3. Channel attention module.

to $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of pixels, after which the transpose of the reshaped A is multiplied with the reshaped B and then the spatial attention map $\mathrm{M} \in \mathbb{R}^{N \times N}$ is obtained by softmax

$$m_{ji} = \frac{\exp(A_i \cdot B_j)}{\sum_{i=1}^{N} \exp(A_i \cdot B_j)} \qquad (3)$$

where m_{ji} measures the ith position's impact on the jth position. Note that the more similar feature representations of the two positions contribute to greater correlation between them.

Then, we perform a matrix multiplication between C and the transpose of M and reshape the result to $\mathbb{R}^{C \times H \times W}$. Finally, it is multiplied by the scale factor α and summed with X to get the final output spatial attention matrix E

$$E_j = \alpha \sum_{j=1}^{N} (m_{ji} C_i) + X_j \qquad (4)$$

where α is initialized as 0 and gradually learns to assign more weight. It can be inferred from (4) that the resulting feature E at each position is a weighted sum of the features at all positions and original features.

2) Channel Attention Module: The structure of CAM is illustrated in Fig. 3. Different from the SAM, we directly calculate the channel attention map $\mathbf{S} \in \mathbb{R}^{C \times C}$ from the original features $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$. Specifically, we reshape X to $\mathbb{R}^{C \times N}$, and then perform a matrix multiplication between X and the transpose of X. Finally, we apply a softmax layer to obtain the channel attention map $\mathbf{S} \in \mathbb{R}^{C \times C}$:

$$S_{ji} = \frac{\exp(X_i \cdot X_j)}{\sum_{i=1}^{C} \exp(X_i \cdot X_j)} \qquad (5)$$

where s_{ji} measures the ith channel's impact on the jth channel. In addition, we perform a matrix multiplication between the transpose of S and X and reshape their result to $\mathbb{R}^{C \times H \times W}$.
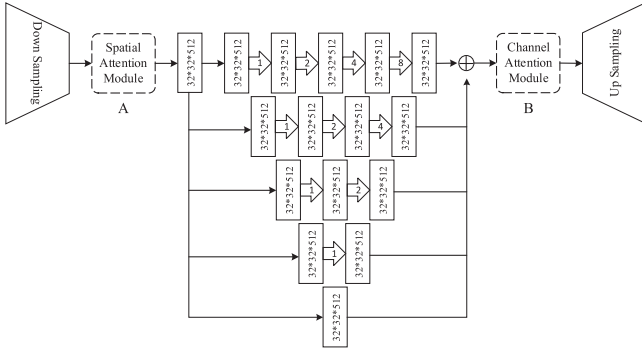
Fig. 4. Architecture of the proposed DAD-LinkNet.

Then, we multiply the result by a scale parameter β and perform an element-wise sum operation with X to obtain the final output $E \in \mathbb{R}^{C \times H \times W}$

$$E_j = \beta \sum_{i=1}^{C} (s_{ji} X_i) + X_j \qquad (6)$$

where β gradually learn a weight from 0. The Equation shows that the final feature of each channel is a weighted sum of the features of all channels and original features, which models the long-range semantic dependencies between feature maps. It emphasizes class-dependent feature maps and helps to boost feature discriminability.

In addition, the original D-LinkNet model's segmentation network structure is a direct sampling operation using linear interpolation, which can be considered as a reproduction of the classical end-to-end neural network architecture, and the entire model is built without introducing redundant learning parameters and without the need for extensive additional computation for other parameters. Therefore, when introducing the attention mechanism, the attention module can be directly added to the D-LinkNet network without worrying about the learning time due to the explosive increase in computation and running time, which may lead to a decrease in the feasibility of the algorithm.

C. DAD-LinkNet Architecture

The architecture of the proposed DAD-LinkNet is an encoder–decoder model. The network is built with D-LinkNet architecture and has dilated convolution layers in its center part. Dilation convolution is a powerful tool that can enlarge the receptive field of feature points without reducing the resolution of the feature maps.

This study introduces the self-attentive mechanism by referring to the dual-attention network structure, which adopts a cascade as the connection method to extract the road information from the spatial domain of the vehicle trajectory and the channel domain of the image to obtain the road features in the image more accurately.

As shown in the Fig. 4, the dotted box A and B parts are two attention modules. In the modified version of D-LinkNet structure, the attention module A is connected to the left side of the pretrained ResNet34 structure as an encoder to preactivate

the feature extraction capability of the network, and the five branches in the central region of the model are added to the dilated convolution operation module. Through the "pooling" operation of dilated convolution, the range of perceptual field can be increased without losing feature information, so that the rod features can be extracted more efficiently and the fusion of features with multiple depths and sizes can be accomplished. The fusion method of feature "stacking" is adopted, and there is no postprocessing operation.

