Aerial Images Meet Crowdsourced Trajectories: A New Approach to Robust Road Extraction

Lingbo Liu, Zewei Yang, Guanbin Li, Kuo Wang, *Member, IEEE*, Tianshui Chen, and Liang Lin, *Senior Member, IEEE*

***Abstract*—Land remote-sensing analysis is a crucial research in earth science. In this work, we focus on a challenging task of land analysis, i.e., automatic extraction of traffic roads from remote-sensing data, which has widespread applications in urban development and expansion estimation. Nevertheless, conventional methods either only utilized the limited information of aerial images, or simply fused multimodal information (e.g., vehicle trajectories), thus cannot well recognize unconstrained roads. To facilitate this problem, we introduce a novel neural network framework termed cross-modal message propagation network (CMMPNet), which fully benefits the complementary different modal data (i.e., aerial images and crowdsourced trajectories). Specifically, CMMPNet is composed of two deep autoencoders for modality-specific representation learning and a tailor-designed dual enhancement module for cross-modal representation refinement. In particular, the complementary information of each modality is comprehensively extracted and dynamically propagated to enhance the representation of another modality. Extensive experiments on three real-world benchmarks demonstrate the effectiveness of our CMMPNet for robust road extraction benefiting from blending different modal data, either using image and trajectory data or image and light detection and ranging (LiDAR) data. From the experimental results, we observe that the proposed approach outperforms current state-of-the-art methods by large margins. Our source code is resealed on the project page http://lingboliu.com/multimodal_road_extraction.html.**

***Index Terms*—Aerial images, crowdsourced trajectories, land remote sensing, road network extraction.**

I. Introduction

EARTH science [1], [2] is a complex and huge subject that has been researched for decades or even centuries. As a subbranch of geoscience, geoinformatics [3] recently has received increasing interests with the rapid development of satellite and computer technologies. Accurately obtaining land surface information (e.g., trees, lakes, buildings, roads, and so on) from remote-sensing data can help us to better understand our earth. Among these objects, traffic roads are very difficult to recognize, since they are threadlike and unimpressive in aerial images. To promote land analysis, in this work we aim to recognize traffic roads automatically from remote-sensing data. Such a geoinformatics task not only facilitates a series of practical applications [4]–[6] for urban development, but also helps to estimate the urban expansion trend to analyze potential impacts of human activities on earth lands.

In literature, numerous algorithms have been proposed to extract traffic roads from aerial images. Most early works [7]–[9] extracted handcrafted features (e.g., texture and contour) and applied shallow models (e.g., support vector machine [10] and Markov random field [11]) to recognize road regions. Recently, deep convolutional networks have become the mainstream in this field and achieved remarkable progresses [12]–[14] due to their great capacities of representation learning. However, aerial image-based traffic road extraction remains a very challenging problem, especially in the face of the following circumstances. **First**, some roads are extremely occluded by trees, as shown in Fig. 1(a). Relying solely on visual information, these roads are hard to be detected from aerial images. **Second**, some infrastructures (e.g., train tracks, building tops, and river walls) have similar appearances of traffic roads, as shown in Fig. 1(b). Without extra information, it is hard to distinguish roads from these structures, which may result in false negatives and false positives. **Third**, in some bad meteorological conditions (e.g., thick fog/haze), it's very difficult to recognize traffic roads due to poor visibility, as shown in Fig. 1(c). Nevertheless, road maps have low tolerance for errors, since incorrect routes would seriously affect the transportation's operation efficiency. Therefore, some robust methods are desired to accurately extract traffic roads.

Fortunately, we observe that some data for nonvisual modalities, such as vehicle trajectories, can also help discover traffic roads. Intuitively, a region with a large number of trajectories is likely to be a road segment [15]–[17]. In recent years, vehicle ownership has grown dramatically and most vehicles have been equipped with GPS devices, which greatly increases the availability of large-scale trajectory datasets and boosts the feasibility of trajectory-based road extraction. Despite substantial progress [18], [19], this research direction still suffers from many challenges. **First**, crowdsourced trajectories have excessive noises (e.g., positioning drift) caused by the uneven

Manuscript received January 15, 2021; revised July 27, 2021 and December 30, 2021; accepted January 2, 2022. This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515020048; in part by the National Natural Science Foundation of China under Grant 61976250, Grant U1811463, and Grant 61836012; and in part by the Guangzhou Science and Technology Project under Grant 202102020633. *(Corresponding author: Liang Lin.)*

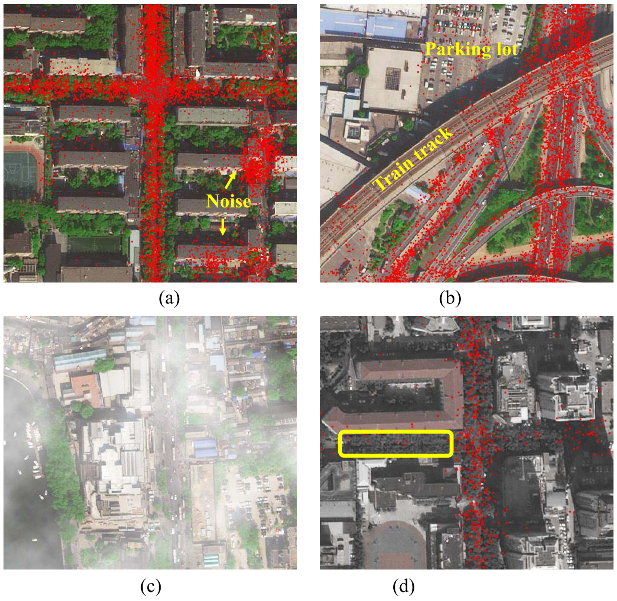
Lingbo Liu, Guanbin Li, Kuo Wang, and Liang Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China (e-mail: liulingb@mail2.sysu.edu.cn; liguanbin@mail.sysu.edu.cn; wangk229@mail2.sysu.edu.cn; linliang@ieee.org).

Zewei Yang is with the School of Mathematics, Sun Yat-sen University, Guangzhou 510000, China (e-mail: yangzw7@mail2.sysu.edu.cn).

Tianshui Chen is with the School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: tianshuichen@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3141821.

Digital Object Identifier 10.1109/TNNLS.2022.3141821



(a) (b) (c) (d)

Fig. 1. (a) Traffic roads are usually occluded by trees. Although crowdsourced trajectories can help discover roads, excessive noises are also introduced. (b) Train tracks and traffic roads have similar appearances, thus it is hard to distinguish them only using visual cues. When only using trajectories, some parking lots are easily mistaken for roads. (c) It's difficult to directly recognize roads from aerial images when the studied city has poor visibility in fog/haze weather. (d) Only using local information, we may fail to recognize some road regions that are heavily occluded and have very few trajectories, as shown in the yellow box. (a) Occlusion. (b) Similar appearance. (c) Poor visibility. (d) Limitation of local information.

quality of GPS devices, as shown in Fig. 1(a). Although various preprocessing techniques (e.g., clustering and K-nearest neighbors) were used Wang *et al.* [20], Shan *et al.* [21], and Karagiorgou *et al.* [22], the noise problem has not been well solved. **Second**, some nonroad areas, such as the parking lot in Fig. 1(b), also have lots of trajectories and they are easily mistaken for roads without auxiliary information. Most conventional works [23]–[26] have not explicitly distinguished these areas. **Third**, previous trajectory-based methods mainly extracted the topologies of road networks. Because of mass erratic trajectories, it is difficult to obtain the accurate width of roads, which can be easily computed in high-resolution aerial images.

In general, image-based methods and trajectory-based methods have individual strengths and weaknesses. It is very natural to incorporate aerial images and crowdsourced trajectories to extract traffic roads robustly. However, there are very limited works [27], [28] that simultaneously utilized the two modalities mentioned above. Moreover, these works directly fed the concatenation of aerial images and rendered maps of trajectories or their features into convolutional neural networks (CNNs), which is a suboptimal strategy for multimodal fusion. Recently, Wu *et al.* [29] designed a gated fusion module to fuse multimodal features, but not refine features mutually, thus the complementarities of images and trajectories have not been fully exploited. Furthermore, all above-mentioned methods performed road extraction only with local features/information, thus may fail to recognize some road regions that are heavily occluded and meanwhile have very few trajectories, such as the yellow box in Fig. 1(d). When considering all information of the whole image and trajectories, we can correctly infer that this region is a road segment. Therefore, both the local and global information should be explored for traffic road extraction.

To facilitate road extraction, we propose a novel framework termed cross-modal message propagation network (CMMPNet), which fully explores the complementarities between aerial images and vehicle crowdsourced trajectories. Specifically, our CMMPNet is composed of: 1) two deep autoencoders for modality-specific feature learning, in which one takes an aerial image as input and the other one uses the rendered trajectory heat-map and 2) a dual enhancement module (DEM) that refines the features of different modalities mutually with a message passing mechanism. In particular, our DEM propagates both the local detail information and global structural information dynamically with two progress propagators. **First**, a nonlocal message (NLM) propagator extracts the local and global messages embedded in the features of each modality, which are utilized to refine the features of another modality. Thereby, image features and trajectories features can be enhanced mutually. Moreover, the limitation of local information is also well eliminated. **Second**, a gated message propagator employs gate functions to dynamically determine the final propagated messages, so that the beneficial messages are transmitted and the interferential messages (e.g., visual cues of train tracks and the noises of trajectories) are abandoned. For further improving the robustness, our DEM is integrated into different layers of CMMPNet to enhance the image features and trajectory features hierarchically. Finally, the last outputs of two autoencoders are concatenated to accurately predict the high-resolution traffic road maps.

The proposed CMMPNet has three appealing properties. **First**, through refining modality-specific features mutually, our method can better explore the complementarities of aerial images and crowdsourced trajectories, compared with previous works that directly taken their concatenation as input or simply fused their features. **Second**, thanks to the tailor-designed DEM, our method is more robust to extract traffic roads. With the aid of visual information, some useless and noisy trajectories can be effectively eliminated, while occluded roads are easily discovered with the trajectory information and some delusive nonroad regions are also well distinguished. **Third**, it is worth noting that our method is very general for robust road extraction by utilizing multimodal information. Furthermore, CMMPNet can also be generalized to combine image and light detection and ranging (LiDAR) data for road extraction. Extensive comparisons on three real-world benchmarks two for image and trajectory data and the other for image and LiDAR data) demonstrate the advantage of our proposed method. In summary, this article makes the following contributions.

1) It proposes a novel CMMPNet for land remote-sensing analysis, which extracts traffic roads robustly by explicitly capturing the complementarities among different modal data.
2) It introduces a dual refinement module for multimodal representation learning, where the complementary information of each modality is dynamically propagated to effectively enhance other modal features based on the message passing mechanism.

3) It presents sufficient experiments and comparisons on three multimodal benchmarks for showing the superiority and generalization of our approach against existing state-of-the-art methods.

The rest of this article is organized as follows. First, we review some related works of earth science research and traffic road extraction in Section II. We then provide some preliminaries in Section III and introduce the proposed CMMPNet in Section IV. Extensive evaluations and generalization analysis are conducted in Section V and in Section VI. Finally, we conclude this article and discuss future works in Section VII.

II. Related Works

A. Earth Science Research

Earth science [1], [2] is a crucial subject that studies the physical, chemical, and biological characterizations of our earth for better understanding various physical phenomena and natural systems. Earth science is also a complex subject and it contains a lot of research branches [30]. For instance, meteorologists [31] study the atmosphere for dangerous storm warnings and hydrologists [32] examine hydrosphere for flood warnings. Seismologists [33] study earthquakes and forecast where they will strike, while geologists [34] study rocks and help to locate useful minerals. Among all the subbranches of geoscience, geoinformatics [3] recently has attracted widespread interests with the rapid development of satellite and computer technologies, since it can greatly facilitate other research branches, e.g., monitoring storm/flood from remote-sensing data and forecasting their evolutionary trend. In this work, we inherit the research content of geoinformatics and apply computer technologies to land remote-sensing analysis, e.g., extracting the traffic road network from aerial images and some complementary modalities. This problem has important applications in transportation navigation and public management. Moreover, we can also compare the road networks at different times and estimate the urban expansion tendency, thereby analyzing the potential impacts of human activities on earth lands.

B. Traffic Road Extraction

As a crucial foundation in intelligent transportation systems, automatic road extraction has been studied for decades [35]. On the basis of the modality of input data, previous approaches can be divided into four categories and we would investigate the related works of each category.

1) Aerial Image-Based Road Extraction: In industrial communities, a large number of high-quality aerial images can be accessed easily, with the rapid development of remote-sensing imaging technologies equipped in artificial satellites [36], [37]. Numerous methods were proposed to extract traffic roads from these aerial images. Early works [38]–[41] usually fed hand-crafted features (e.g., texture and contour) into shallow models (e.g., deformable model and Markov Random Field) to recognize road regions. However, most of them only worked in constraint scenarios. In recent years, due to the great capacity for representation learning, deep neural networks [42] have become the mainstream in this field. For instance, Cheng *et al.* [12] proposed a cascaded end-to-end CNN to cope with the road detection and centerline extraction simultaneously with two cascaded CNN. Zhang *et al.* [43] developed a semantic segmentation neural network, which combined the residual learning and U-Net to extract road areas. Zhou *et al.* [44] utilized dilation convolutions to enlarge the receptive field of Linknet [45] and then employed this enhanced model to extract road regions from high-resolution aerial images. Fu *et al.* [46] predicted the category of each pixel with a multiscale fully convolutional network and refined the output density map with a conditional random fields' postprocessing. Despite substantial progress, they may still fail in complex scenarios, especially in the face of extreme occlusions. As analyzed above, it is very difficult to perfectly extract traffic roads only with the visual information of aerial images. Therefore, more complementary information should be delved from other modalities for facilitating road extraction.

2) Trajectory-Based Road Extraction: Intuitively, a geographical region with mass vehicle trajectories is likely to be a road area. Based on this observation, some researchers have attempted to recognize traffic roads from crowdsourced trajectories. Since trajectory data has excessive noise, most previous works focused on how to eliminate the GPS noises and uncertainties. Conventional methods can be divided into three categories. The first category is clustering-based models [24], [47], [48]. In these works, the task of road extraction is formulated as a network alignment optimization problem where both the nodes and edges of road networks have to be inferred. Specifically, nodes or short edges are first identified from raw GPS points with spatial clustering algorithms and then connected to form the final road networks. The second category is trace-merging-based models [49], [50], which either merge each trajectory to an existing road segment or generate a new segment if no existing segment is matching. The third category is kernel density estimation (KDE)-based methods [19], [51], which first apply KDE [52] to convert trajectories into a density map for reducing the influences of noise, and then employs image processing techniques to extract road centerlines. Recently, deep neural networks have also been applied to this task. For instance, Ruan *et al.* [26] proposed a deep learning-based map generation framework, which extracts features from trajectories in both spatial view and transition view to infer road centerlines. Although various techniques are used, GPS noises still cannot be well eliminated and the extracted road networks are far from satisfactory, due to the limited information of crowdsourced trajectories.

3) LiDAR-Based Road Extraction: Compared with aerial images, LiDAR data have two specialties. First, LiDAR data contains depth or distance information. Second, different objects (e.g., buildings, trees, and roads) have different reflectivity to the laser. Because of these specialties, roads are mostly defined by flatness in the aerial viewpoint, which can help to distinguish the road proposals from buildings and trees. In literature, there also exist some algorithms [53]–[55] that identified traffic road from LiDAR data. For instance, Hu *et al.* [56] first filtered the nonground LiDAR points and then detected road centerlines from the remaining ground points. After obtaining the ground intensity images, Hu *et al.* [56] designed structure templates to search for roads and determined road widths and orientations with a subsequent voting scheme. Despite some progress, LiDAR-based road extraction remains a very challenging problem and existing

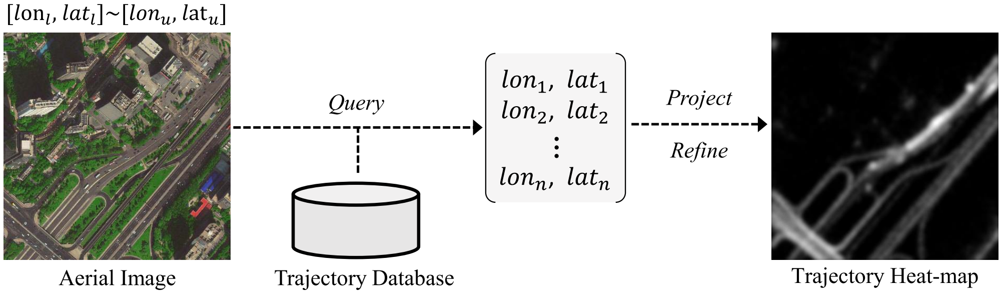


Fig. 2. Illustration of trajectory heat-map generation. Given an aerial image, we first query all trajectory samples in the corresponding geographical region and then generate a 2-D trajectory heat-map by counting the number of samples projected at every pixel. Finally, a logarithm-based normalized function and a Gaussian kernel filter are applied to refine the trajectory heat-map.

methods perform poorly in complex scenarios, suffering from the sparsity of LiDAR data and the noise points [57].