The arrows with numbers in the Fig. 4 indicate the depth of the convolutional neural network, the depth of the five branches is from 5 to 1 in order, and the size of the receptive field is set to 15, 7, 3, 1, and 0. Afterward, a 1*1 convolutional layer is added for feature fusion. Finally, the road prediction probability map with the same size as the initial input image is obtained by using the normalization function and activation function. The final segmented road binary image is obtained.

D. Loss Function

Loss function is a nonnegative real-valued function used in the machine learning process to estimate the difference between the predicted value and the true value obtained by the model after learning. The smaller the loss function, the better the performance of the model. In the current research field of image segmentation, the main two ways of constructing the loss function are binary cross entropy loss (BCE) and dice coefficient loss (DICE).

BCE enables the output prediction to match the road dataset to the maximum extent, which can satisfy the optimization to reduce the error; while DICE, as a function to measure the similarity between sets, can be used to compare the similarity between two samples. The gradients of these two loss functions are calculated differently, and the gradient value of DICE is much larger than the gradient value of BCE. Referring to the parameter settings in DANet, a concept of hyperparametric loss function is introduced here. The new hyperparametric loss function is weighted by the two loss functions BCE and DICE, and the different weighting ratios of these two loss functions are adjusted by conducting several experiments, to optimize the final road segmentation results and find the best parameter values to make the results with the best accuracy. The hyperparametric loss is calculated by

$$H_{\text{loss}} = H(p, q) + \lambda \times \text{DiceLoss} \qquad (7)$$

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log(q(x_i)) \qquad (8)$$

$$\text{DiceLoss} = 1 - \frac{2|X \cap Y|}{|X| + |Y|} \qquad (9)$$

where H_{loss} denotes the hyperparametric loss, $H(p, q)$ is the BCE loss, *DiceLoss* is the DICE loss, λ denotes the weight ration of DICE to BCE, $|X|$ denotes the ground truth, $|Y|$ denotes the predict road mask.

E. Transfer Learning for Pretrained Model

Transfer learning is an extremely effective method currently used in the field of computer vision. Its basic idea is to make full use of the correlation between existing knowledge and learning objectives, and transfer knowledge from the existing model to the new learning model. In deep learning, transfer learning involves sharing the expressions and connection weights of some common features learned by deep neural networks.

Because the common features of the same data type (e.g., urban roads) are not much different, we first select some urban road datasets from the Massachusetts Road Dataset and NWPU VHR-10 Dataset for pretraining and use the weight sharing method to make the model parameters initialized at the position near the optimal parameters. As a result, transfer learning can reduce the initial error (loss), speed up the convergence of training, and improve accuracy.

IV. EXPERIMENTAL RESULTS

To evaluate the proposed method, we carry out comprehensive experiments on three datasets. To assess extraction accuracy, we use a confusion matrix to evaluate the model performance. Four metrics are utilized to evaluate the extraction performance, including the following:

1) Precision (P): $P = \frac{TP}{TP+FP}$
2) Recall (R): $R = \frac{TP}{TP+FN}$
3) F1-score (F): $F = \frac{2TP}{2TP+FP+FN}$
4) IoU: $IoU = \frac{TP}{TP+FP+FN}$.

Here, TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

Experimental results demonstrate that DAD-LinkNet achieves state-of-the-art performance on three datasets. In the next sections, we first introduce the datasets and implementation details, then we perform a series of road extraction experiments on the three datasets.

A. Dataset Description

In our experiments, two well-known remote-sensing datasets: Massachusetts Road Dataset [38], NWPU VHR-10 dataset [39], and a Wuhan dataset containing satellite image and floating vehicle data are used for evaluation.

Massachusetts Road Dataset: This dataset is a common dataset in the field of road extraction, which contains road network images of major roads in Marseilles, and it includes various road types such as urban roads, suburban roads, and rural roads, as well as some disturbing factors (pedestrians, vehicles, standing water, etc.) commonly found in the field of road network extraction. The 1171 road images of the Massachusetts road dataset were cleaned to remove the 398 images with severe missing images, and 700 of them were used as the sample data source for pretraining.

NWPU VHR-10 Dataset: This dataset contains 800 high-resolution satellite images that were cropped from Google Earth and Vaihingen datasets and then manually annotated by experts. The dataset is divided into 10 categories (aircraft, ships, storage tanks, baseball fields, tennis courts, basketball courts, surface
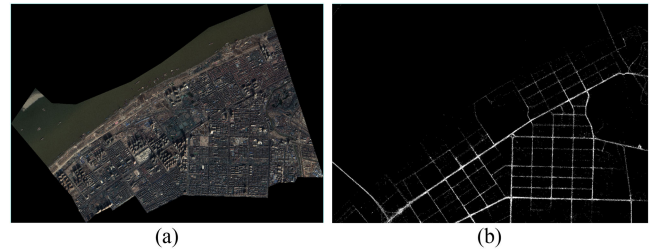

Fig. 5. The spatial extent of Wuhan dataset. (a) is the imagery data, (b) is the floating vehicle trajectory data.

runways, ports, bridges, and vehicles). For better road extraction experiments, 325 of these images suitable for road extraction learning are manually selected as samples.

Wuhan Dataset: This dataset, as shown in Fig. 5, consists of a four-band satellite image and floating vehicle data. The imagery was collected by the Pleiades satellite and shows an area of Wuhan city. The image has a spatial size of 9331 pixels × 13367 pixels and a spatial resolution of 0.5 m/pixel. The floating vehicle trajectory data was collected through GPS instruments on cabs in Wuhan. The collection period was from 04/24/2015 to 04/30/2015.

B. Data Preprocessing

A series of preprocessing operations are performed on different images in the dataset to reduce the learning time. Since the training time increases exponentially due to the large resolution of the target image, the images used in the dataset are first segmented. The segmented images are uniformly processed into square image blocks of 512 pixels in length and width. Second, to improve the accuracy of the result, a small block of 512 pixels in length and width is also cropped to achieve a simple data enhancement. Finally, some of the images are randomly selected to be rotated, mirrored, and color adjusted to further enhance the data enhancement effect.

The floating car trajectory data is processed by thinning algorithm, a buffer zone is set in the trajectory data map to select samples, the vector point data is converted into corresponding rasterized data by ArcGIS software, and vectorized road network geometric data can be obtained by coordinate transformation and accuracy calibration operations, and then this part of data is processed according to the format of image dataset as the expansion of dataset.

C. Implementation Details

We conducted all the experiments on an HP Omen Station with the following specification: Central processing Unit Intel i7-7700HQ 2.80GHz with a RAM of 32GB and an NVIDIA GeForce GTX 2080 Graphical Processing Unit (with a RAM of 16GB) . All codes were implemented using Pytorch 1.5, and the programming environment is Python 3.6.

We employ a poly learning rate policy, where the initial learning rate is multiplied by $(1 - \frac{iter}{total\_iter})^{0.9}$ after each iteration. The base learning rate is set to 0.01 for Massachusetts Road
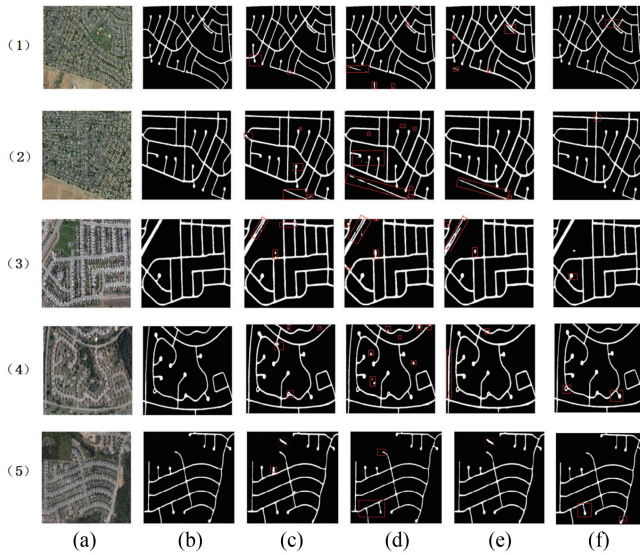
Fig. 6. Road extraction results of experiment 1. (a) original image. (b) ground truth. Road extraction results using (c) D-Link Net, (d) U-Net, (e) DAD-LinkNet (weighted loss function), (f) DAD-LinkNet (unmodified loss function).

TABLE I
RESULTS OF ROAD EXTRACTION IN EXPERIMENT 1

Methods	Precision	Recall	F1-score	IoU
U-Net	0.7723	0.7604	0.7138	0.6564
D-LinkNet	0.8308	0.7498	0.7719	0.6447
DAD-LinkNet (unmodified loss function)	**0.8553**	0.7702	0.7802	0.6576
DAD-LinkNet (weighted loss function)	0.8430	**0.7892**	**0.7814**	**0.6774**

Dataset. Momentum and weight decay coefficients are set to 0.95 and 0.001, respectively.

The proposed model is trained with synchronized BN. Batch-size are set to 4 for Massachusetts Road Dataset and 6 for other datasets. When adopting multiscale augmentation, we set training time to 120 epochs for NWPU VHR-10 Dataset and 140 epochs for other datasets. The weighted loss function on the end of the network is adopted when two attention modules are used.

D. Experiment 1

In the first experiment, the presented method is validated using the images from Massachusetts Road Dataset and NWPU VHR-10 Dataset. In the experiment, a total of 1315 images with roads were used and increased to 3010 images by data enhancement and complementation, and these datasets were divided into training, testing, and validation sets in the ratio of 7:2:1.