4) Multimodal Road Extraction: As analyzed above, each modality has individual benefits and drawbacks, so it's wise to aggregate their complementary information for extracting traffic roads effectively. In literature, numerous methods have been proposed to identify road areas using both aerial images and LiDAR data, because of the accessibility of these data. For instance, Hu *et al.* [58] first segmented the primitives of roads from both optical images and LiDAR data, and then detected road stripes with an iterative Hough transform algorithm to form the final road network by topology analysis. Parajuli *et al.* [59] developed a modular deep convolution network called TriSeg, in which two SegNet [60] were used to extract features, respectively, from aerial images and LiDAR data, and another SegNet fused modular features to estimate the final road maps. However, neither of aerial images and LiDAR data can provide sufficient information to discover the traffic roads heavily occluded by trees, thus some recent works incorporated aerial images and vehicle trajectories to identify road areas. For instance, Sun *et al.* [27] fed the concatenation of rendered trajectory heat-maps and aerial images into different backbone networks (e.g., UNet [61], Res-UNet [43], LinkNet [45] and D-LinkNet [44]) to estimate those traffic roads. In [29], trajectory maps and aerial images were first fed into different networks, respectively, for feature extraction, and then the modular features at different layers were fused to predict the final roads. Despite progress, such a concatenation or fusion manner cannot fully exploit the complementarities of different modalities, and more effective methods are desired for multimodal road extraction.

C. Message Passing Mechanism

In the field of machine learning, message passing [62], [63] refers to information interactions between different entities. A large number of works have shown that such a mechanism can effectively facilitate deep representation learning. For instance, Wang *et al.* [64] introduced an interview message passing module to enhance the view-specific features for action recognition, while Liu *et al.* [65] propagated information among multiscale features to model the scale variations of people. In graph convolution networks, the message passing mechanism is usually embedded to aggregate information from neighboring nodes [66]–[70]. Recently, this mechanism has also been adopted for cross-modal representation learning. For instance, Wang *et al.* [71] addressed the text-image retrieval problem by transferring multimodal features and computing their matching scores. Nevertheless, most of these previous methods propagated information in a local manner (e.g., at short range). Without capturing global information, these methods may fail to discover the occluded roads that meanwhile have very few trajectories, as shown in Fig. 1(d). Therefore, more effective approaches are desired to fully exploit the complementary information of aerial images and crowdsourced trajectories for traffic road extraction.

III. Preliminaries

In this section, we first introduce how to generate trajectory heat-maps from raw GPS data and then formally define the problem of image + trajectory-based road extraction.

A. Raw Trajectory Samples

With the rapid growth of vehicle ownership, we can easily collect a mass of vehicles' GPS trajectories to construct a large-scale trajectory database [72]–[74]. In this database, each trajectory sample can be represented as a tuple $\{\mathrm{vid}, \mathrm{lon}, \mathrm{lat}, t, \mathrm{sp}, \mathrm{si}\}$, where vid is the ID of a vehicle, lon and lat are the longitude and latitude at timestamp t. Term sp denotes the vehicle's speed. si is the trajectory sampling interval and different vehicles have different sampling settings. We notice that some early works [18] manually generated some virtual samples on the line segment between two consecutive real samples to augment the trajectory quantity. Nevertheless, this would cause a lot of noise in complex scenarios, since the real-world vehicles may have large sampling intervals (such as si is mainly set to 10, 60, 180, and 300 s in Beijing [27]) and it is difficult to accurately infer the virtual trajectories under these settings. Thanks to the crowdsourcing mechanism, adequate trajectories can be easily collected nowadays. Therefore, we only use the real trajectory samples in this work.

B. Trajectory Heat-Map Generation

For deep neural networks, matrix or tensor is one of the most common formats of input. Thus we need to transform the raw GPS data into 2-D trajectory heat-maps before feeding them into networks. The whole transformation process is shown in Fig. 2. Specifically, give an aerial image with a resolution

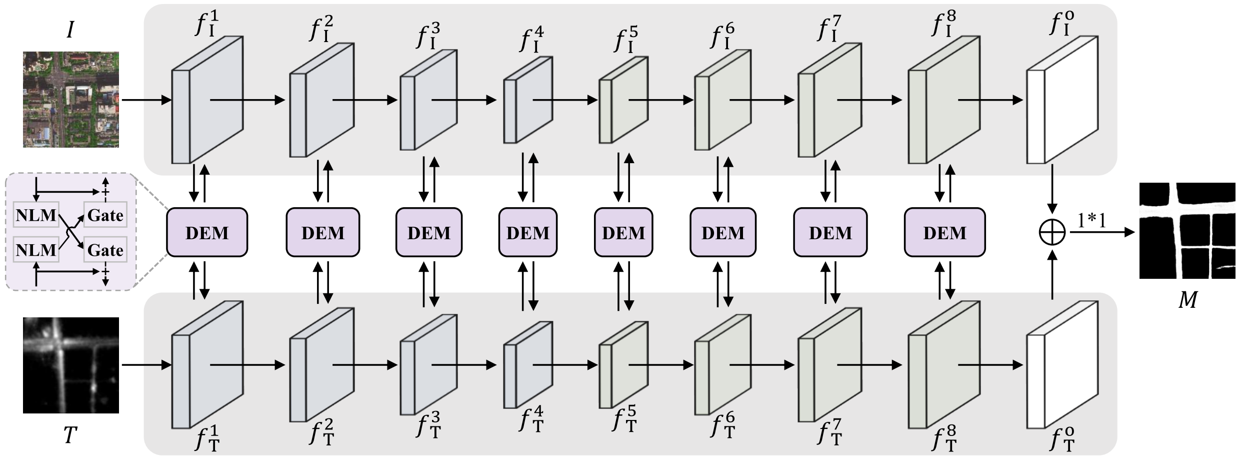


Fig. 3. Architecture of the proposed CMMPNet for multimodal road extraction. Specifically, our CMMPNet is composed of 1) two deep autoencoders that take an aerial image and a trajectory heat-map, respectively, to learn modality-specific features and 2) a DEM that dynamically propagates the NLMs (i.e., local one and global one) of every modality with gated functions to enhance the representation of another modality. The final features of the image and trajectory heat-map are concatenated to generate a traffic road map.

$H \times W$, we first search out all trajectory samples in its coordinate range $[\text{lon}_l, \text{lat}_l]$*$[\text{lon}_u, \text{lat}_u]$, where the subscripts l and u denote the lower and upper bounds, respectively. These samples are then projected into a $H \times W$ greyscale map by counting the number of samples projected at every pixel. In this map, the pixels of road areas usually have high values, while the pixel values in nonroad regions are very small, even zero. This would facilitate the discovery of traffic roads. However, we find that such a projected map has two minor defects. First, some infrequently-traveled roads are so pale that they are hard to be recognized. Second, this map is too coarse and sharp. For example, two adjacent pixels in road areas may have values of different scales, or even a road pixel does not match any projected samples. Inspired by KDE frequently used for trajectory processing [18], we normalize the projected map with a logarithm function and apply a 3×3 Gaussian kernel filter for smoothing. The involved Gaussian filter can also eliminate trajectory noises to a certain degree [75]. In this way, the final trajectory heat-map becomes smooth and traffic roads are more distinct from backgrounds.

C. Image + Trajectory-Based Road Extraction

Given a $H \times W$ aerial image I and the corresponding trajectory heat-map T, our goal is to automatically predict a $H \times W$ binary road map

$$M = \mathcal{F}(\{I, T\}, \theta) \tag{1}$$

where $\mathcal{F}(\cdot)$ is a mapping function with learnable parameters θ. Specifically, the pixels within road areas are supposed to have high-response value (i.e.. 1), while the response values of background pixels should be 0.

IV. Methodology

A. Framework Overview

As mentioned above, aerial images and vehicle crowdsourced trajectories are complementary for traffic road extraction. To recognize unconstrained roads effectively and robustly, we propose a CMMPNet, which mutually enhances the hierarchical features of different modalities for better capturing their complementary information. As shown in Fig. 3, our CMMPNet is composed of 1) two deep autoencoders for modality-specific feature learning and 2) a DEM for cross-modal feature refinement. In this section, we mainly introduce the architecture of CMMPNet, whose specific components are described in Section IV-C.

Specifically, given an aerial imagery I and a trajectory heat-map T with a resolution $H \times W$, we first explicitly learn modality-specific representations by feeding them into different autoencoders, each of which consists of four encoding blocks and four decoding blocks. As shown in Fig. 3, the first autoencoder takes I as input and extracts a group of image features

$$f_I = \{f_I^1, f_I^2, f_I^3, f_I^4, f_I^5, f_I^6, f_I^7, f_I^8\} \tag{2}$$

where the first four features are the outputs of encoding blocks and the remaining four features are the output of decoding blocks. With the same architecture, the second autoencoder extracts a group of trajectory features

$$f_T = \{f_T^1, f_T^2, f_T^3, f_T^4, f_T^5, f_T^6, f_T^7, f_T^8\} \tag{3}$$

from the input trajectory heat-map T.