Fig. 6 shows the visual comparison among the proposed model, i.e., DAD-LinkNet, with U-Net, and D-LinkNet. It can be found that incomplete extractions exist in all methods. However, in terms of road integrity, DAD-LinkNet can obtain a more complete road structure. Table I records the quantitative results

of the comparison among the selected networks. From the experimental metrics, DAD-LinkNet achieves higher scores than other methods in all evaluating metrics. This phenomenon can be explained by the fact that the image features extracted by the encoder component are shared in the latter dual attention module, leading to the correlation between spatial and channel domain output results to a certain extent.

Due to the complexity of the roads in the original image, the road extraction results obtained based on D-LinkNet network and U-Net network have some blurring, adhesion, and misjudgment of the adjacent roads, such as the parts marked in red box in Fig. 6(c) and (d). In particular, the extraction results based on the D-LinkNet network may even result in "broken roads" due to misjudgment, and there is also the problem of incomplete road structure compared with the ground truth due to omission in the extraction and acquisition of road structure information. DAD-LinkNet can effectively extract the main structure of the road, preserving the details even in some challenging scenes, which proves the effectiveness and superiority of our proposed method.

E. Experiment 2

The floating vehicle trajectory data in Wuhan dataset was used in the second experiment. It is composed of a series of trajectory points, each trajectory point contains the coding information, position information, and time information of the vehicle. Therefore, the spatial information of the road network is extracted from these data, and the corresponding trajectory lines are formed by connecting the trajectory points, followed by making a buffer zone of distance γ for each trajectory line and fusing them into the surface elements of single components. Because the trajectories of local roads are sparse, when the maximum distance between trajectories exceeds twice the radius of buffer γ, the fused surface elements of road buffers are easy to form holes inside the roads, and it is necessary to fill the holes according to the area of holes using mathematical morphology before binarization. The rasterized floating vehicle trajectory data can reduce the computational complexity on the one hand, and at the same time can compensate for the shadows and occlusions in the remote sensing images, providing complementary road features.

Fig. 7 shows the results extracted from the remote sensing image and the floating vehicle trajectory data, and the fused results, respectively. Based on the proposed DAD-LinkNet model, it can better extract small roads and complex road networks that may exist in the original image, accurately segment narrow road edges on the extracted road contours without blurred pixels, and more efficiently, accurately, and completely preserve the road details in remote sensing images. It can also cope with situations, where there is no floating vehicle trajectory. However, there are some discontinuities in the road results extracted based on DAD-LinkNet model due to shadows or vegetation occlusion, etc. And the roads obtained based on floating vehicle trajectory data can make up for the deficiencies in the results extracted by DAD-LinkNet model, and more complete and accurate results can be obtained by fusing the two results. According to the
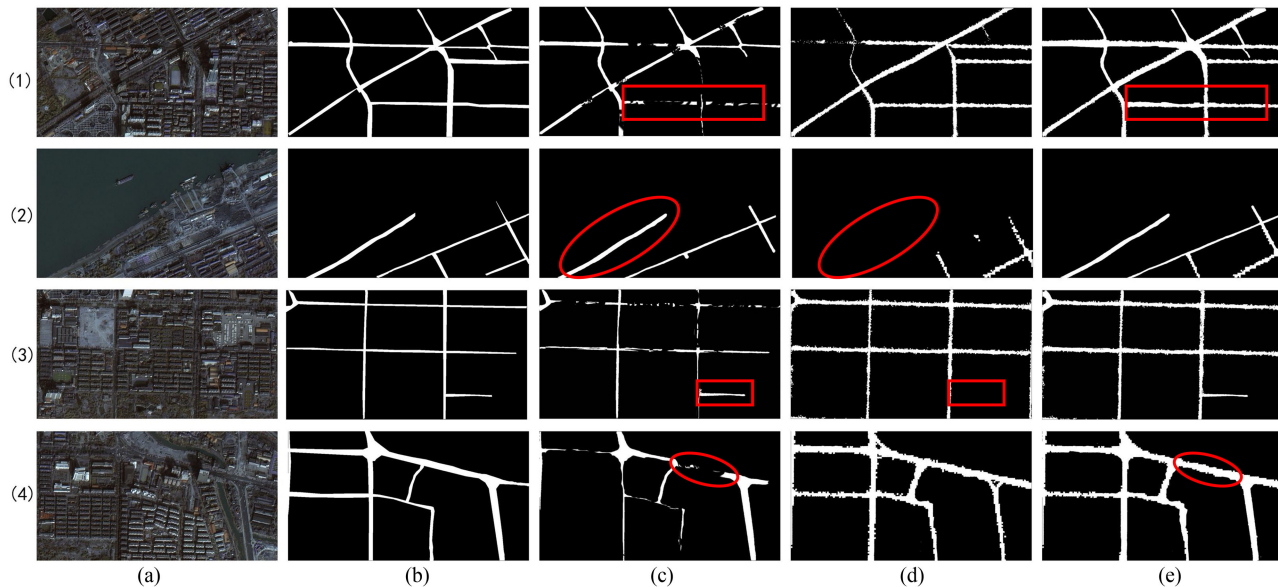
Fig. 7. Road extraction results of experiment 2. (a) original image. (b) ground truth. Road extraction results using (c) DAD-LinkNet, (d) rasterized GPS data, (e) DAD-LinkNet +GPS data.