Rather than directly fuse image and trajectory features with concatenation [28] or weighted addition [29], we fully capture the multimodal complementary information through enhancing their features mutually with a message passing mechanism. For each pair of multimodal feature $\{f_I^i, f_T^i\}$, we employ the proposed DEM to generate two enhanced features $\{\widehat{f_I^i}, \widehat{f_T^i}\}$ with their complementary information. This process can be formulated as

$$\widehat{f_I^i}, \widehat{f_T^i} = \text{DEM}\big(f_I^i, f_T^i\big), \quad i = 1, 2, \ldots, 8. \tag{4}$$

These enhanced features are then fed into the next block of individual autoencoder, respectively, for further representation learning. For convenience, the final outputs of image autoencoder and trajectory autoencoder are denoted as f_I^o and f_T^o, and they have the same resolution $H \times W$. Here, f_I^o and f_T^o

TABLE I

CONFIGURATION OF OUR AUTOENCODER. IN THE FIRST CONVOLUTIONAL LAYER, THE INPUT CHANNEL C_i IS SET TO 3 FOR AERIAL IMAGES AND 1 FOR TRAJECTORY HEAT-MAPS, AND THE STRIDE IS SET TO 2. IN EACH BLOCK, DR DENOTES THE DOWNSAMPLING RATIO OF RESOLUTION AND C_o IS THE CHANNEL NUMBER OF OUTPUT. MP DENOTES A 2 × 2 MAX-POOLING LAYER. *Res*, *Up* AND *Inter* REFER TO THE RESIDUAL UNIT, UPSAMPLING UNIT AND INTERIM UNIT DESCRIBED IN FIG. 4

Block	**Configuration**	**Output**		
		Sign	**DR**	C_o
-	Conv(7,C_i,64,s=2)		1/2	64
encoding-1	MP→3*Res(64, 64)	f^1	1/4	64
encoding-2	MP→Conv(64,128)→3*Res(128,128)	f^2	1/8	128
encoding-3	MP→Conv(128,256)→5*Res(256,256)	f^3	1/16	256
encoding-4	MP→Conv(256,512)→2*Res(512,512)	f^4	1/32	512
-	Inter(512,512)		1/32	512
decoding 1	Up(512,256) + f_3	f^5	1/16	256
decoding 2	Up(256,128) + f_2	f^6	1/8	128
decoding 3	Up(128, 64) + f_1	f^7	1/4	64
decoding 4	Up(64, 64)	f^8	1/2	64
-	TConv(4,64,32,2)→Conv(3,32,32)	f^o	1	32

are jointly utilized to predict a probability map $M \in R^{H\times W}$ for traffic roads with the following formulation:

$$M = \mathrm{Conv}\big(f_I^o \oplus f_T^o, \mathbb{W}_{1*1}\big) \tag{5}$$

where $\oplus$ denotes feature concatenation and $\mathbb{W}_{1*1}$ refers to the parameters of a 1*1 convolutional layer. For each position (x, y), it can be regarded as a road region only when $M(x, y)$ is greater than a given threshold.

It's worth noting that our method is universal for multimodal road extraction. Except for image + trajectory data, the proposed CMMPNet can also be directly employed to recognize traffic roads with image + LiDAR data. The university of our method would be verified in Sections V and VI.

B. Modality-Specific Feature Learning

In the previous work [27], aerial images and trajectory heat-maps were directly concatenated to feed into the same network, which caused that their features were over mixed and their complementarities were missed to some extent. To address this problem, we feed the given aerial image and the corresponding trajectory heat-map into different networks to learn modality-specific features. Optimized with individual parameters, these features well preserve the specific information of each modality, thus can be further utilized for mutual refinement.

To maintain the high resolution of final outputs, two autoencoders are adopted intentionally to extract modality-specific features. Notice that various autoencoders (e.g., ResUNet [43], LinkNet [45], and D-LinkNet [44]) are suitable to serve as the backbone network of our framework. Since these networks have similar architectures, we take D-LinkNet-based autoencoder as an example to demonstrate the details of modality-specific feature learning. As shown in Table I, both the image autoencoder and trajectory autoencoder are mainly composed of four encoding blocks and four decoding blocks. Specifically, we first use a convolutional layer to extract initial features and then feed them into the following four encoding blocks, each of which consists of a 2 × 2 max-pooling layer and multiple residual units. As shown in Fig. 4(a), each residual unit contains two 3 × 3 convolutional layers and a skip layer. After the encoding stage, an interim unit is adopted to capture more spatial context by expanding the receptive field with four dilated convolutional layers. At the decoding stage, four decoding blocks are utilized to enlarge the resolutions of features progressively. Specifically, each decoding block is developed as an upsampling unit, which consists of two convolutional layers for channel adjustment and a transposed convolutional layer for feature upsampling, as shown in Fig. 4(c). To simultaneously exploit the lower-level information and high-level information, we incorporate the features of encoding blocks and decoding blocks with element-wise addition. Finally, we fully restore the resolution of feature to $H \times W$ with a transposed convolution and apply a 3×3 convolutional layer to generate the final modality-specific feature $f^o \in R^{H\times W\times 32}$. Note that our image autoencoder and trajectory autoencoder have individual parameters, thus they can effectively capture and preserve the specific information of each modality.

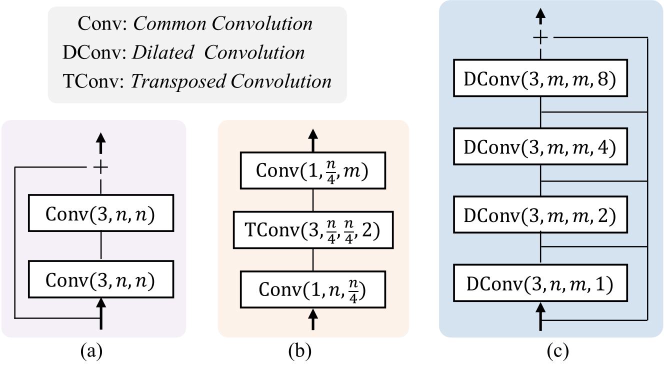


Fig. 4. Architecture of residual unit, upsampling unit, and interim unit. Conv(k, n, m) denotes a $k \times k$ standard convolution, whose input channel is n and output channel is m. DConv(k, n, m, r) refers to a dilated convolution with a dilated ratio r and TConv(k, n, m, s) is a transposed convolution with a stride s. (a) Residual unit. (b) Upsampling unit. (c) Interim unit.

C. Cross-Modal Feature Refinement

After modality-specific feature learning, we refine these features mutually with a DEM based on the message passing mechanism. In this module, a NLM propagator and a gated message propagator are integrated to dynamically transmit the local and global message of each unimodal feature to complement the feature of another modality. Absorbing the complementary information of other modalities, each unimodal feature becomes more reasonable and robust. In this section, we take the refinement of features f_I^i and $f_T^i \in R^{h\times w\times c}$ as an example to demonstrate the working mechanism of the tailor-designed DEM. Note that h, w, and c are the height, width, and channel number of these features.

1) NLM Propagator: Unlike previous works [65], [77] that only used local cues, our method explores both the local and global information for feature enhancement. Here we mainly introduce how to utilize the information of trajectory feature f_T^i to enhance the image feature f_I^i. The refinement of f_T^i is performed with the same process.

As shown in Fig. 5, we first extract a local information map $L_T^i \in R^{h\times w\times c}$ by feeding f_T^i into a 3×3 convolutional

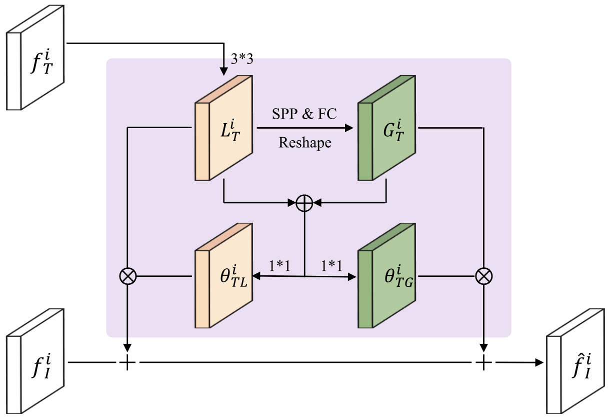


Fig. 5. Architecture of DEM. This figure mainly illustrates how to enhance the image feature f_I^i with the information extracted from the trajectory feature f_T^i. The cross-modal information from f_T^i to f_I^i is obtained by dynamically fusing the local information L_T^i and global information G_T^i with the learnable fused weights θ_{TL}^i and θ_{TG}^i. This architecture can also be employed to enhance f_T^i. SPP and FC are the abbreviations of spatial pyramid pooling [76] and fully connected layer, respectively. $+$ and $\otimes$ denote the element-wise addition and multiplication, and $\oplus$ refers to feature concatenation.

layer. Then we aggregate the local information at different locations to generate a global information map. Rather than use the compute-intensive nonlocal module proposed in [78], we employ a lightweight N-level spatial pyramid pooling (SPP [76]) and a fully connected (FC) layer for global information generation. Specifically, at the ith level ($i = 1, 2 \ldots, N$), L_T^i is divided into $2^{i-1} \times 2^{i-1}$ regions, each of which has a dimension of $(h/2^{i-1}) \times (w/2^{i-1}) \times c$ and is fed into a $(h/2^{i-1}) \times (w/2^{i-1})$ max-pooling layer to obtain a $1 \times 1 \times c$ information vector. Further, the information vectors at all levels are concatenated and fed into the FC layer with c output neurons to generate a global information vector. This global vector is copied $h \times w$ times and reshaped to form the global information map $G_T^i \in R^{h \times w \times c}$. After obtaining the local map L_T^i and global map G_T^i, we can easily propagate them to refine f_I^i with the following formulation:

$$\widehat{f_I^i} = f_I^i + L_T^i + G_T^i \tag{6}$$

where $\widehat{f_I^i}$ is the enhanced image feature and the operator "$+$" denotes a element-wise addition. In the same way, we can compute the enhanced trajectory feature $\widehat{f_T^i}$ as follows:

$$\widehat{f_T^i} = f_T^i + L_I^i + G_I^i \tag{7}$$

where L_I^i and G_I^i are the extracted local information map and global information map of image feature f_I^i.

2) Gated Message Propagator: In the previous propagator, the local and global information is transmitted statically, which is not optimal for cross-modal refinement and even has disturbing effects at some locations. To alleviate this issue, a gated message propagator is introduced to adaptively determine and propagate the complementary information. With multiple learnable gate functions, the beneficial information is transmitted and the disturbing information (e.g., visual cues of train tracks and the noises of trajectories) is suppressed.

Specifically, we first introduce the computation of the gated weights of different information. As shown in Fig. 5, the trajectory information L_T^i and G_T^i is concatenated and fed into two 1×1 convolutional layers

$$\begin{aligned} \theta_{\mathrm{TL}}^i &= \mathrm{Sigm}\big(\mathrm{Conv}\big(L_T^i \oplus G_T^i, \mathbb{W}_{\mathrm{TL}}^i\big)\big) \\ \theta_{\mathrm{TG}}^i &= \mathrm{Sigm}\big(\mathrm{Conv}\big(L_T^i \oplus G_T^i, \mathbb{W}_{\mathrm{TG}}^i\big)\big) \end{aligned} \tag{8}$$

where $\theta_{\mathrm{TL}}^i, \theta_{\mathrm{TG}}^i \in R^{h \times w \times c}$ are the gated weights of L_T^i and G_T^i, respectively. $\mathbb{W}_{\mathrm{TL}}^i$ and $\mathbb{W}_{\mathrm{TG}}^i$ are the parameters of convolutional layers, and Sigm() is an element-wise sigmoid function. In this same way, we can also compute the gated weights $\theta_{\mathrm{IL}}^i, \theta_{IG}^i \in R^{h \times w \times c}$ for the information L_I^i and G_I^i. Finally, we reweight each information with individual gated weight and then preform the dynamic message propagation. Therefore, (6) and (7) have become

$$\begin{aligned} \hat{f}_I^i &= f_I^i + \theta_{\mathrm{TL}}^i \otimes L_T^i + \theta_{\mathrm{TG}}^i \otimes G_T^i \\ \hat{f}_T^i &= f_T^i + \theta_{\mathrm{IL}}^i \otimes L_I^i + \theta_{IG}^i \otimes G_I^i \end{aligned} \tag{9}$$

where $\otimes$ denotes an element-wise multiplication.

D. Implementation Details

In this work, we implement the proposed CMMPNet on the representative deep learning platform PyTorch [79]. First, we perform data augmentation to alleviate the overfitting issue. Specifically, all training samples including the satellite images, trajectory heat-maps and ground-truth (GT) maps are 1) flipped horizontally or vertically; 2) rotated by 90°, 180°, 270°; and 3) randomly cropped with a size range of [0.7, 0.9] and resized to the original resolution. After augmentation, the number of training samples is enlarged by seven times. We then determine the hyperparameters of our framework. The filter weights of all convolutional layers and FC layers are uniformly initialized by Xavier [80]. The batch size is set to 4 and the learning rate is set to 0.0002. Finally, we apply the Adam [81] optimizer to train our CMMPNet for 30 epochs by minimizing the binary cross-entropy loss between the generated road maps and the corresponding GT maps.

V. Experiments

In this section, we first introduce the experiment settings of image + trajectory-based road extraction. We then compare the proposed CMMPNet with existing state-of-the-art approaches and finally conduct extensive ablation studies to verify the effectiveness of each component in our network.

A. Settings

1) Datasets: In this work, our experiments are mainly conducted on the BJRoad dataset [27], which is captured in Beijing, China. Specifically, this benchmark consists of 350 high-resolution aerial images that cover a large geographic area of about 100 km^2 and around 50 million trajectory records of 28 000 vehicles. The resolution of aerial images is 1024×1024 and each pixel denotes a 0.5 m × 0.5 m region in the real world. For each aerial image, a 1024×1024 trajectory heat-map is generated with the preprocessing described in Section III, and the corresponding GT map is manually created by masking out the pixel of traffic roads. Finally, this dataset is officially divided into three partitions: 70% samples are adopted for training, 10% for validation, and the rest 20% for testing.

Following the previous work [29], we also perform experiments on the Porto dataset, which covers a geographic area of about 209 km^2 in Porto, Portugal. This dataset contains a mass of crowdsourced trajectories generated by 442 taxis from 2013 to 2014. On this dataset, we adopt a fivefold cross-validation setting, since the details of training/testing sets are not provided in [29]. Specifically, the aerial image of the whole area is first cut into 6048 nonoverlapping subimages with a resolution of 512 × 512. These subimages are then randomly divided into five equal parts. For the ith validation, the ith part is used for testing, and the remaining parts are used for training. Finally, the mean and variance of five validations are reported.

2) Evaluation Details: Given a probability map M, we need to determine an estimated road map $M_e \in R^{H\times W}$ before evaluation. Same to [27], a pixel (x, y) is predicted as a road region in our work, if the response value of $M(x, y)$ is greater than 0.5. Following previous works [44], [82], we adopt Intersection over Union[1] (IoU) to evaluate the performance for road extraction. Specifically, the IoU score between an estimated map M_e and its corresponding GT map M_g is computed by

$$\mathrm{IoU}(M_e, M_g) = \frac{|M_e \cap M_g|}{|M_e \cup M_g|} \tag{10}$$

where $|M_e \cap M_g|$ denotes the pixel number in the intersection set of M_e and M_g, and $|M_e \cup M_g|$ is the pixel number in their union set. There are two manners for computing the IoU of all testing samples. The first manner is to compute the IoU of each sample and then average the IoU of all samples. Such a metric is termed as average IoU (A_IoU). The second manner is to stitch the estimated maps of all samples into a global map and then compute an IoU score. This metric is termed as global IoU (G_IoU). Since different IoU metrics were used in previous works, we would report the results of both A_IoU and G_IoU in the following sections.

B. Comparison With State-of-the-Art Methods

In this section, we compare our CMMPNet with seven deep learning-based approaches, including DeepLab (v3+) [83], UNet [61], Res-UNet [43], LinkNet [45], D-LinkNet [44], Sun *et al.* [27], and DeepDualMapper [29]. Specifically, these compared methods are reimplemented for multimodal road extraction. In particular, DeepDualMapper feeds aerial images and trajectory heat-maps into different backbone networks[2] and then fuses their features with a gated fusion module, while other methods directly take the concatenation of aerial images and trajectory heat-maps as input. Moreover, all the compared methods except DeepLab (v3+) and DeepDualMapper are equipped with 1-D transpose convolution to better model traffic roads [27]. Notice that the first six compared methods were implemented by Sun *et al.* [27], and we utilize the official code of [27] to implement DeepDualMapper and our method with the same data partition. As mentioned above, our CMMPNet can be developed with various autoencoders. We hence evaluate multiple implementations of CMMPNet based on different autoencoders, such as Res-UNet, LinkNet, and D-LinkNet.