TABLE II
RESULTS OF ROAD EXTRACTION IN EXPERIMENT 2

Methods	Precision	Recall	F1-score	IoU
GPS data	0.6467	0.5477	0.5777	0.4327
DAD-LinkNet	**0.9526**	0.6307	0.7549	0.6111
DAD-LinkNet+GPS data	0.8079	**0.7941**	**0.8012**	**0.6592**

TABLE III
EXPERIMENTAL RESULTS OF ROAD EXTRACTION WITH DIFFERENT WEIGHTS

Ratios of a and b	1：2	1：3	1：4	1：5
IoU	0.963	0.972	0.979	0.967
Recall	0.763	0.769	0.779	0.774
Precision	0.834	0.837	0.847	0.842

experimental results, the quality and completeness of the whole road extraction are improved by fusing the road results extracted from the satellite images and the floating vehicle trajectory data. This is also verified from the statistics in Table II.

V. DISCUSSION

In this section, we explore the influence of different weight parameter selections and the attention mechanism and weighted loss function on the experimental results.

A. Optimal Weighting Parameter Test

For the selection of the weight parameters of the loss function, to obtain the most suitable parameters, let the weight coefficient of BCE loss be a and the weight coefficient of DICE loss be b. For the convenience of implementation, set the value of a to 1 and adjust the value of b. Since the impact of DICE loss function on the final loss of the network model is much smaller than the impact of BCE function on the final loss of the model, several groups of ratios of a and b are preset here for the test as 1:2, 1:3, 1:4, and 1:5, respectively. These ratios are used as the weight ratios for the a priori test, and the final error values are calculated under these ratios to find the weight ratio with the best results.

After setting the corresponding ratios of a and b parameters, the learning iterations are performed separately, and the final performance is compared under different parameters. Table III shows the different evaluation metrics of the road extraction results with different ratios, from which it can be found that the improved network model has the best performance when the ratio of a and b is 1:4.

B. Attention Mechanism and Weighted Loss Function Can Improve the Quality of Road Extraction

To investigate the effects of attention mechanism and loss function on road extraction results, U-Net network, D-LinkNet network, and DAD-LinkNet network (unmodified loss function) and DAD-LinkNet network (weighted loss function) were used for road extraction, and the changes of precision, recall, and IoU of different network models after increasing the number of iterations were recorded, and the results are shown in the following figures.

1) In terms of precision. As can be seen from Fig. 8, when the number of iterations is less than 60, the D-LinkNet network, DAD-LinkNet network (unmodified loss function), and DAD-LinkNet network (weighted loss function) have little difference in precision, but all significantly outperform the U-Net network. After 90 iterations,
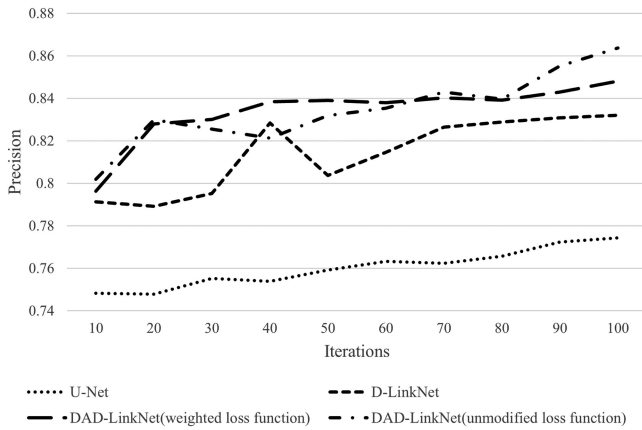
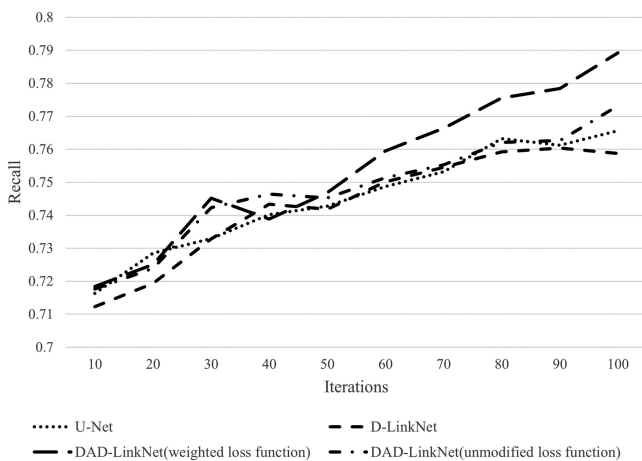Fig. 8.   Variation curve of precision index with iteration.