TABLE II

PERFORMANCE OF DIFFERENT METHODS ON THE TESTING SET OF BJROAD DATASET. OUR CMMPNET OUTPERFORMS ALL EXISTING APPROACHES WITH LARGE MARGINS

Method	A_IoU (%)	G_IoU (%)
DeepLab (v3+) [83]	50.81	-
UNet [61]	54.88	-
Res-UNet [43]	54.24	-
LinkNet [45]	57.89	-
D-LinkNet [44]	57.96	-
Sun et al. [27]	59.18	-
DeepDualMapper [29]	60.91	61.54
Res-UNet+CMMPNet	**62.58**	**63.03**
LinkNet+CMMPNet	**63.09**	**63.46**
D-LinkNet+CMMPNet	**62.85**	**63.39**

TABLE III

PERFORMANCE OF DIFFERENT METHODS ON THE PORTO DATASET. FIVEFOLD CROSS-VALIDATION IS CONDUCTED ON THIS DATASET. THE MEAN AND VARIANCE OF FIVE VALIDATIONS ARE REPORTED IN THIS TABLE

Method	A_IoU (%)	G_IoU (%)
D-LinkNet [44]	72.82±0.47	72.92±0.45
Sun et al. [27]	72.94±0.71	73.04±0.63
DeepDualMapper [29]	73.67±0.51	73.91±0.51
D-LinkNet+CMMPNet	74.56±0.46	74.66±0.41

The performance of all methods on the BJRoad dataset is summarized in Table II. We can observe that DeepLab (v3+) obtains the worst A_IoU 50.81% probably because of parameter overfitting. Sun *et al.* [27] utilized various techniques (e.g., different sampling intervals and GPS augmentation) to obtain an improved A_IoU 59.18%. However, just directly feeding the concatenation of aerial images and trajectory heat-maps into networks, these methods have limited capabilities to capture the multimodal information, thus none of them can acquire an A_IoU above 60%. By fusing image and trajectory features with a gated fusion module, DeepDualMapper obtains a competitive A_IoU 60.91% and G_IoU 61.54%. Despite the progress, DeepDualMapper only use a fusion strategy rather than a mutual refinement strategy, thereby cannot address this task well. In contrast, when learning modality-specific features explicitly and enhancing cross-modal features mutually, our method can fully exploit the complementary information of aerial images and crowdsourced trajectories. For this reason, the proposed CMMPNet outperforms all previous methods with large margins. For instance, Res-UNet + CMMPNet achieves a competitive A_IoU 62.58% and obtains a relative improvement of 15.37%, compared with the original Res-UNet. By improving the A_IoU from 57.96% to 62.85%, our D-LinkNet + CMMPNet also obtains a substantial improvement of 8.4%, compared with the baseline D-LinkNet. Finally, with an impressive A_IoU 63.09% and a G_IoU 63.46%, our LinkNet + CMMPNet becomes the best-performing model. We notice that the performance of D-LinkNet + CMMPNet is slightly lower than that of LinkNet + CMMPNet. This is probably because D-LinkNet + CMMPNet contains two extra

[1] https://en.wikipedia.org/wiki/Jaccard_index

[2] In DeepDualMapper, the original backbone network is UNet. However, our reimplemented DeepDualMapper based on UNet performs poorly. Thus, in this work, we adopt D-LinkNet as the backbone to reimplement DeepDualMapper and this model can obtain competitive performance on different datasets.

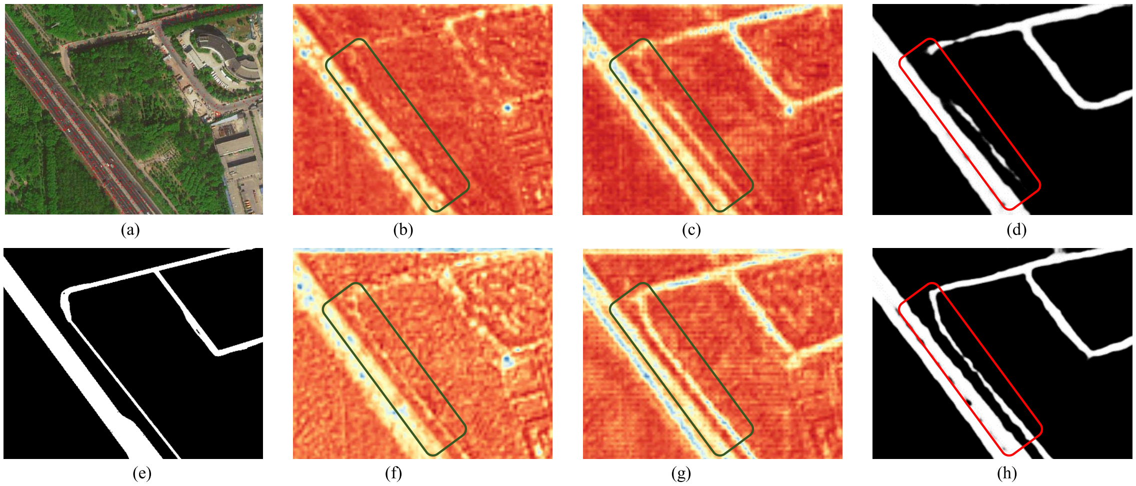


Fig. 6. Visualization of the feature maps and traffic road maps generated with/without global information on the testing set of BJRoad dataset. (a) Is the input aerial image with trajectory points and (e) is the GT road map. (b) and (c) Are the average maps of image feature and trajectory feature after the first/seventh DEM without the global message, while (d) is the generated road network without the global message. (f)–(h) Are the generated feature maps and the road network using both the local message and global message. We can observe that our method can generate more discriminative features and recognize the occluded/unimpressive traffic roads effectively when performing road extraction with global information. (a) Image and trajectories. (b) Average feature after the first DEM W/O global message. (c) Average feature after the seventh DEM W/O global message. (d) Generated road network W/O global message. (e) GT road map. (f) Average feature after the first DEM W/-global message. (g) Average feature after the seventh DEM W/-global message. (h) Generated road network W/-global message.

interim units and suffers from certain overfitting, although data augmentation has been performed.

Moreover, we compare the performance of our CMMPNet with three competitive models including D-LinkNet [44], Sun *et al.* [27], and DeepDualMapper [29] on the Porto dataset. As shown in Table III, all methods obtain much better results on this dataset, compared with their performance on the BJRoad dataset. The main reason is that the aerial images of Porto are clearer and the noises of trajectories are smaller [29]. Despite the existing benchmarks are high, our CMMPNet still can boost the IoU with substantial margins, ranking first in performance on the Porto dataset. In summary, these comparisons greatly demonstrate the effectiveness of the proposed CMMPNet for image + trajectory-based road extraction.

C. Component Analysis

After external comparison, we then perform extensive internal experiments to analyze the effectiveness of each module in the proposed CMMPNet. In this section, D-LinkNet is adopted as the backbone network and our implementation details have been described in Section IV-D.

1) Effect of Global Message: In previous works [84], [85], local information is widely adopted, but global information is neglected. In this section, we implement several variants of CMMPNet to verify the effectiveness of global information. As shown in Table IV, when propagating the global information extracted by SPP and FC layer, "Local + Global" model obtains an A_IoU 61.98% and a G_IoU 62.43%, and is better than "Local" model. With an A_IoU 62.85% and a G_IoU 63.39%, "Local + Global + Gate" model also outperforms "Local + Gate" model, whose A_IoU is 62.32% and G_IoU is 62.78%. Except for quantitative results, we also visualize some feature maps and traffic road maps generated by "Local + Gate" model and "Local + Global + Gate" model in Fig. 6. Note that those visualized feature maps are the channel-wise average of image features and trajectory features after DEM. We can observe that incorporating global information can generate more discriminative features and better recognize traffic roads, especially when the roads are occluded/unimpressive and the vehicle trajectories are rare in local regions. These quantitative and qualitative experiments show that global information is greatly effective for traffic road extraction.

TABLE IV
INFLUENCE OF NLM PROPAGATOR AND GATED MESSAGE PROPAGATOR ON THE TESTING SET OF BJROAD DATASET

Local	Global	Gate	A_IoU (%)	G_IoU (%)
✓			61.62	62.09
✓		✓	62.32	62.78
✓	✓		61.98	62.43
✓	✓	✓	62.85	63.39

2) Effect of Gated Message Propagator: In this propagator, multiple gate functions are employed to dynamically propagate the complementary information. In this section, we also implement several variants to verify the effectiveness of this mechanism. As shown in Table IV, after applying gate functions on "Local" model, A_IoU increases from 61.62% to 62.32% and G_IoU increases from 62.09% to 62.78%. Further, we can obtain a more substantial improvement (around 1% on both A_IoU and G_IoU), when performing gate functions on "Local + Global" model. These comparisons show that this proposed propagator can facilitate robust road extraction using multimodal information.

3) Configuration of SPP: In NLM Propagator, we employ a N-level SPP and an FC layer to extract global information. In this section, we explore the effect of the level number for road extraction using multimodal information. As shown in

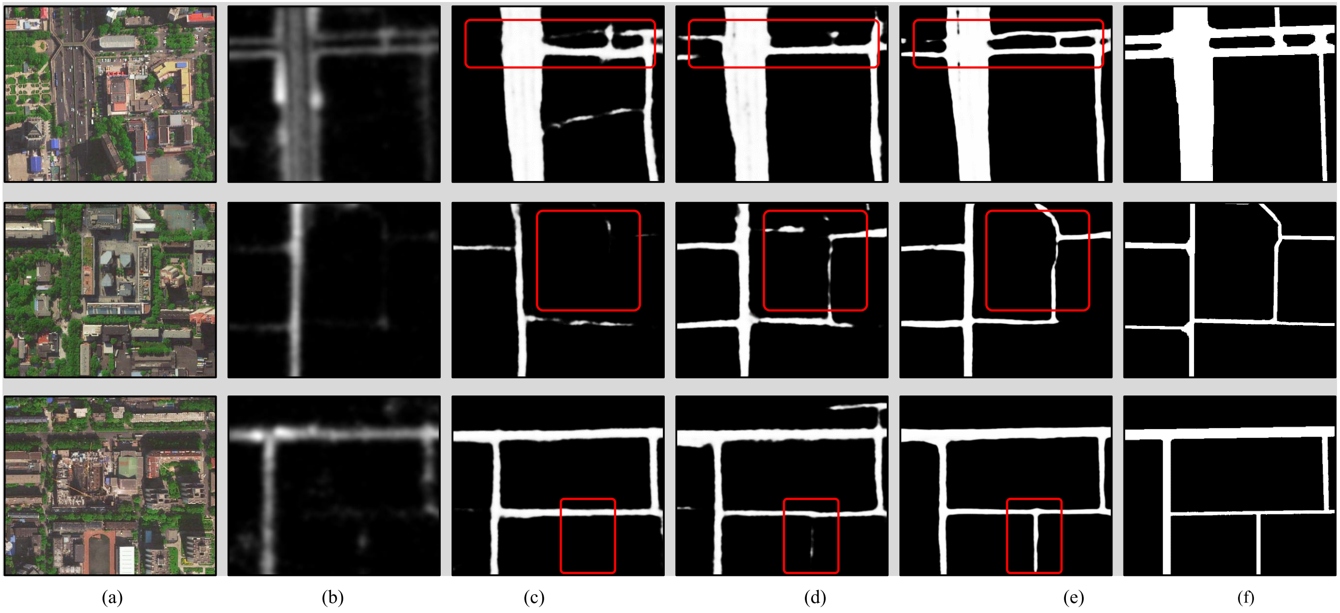


Fig. 7. Visualization of the traffic road networks generated by different methods on the testing set of BJRoad dataset. (a) and (b) Are the input aerial images and trajectories heat-maps. (c) Are the results that only aerial images are taken as input, while (d) are the results that the concatenation of images and heat-maps are taken as input. As shown in (e), results of our CMMPNet are more accurate and are very similar to the GT road networks. (a) Aerial image. (b) Trajectories heat-map. (c) Image-based result. (d) Early fusion result. (e) Our result. (f) GT road map.

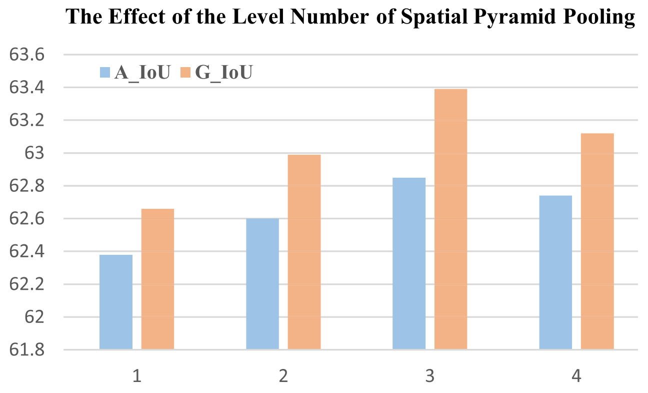


Fig. 8. Influence of the level number N of SPP layer in DEM on the testing set of BJRoad dataset. Our method achieves the best performance when N is set to 3.

Fig. 8, when applying a global max-pooling ($N = 1$), our CMMPNet obtains an A_IoU 62.38% and is slightly better than the "Local + Gate" model in Table IV, since global pooling can only provide some coarse and limited information. As the level number increases, the performance also gradually increases, and our method achieves the best A_IoU 62.85% and G_IoU 63.39% when N is equal to 3. When n increases to 4, the performance slightly drops, probably because of overfitting, i.e., the amount of parameters of the FC layer in DEM increases sharply as the level number increase. Therefore, the level number N of SPP is uniformly set to 3 for road extraction.

D. More Discussion

1) Unimodal Data Versus Multimodal Data: We first explore whether multimodal data is reliably useful for traffic road extraction. As shown in Table V, when only feeding trajectory heat-maps into a D-LinkNet, we obtain a poor performance (A_IoU 52.38%, G_IoU 52.90%) on the BJRoad dataset. When only utilizing aerial images, we obtain an A_IoU 59.79% and a G_IoU 60.24%, which indicates that image data is more crucial than trajectory data. In contrast, when using the aerial images and trajectory heat-maps simultaneously, our CMMPNet and the early/late fusion models described in the next paragraph outperform the unimodal models consistently with an improvement of at least 1% on IoU. This comparison demonstrates that multimodal data is more effective for traffic road extraction, because aerial images and vehicle crowdsourced trajectories have rich complementarities.

2) Which Multimodal Learning Manner Is Better?: We then explore the effects of different multimodal learning manners. Except for the proposed CMMPNet, we also implement another two commonly-used manners, i.e., early fusion model and late fusion model. Specifically, the former feeds the concatenation of aerial images and trajectory heat-maps into a D-LinkNet. In the latter, aerial images and trajectory heat-maps are, respectively, fed into individual D-LinkNet, and their final features are concatenated to estimate the road maps. As shown in Table V, the early fusion model obtains an A_IoU 61.11% and a G_IoU 61.53%, slightly outperforming the late fusion model (A_IoU 60.78%, G_IoU 61.24%). This is because the multimodal information is utilized at different layers in the former, but just utilized once in the latter. Compared with these two models, our CMMPNet is more reasonable to learn modality-specific features and propagate cross-modal information hierarchically. For this reason, our method achieves an impressive A_IoU 62.85% and G_IoU 63.39%, and outperforms early/late fusion models with a large margin, as shown in Fig. 7. This comparison shows the effectiveness of our CMMPNet for multimodal representation learning.

TABLE V

PERFORMANCE OF DIFFERENT INPUTS AND DIFFERENT REPRESENTATION LEARNING MANNERS ON THE TESTING SET OF BJROAD DATASET

Input	Learning Manner	A_IoU (%)	G_IoU (%)
Trajectory	-	52.38	52.90
Image	-	59.79	60.24
Image+Trajectory	Late Fusion	60.78	61.24
	Early Fusion	61.11	61.53
	CMMPNet	62.85	63.39

TABLE VI

PERFORMANCE OF TRAFFIC ROAD EXTRACTION BASED ON FOGGY IMAGES AND VEHICLE TRAJECTORIES ON THE TESTING SET OF THE FOGGY BJROAD DATASET

Input	Manner Way	A_IoU (%)	G_IoU (%)
Trajectory	-	52.38	52.90
Fog_Img	-	54.54	55.27
Fog_Img + Trajectory	Early Fusion	57.98	58.49
	CMMPNet	60.45	61.06

3) Significance of Crowdsourced Trajectories: Although the IoU of aerial images is much better than that of trajectories, we argue that vehicle trajectories are crucial for the robustness of road extraction, especially when some cities (e.g., Chongqing and Chengdu, China) are greatly covered by fog and mist in aerial images. So here we explore to extract traffic roads from foggy images and crowdsourced trajectories. Since there are no foggy images in the BJRoad dataset, we need to generate some aerial images with heavy fog in advance. Specifically, for each cloudless image in BJRoad, we employ a fog effect renderer of Photoshop to generate a foggy image. After augmenting the training samples as described in Section IV-D, we reimplement the proposed CMMPNet and three other compared methods, including 1) two unimodal models which feed foggy images or trajectory heat-maps into D-LinkNet and 2) an early fusion D-LinkNet model which takes the concatenation of foggy images and trajectory heat-maps as input.

The results of all methods are summarized in Table VI. We can observe that the unimodal D-LinkNet only obtains an A_IoU 54.54% and a G_IoU 55.27% when only using foggy images. Compared with the corresponding model using cloudless images, this model has a dramatic drop in performance, since traffic roads may be invisible in foggy images. When utilizing foggy images and trajectories simultaneously, the early fusion model obtains an A_IoU 57.98% and a G_IoU 58.49%. Based on the same D-LinkNet, our CMMPNet achieves a competitive A_IoU 60.45% and G_IoU 61.06%, having a performance improvement of at least 3% compared with other models. Moreover, the visualizations in Fig. 9 shows that our CMMPNet can still generate high-quality road maps in foggy weather conditions. This is attributed to the fact that the vehicle trajectories can provide rich information to remedy the limitation of aerial images, and our method can fully capture their complementary information. In summary, crowdsourced trajectories are very crucial and beneficial for robust road extraction.

VI. APPLY TO IMAGE + LIDAR-BASED EXTRACTION

As mentioned above, our method is general for road extraction by exploiting multimodal information. In this section, we employ the proposed CMMPNet to recognize traffic roads from aerial images and LiDAR data. As shown in Fig. 11(a) and (b), LiDAR data can help to discover some occluded or inconspicuous roads in aerial images. Here we conduct extensive experiments on the Tallahassee-Leon County GIS (TLCGIS) [59] dataset, which consists of 5860 pairs of aerial images and LiDAR images rendered from raw LiDAR point cloud data. The resolution of these images is 500 × 500 and the geographical length of each pixel is 0.5 feet. This dataset is officially divided into training, test, and validation sets with each having 2640, 2400, and 240 samples, respectively. On this dataset, we also take D-LinkNet as the backbone to develop our CMMPNet and optimize this model with the process described in Section IV-D.