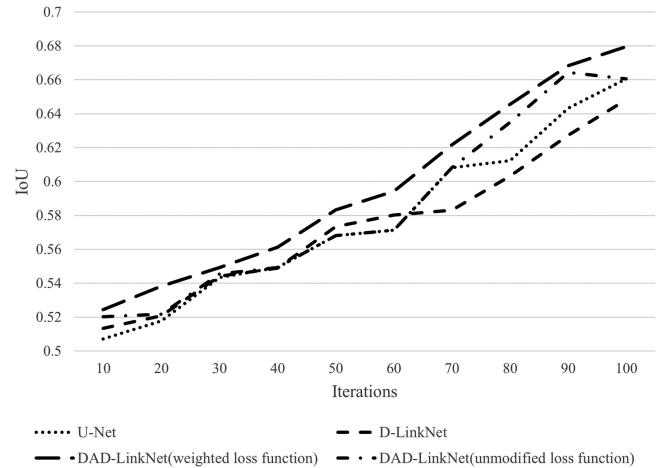Fig. 9.   Variation curve of recall index with iteration.

Fig. 10.   Variation curve of IoU index with iteration.

Moreover, after the number of learning iterations reached 100, the recall rates of the two improved DAD-LinkNet networks improved compared to the U-Net and D-LinkNet network models.

3) In terms of IoU. As can be seen from Fig. 10, the IoU of DAD-LinkNet network (weighted loss function) exceeds the other three networks after the number of learning iterations is greater than 65, which illustrates the rationality of improving the D-LinkNet network based on the attention mechanism in this paper; meanwhile, the IoU of DAD-LinkNet network (weighted loss function) has been greater than the IoU of DAD-LinkNet network (unmodified loss function) since the beginning of the iteration, which illustrates that the weighted loss function proposed in this article can improve the quality of road extraction results.

Referring to the above three evaluation metrics, it can be found that the comprehensive performance of the two improved DAD-LinkNet networks has been significantly optimized over both U-Net and D-LinkNet networks. Moreover, the DAD-LinkNet network with weighted loss function is significantly better than the network with the original version of loss function, which proves the superiority of the DAD-LinkNet network proposed in this article.

VI. CONCLUSION

In this article, we have presented a modified road extraction network DAD-LinkNet for multisource heterogeneous remote sensing data, which adaptively integrates local semantic features using the self-attention mechanism. To address the distribution differences between the source and target domains, we introduce a dual attention mechanism to capture heterogeneous road features and improve the loss function by introducing a weight parameter to improve the accuracy of the road extraction results. The experiment results demonstrate that this method can effectively improve the accuracy of road extraction from multisource heterogeneous remote sensing data.

the DAD-LinkNet network (weighted loss function) significantly outperforms the D-LinkNet network in terms of precision and improves the precision compared to the DAD-LinkNet network (unmodified loss function).

2) In terms of recall. From the experimental results shown in Fig. 9, when the number of learning iterations is small, D-LinkNet network, DAD-LinkNet network (unmodified loss function), and DAD-LinkNet network (weighted loss function) do not differ much, but U-Net network significantly outperforms the other methods. However, as the number of iterations increases, both modified DAD-LinkNet networks show a tendency to outperform the U-Net as well as the D-LinkNet networks in terms of recall. Also, by comparing the average rate of change of recall (i.e., the average slope in the image) during the 100 iterations of learning, it can be found that the average rate of change of recall of both improved DAD-LinkNet networks is greater than that of the D-LinkNet network, showing that the improved network structure is more efficient in learning compared with the original version of the D-LinkNet network as well as the U-Net network.

REFERENCES

[1] B. Bhatta, *Analysis of Urban Growth and Sprawl From Remote Sensing Data*. New York, NY, USA: Springer, 2010.

[2] R. P. Greene and J. B. Pick, *Exploring the Urban Community: A GIS Approach*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2012.

[3] W. F. Laurance and A. Balmford, "A global map for road building," *Nature*, vol. 495, no. 7441, pp. 308–309, Mar. 2013.

[4] P. D. Ellis and M. Roberts, *Leveraging Urbanization in South Asia: Managing Spaitial Transformation for Prosperity and Livability*. Washington, DC, USA: World Bank, 2016.

[5] R. Lian, W. Wang, N. Mustafa, and L. Huang, "Road extraction methods in high-resolution remote sensing images: A comprehensive review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 5489–5507, Sep. 2020.

[6] J. Martínez Sánchez, F. Fernández Rivera, J. C. Cabaleiro Domínguez, D. López Vilariño, and T. Fernández Pena, "Automatic extraction of road points from airborne LiDAR based on bidirectional skewness balancing," *Remote Sens.*, vol. 12, no. 12, pp. 2025, Jun. 2020.

[7] Z. Shao, Z. Zhou, X. Huang, and Y. Zhang, "MRENet: Simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images," *Remote Sens.*, vol. 13, no. 2, Jan. 2021, Art. no. 239.

[8] M. Zhou, H. Sui, S. Chen, J. Wang, and X. Chen, "BT-RoadNet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 168, pp. 288–306, 2020.

[9] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

[10] J. Shu, S. Wang, X. Jia, W. Zhang, R. Xie, and H. Huang, "Efficient lane-level map building via vehicle-based crowdsourcing," *IEEE Trans. Intell. Transport. Syst.*, pp. 1–14, 2020.

[11] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.

[12] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.

[13] X. Gao *et al.*, "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, 2018.

[14] Z. Chen, C. Wang, J. Li, N. Xie, Y. Han, and J. Du, "Reconstruction bias U-Net for road extraction from optical remote sensing images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 2284–2294, Jan. 2021.

[15] L. Zhou, C. Zhang, and M. Wu, "D-LinkNet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Salt Lake City, UT, USA, 2018, pp. 192–1924.

[16] Y. Liu, J. Yao, X. Lu, M. Xia, X. Wang, and Y. Liu, "RoadNet: Learning to comprehensively analyze road networks in complex urban scenes from high-resolution remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2043–2056, Apr. 2019.

[17] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan, "Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3322–3337, Jun. 2017.

[18] X. Lu *et al.*, "Multi-scale and multi-task deep learning framework for automatic road extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9362–9377, Nov. 2019.

[19] L. Chen, Q. Zhu, X. Xie, H. Hu, and H. Zeng, "Road extraction from vhr remote-sensing imagery via object segmentation constrained by gabor features," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 9, p. 362, Sep. 2018.

[20] L. Gao, W. Song, J. Dai, and Y. Chen, "Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network," *Remote Sens.*, vol. 11, no. 5, Mar. 2019.

[21] A. Batra, S. Singh, G. Pang, S. Basu, C. V. Jawahar, and M. Paluri, "Improved road connectivity by joint learning of orientation and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10385–10393. Accessed: Jul. 30, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2019/html/Batra_Improved_Road_Connectivity_by_Joint_Learning_of_Orientation_and_Segmentation_CVPR_2019_paper.html

[22] C. Tao, J. Qi, Y. Li, H. Wang, and H. Li, "Spatial information inference net: Road extraction using road-specific contextual information," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 155–166, Dec. 2019.

[23] X. Zhang, X. Han, C. Li, X. Tang, H. Zhou, and L. Jiao, "Aerial image road extraction based on an improved generative adversarial network," *Remote Sens.*, vol. 11, no. 8, Apr. 2019, Art. no. 930.

[24] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[25] Y. Zhang, Z. Xiong, Y. Zang, C. Wang, J. Li, and X. Li, "Topology-aware road network extraction via multi-supervised generative adversarial networks," *Remote Sens.*, vol. 11, no. 9, Apr. 2019, Art. no. 1017.

[26] J. Hu, S. Xiong, J. Zha, and C. Fu, "Lane detection and trajectory tracking control of autonomous vehicle based on model predictive control," *Int. J. Automot. Technol.*, vol. 21, no. 2, pp. 285–295, Apr. 2020.

[27] G. Cui, W. Bian, and X. Wang, "Hidden markov map matching based on trajectory segmentation with heading homogeneity," *GeoInformatica*, vol. 25, no. 1, pp. 179–206, Jan. 2021.

[28] L. Luo, X. Hou, W. Cai, and B. Guo, "Incremental route inference from low-sampling GPS data: An opportunistic approach to online map matching," *Inf. Sci.*, vol. 512, pp. 1407–1423, Feb. 2020.

[29] C. Chen, Y. Ding, X. Xie, S. Zhang, Z. Wang, and L. Feng, "TrajCompressor: An online map-matching-based trajectory compression framework leveraging vehicle heading direction and change," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2012–2028, May 2020.

[30] M. Deng *et al.*, "Generating urban road intersection models from low-frequency GPS trajectory data," *Int. J. Geogr. Inf. Sci.*, vol. 32, no. 12, pp. 2337–2361, Dec. 2018.

[31] B. Lin and J. Su, "One way distance: For shape based similarity search of moving object trajectories," *GeoInformatica*, vol. 12, no. 2, pp. 117–142, Jun. 2008.

[32] N. Wang, S. Gao, X. Peng, and M. Wang, "Research on fast and parallel clustering method for trajectory data," in *Proc. IEEE 24th Int. Conf. Parallel Distrib. Syst.*, Singapore, 2018, pp. 252–258.

[33] X. Niu, T. Chen, C. Q. Wu, J. Niu, and Y. Li, "Label-based trajectory clustering in complex road networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4098–4110, Oct. 2020.

[34] S. Wang, Z. Bao, J. S. Culpepper, T. Sellis, and X. Qin, "Fast large-scale trajectory clustering," *Proc. VLDB Endow.*, vol. 13, no. 1, pp. 29–42, Sep. 2019.

[35] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 25, no. 7, pp. 53–82, Jul. 2010.

[36] L. Tang, L. Niu, X. Yang, X. Zhang, Q. Li, and S. Xiao, "Urban intersection recognition and constrction based on big trace data," *Acta Geodaetica et Cartographica Sinica*, vol. 46, no. 6, pp. 770–779, 2017.

[37] M. Saeedimoghaddam and T. F. Stepinski, "Automatic extraction of road intersection points from USGS historical map series using deep convolutional neural networks," *Int J Geogr. Inf. Sci.*, vol. 34, no. 5, pp. 947–968, May 2020.

[38] V. Mnih, "Machine learning for aerial image labeling," Ph.D. dissertation, Graduate Dep. of Computer Science, Univ. Toronto (Canada), Ottawa, 2014.

[39] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.

[40] L. ZhiYong, T. Liu, J. A. Benediktsson, and N. Falco, "Land cover change detection techniques: Very-high-resolution optical images: A review," *IEEE Geosci. Remote Sens. Mag.*, pp. 2–21, 2021.

Lipeng Gao received the B.S. degree in geographical information system and the M.S. degree in photogrammetry and remote sensing from China University of Mining and Technology, Xuzhou, China, in 2011 and 2014, respectively, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019.

From 2015 to 2016, he was a Research Assistant with the Hong Kong Polytechnic University, Hong Kong. He has authored or co-authored more than 20 peer-reviewed journal/conference papers covering a wide range of topics in remote sensing, machine learning, and spatial big data analytics. His research interests include object extraction, intelligent information processing, big data, and software engineering.

Jingyu Wang was born in Taizhou, China in 1999. He is currently working toward the undergraduate degree in software engineering with the School of Software, Northwestern Polytechnical University, Xi'an, China.

His research interests include computer vision, image processing, and deep learning.

Qixin Wang was born in Xiao County, Xuzhou, China in 1990. He is currently working toward the Ph.D. degree in cartography and geographic information engineering with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include spatial data mining, big data analytics, and deep learning.

Wenzhong Shi received the doctoral degree in natural sciences from the University of Osnabrück, Vechta, Germany, in 1994.

He has authored/co-authored more than 100 research papers and 10 books. His current research interests include GIS and remote sensing, uncertainty and spatial data quality control, and image processing for high-resolution satellite images.

Dr. Shi is a Chair Professor in GIS and remote sensing with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. He was a recipient of the State Natural Science Award from the State Council of China, in 2007 and The Wang Zhizhuo Award from International Society for Photogrammetry and Remote Sensing, in 2012.

Jiangbin Zheng received the B.S., M.S., and Ph.D. degrees in computer science from Northwestern Polytechnical University, Xi'an, China, in 1993, 1996, and 2002, respectively.

From 2000 to 2001 and 2002, he was a Research Assistant with The Hong Kong Polytechnic University, Hong Kong. From 2004 to 2005, he was a Research Assistant with The University of Sydney, Sydney, Australia. Since 2009, he has been a professor and Ph.D. supervisor with the School of Computer Science, Northwestern Polytechnical University. He has authored or co-authored more than 100 peer-reviewed journal/conference papers covering a wide range of topics in image/video analytics, pattern recognition, machine learning and big data analytics. His research interests focus on intelligent information processing, visual computing, multimedia signal processing, big data, and software engineering.

Hongping Gan received the Ph.D. degree in communication and information engineering from the State Key Laboratory of ISN, Xidian University, Xi'an, China, in 2020.

From 2018 to 2019, he was with the School of Information Technology and Electrical Engineering, The University of Queensland, St. Lucia QLD, Australia as a Joint Training Doctoral Student under CSC scholarship. He is currently an Associate Professor with the School of Software, Northwestern Polytechnical University, China. His current research interests include deep learning, compressive sensing, and image processing.

Zhiyong LV received the M.S. and Ph.D. degrees in cartography and geographic information engineering from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2008 and 2014 respectively.

He was an Engineer of surveying and worked with The First Institute of Photogrammetry and Remote Sensing from 2008 to 2011. He is currently working with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, China. His research interests include multihyperspectral and high-resolution remotely sensed image processing, spatial feature extraction, neural networks, pattern recognition, deep learning, and remote sensing applications.

Honghai Qiao received the M.S. degree in traffic engineering information from the School of Automation, Northwestern Polytechnical University, Xi'an, China, in 2014. He is currently working toward the Ph.D. degree in control science and engineering with the School of Automation, Northwestern Polytechnical University, Xi'an, China.

His research interests include complex networks, social network, deep learning, and control theory.