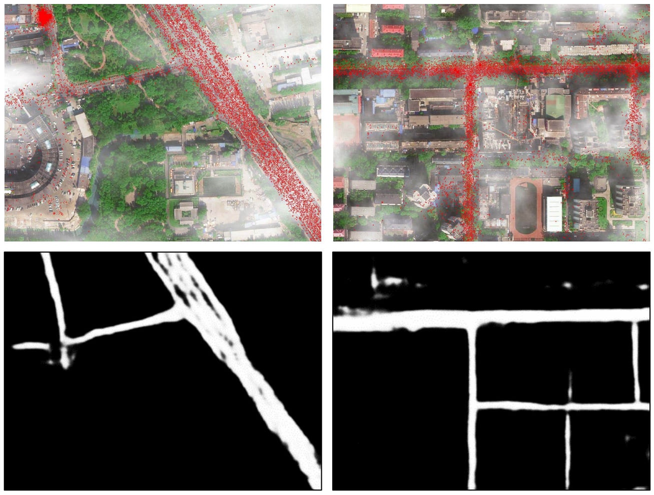


Fig. 9. First row shows some foggy images and mass vehicle trajectories on the testing set of the foggy BJRoad dataset. Although traffic roads are occluded extremely in these images, our CMMPNet can still generate high-quality road network maps by fully exploiting the complementary information of vehicle crowdsourced trajectories, as shown in the second row.

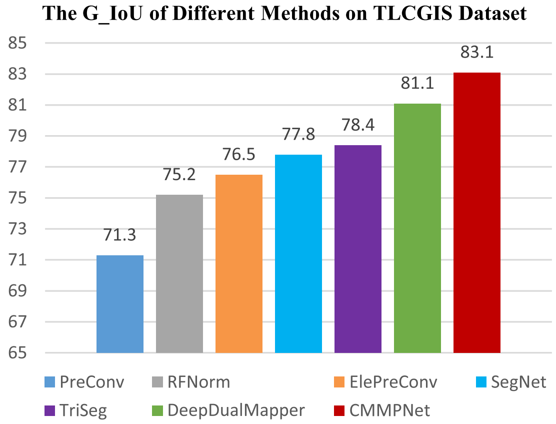


Fig. 10. Performance of different methods on the TLCGIS dataset. The proposed CMMPNet also outperforms all existing approaches for image + LiDAR-based road extraction.

A. Comparison With State-of-the-Art Metods

In this section, we compare our CMMPNet with six state-of-the-art methods on the TLCGIS dataset. The details of these

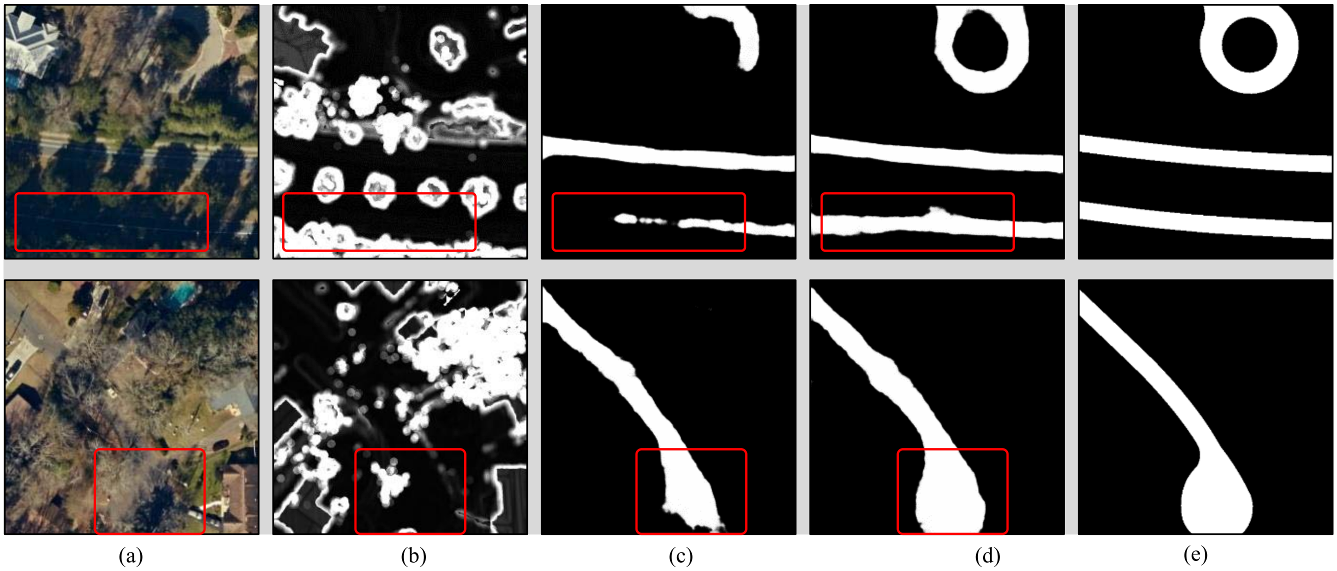


Fig. 11. Visualization of the generated traffic road maps on the TLCGIS dataset. (a) and (b) Are the input aerial images and LiDAR images. (c) Are the results that only aerial images are taken as input. (d) Are the results of our CMMPNet that utilizes both aerial images and LiDAR data. (e) Are the GT road maps. (a) Aerial images. (b) LiDAR data. (c) Image-based results. (d) Our results. (e) GT road maps.

compared methods are described as follows. **SegNet** [60]: As a fully convolutional Autoencoder, SegNet takes the concatenation of aerial images and LiDAR images as input. **PreConv** [59]: The LiDAR images are first fed into a depth convolution unit (DepthCNN) implemented with two convolutional layers. The LiDAR features and aerial images are then concatenated and fed into SegNet. **RFNorm** [59]: Given aerial and LiDAR images, some Random Forest classifiers [86] are first trained to estimate the road probability score at each location. The aerial-based and LiDAR-based score maps are concatenated and fed into SegNet. **ElePreConv** [59]: In this model, LiDAR images are first encoded with two convolutions (eight and four filters), while aerial images are extended with an extra zero-initialized channel. The element-wise addition of four-channel images and LiDAR features are fed into FuseNet [87]. **TriSeg** [59]: This model consists of three SegNets. The first two SegNets, respectively, take aerial or LiDAR images to generate the road probability maps, which are concatenated and fed into the third SegNet for final estimation. **DeepDualMapper** [29]: This model has been described above and here we adopt D-LinkNet as the backbone network to reimplement this model.

The performance of all approaches is summarized in Fig. 10. We can observe that the previous best-performing methods are TriSeg and DeepDualMapper, whose G_IoU are 78.4% and 81.1%, respectively. Thanks to the cross-modal mutual refinement strategy, our CMMPNet achieves a new state-of-the-art G_IoU 83.1% on the TLCGIS dataset and greatly outperforms DeepDualMapper with an absolute improvement of 2%. Moreover, we also visualize some results in Fig. 11. As can be observed, the traffic road maps generated by our method are more accurate in complex scenarios. In summary, these quantitative and qualitative comparisons demonstrate that our CMMPNet is universal and effective to extract traffic roads from aerial images and LiDAR data.

TABLE VII
PERFORMANCE OF DIFFERENT INPUTS AND DIFFERENT REPRESENTATION LEARNING MANNERS ON THE TESTING SET OF TLCGIS DATASET

Input	Learning Manner	G_IoU (%)
Lidar	-	69.12
Image	-	80.96
Image+Lidar	Late Fusion	81.50
	Early Fusion	81.62
	CMMPNet	83.10

B. Internal Analysis

In this section, we verify the effectiveness of each component in the proposed CMMPNet for image + LiDAR-based road extraction. We first explore which manner can better exploit the information of these modalities. As shown in Table VII, we can obtain a G_IoU of 69.12%, when only feeding the rendered LiDAR images into D-LinkNet. When only using aerial images, the G_IoU of D-LinkNet is 80.96%, which indicates that aerial images are more important. Incorporating the information of aerial and LiDAR images simultaneously, the early fusion model obtains a G_IoU of 81.62%, while the late fusion model has a comparable G_IoU of 81.50%. When fully exploring their complementary information with cross-modal message propagators, our CMMPNet achieves an impressive G_IoU 83.10%, outperforming the early/late fusion models with an absolute improvement of 1.5%. This demonstrates that the proposed CMMPNet can also effectively capture the complementary information among aerial images and LiDAR data.

We then explore the effect of global information and gate functions. Similar to Section V-C, we implement several variants of CMMPNet. As shown in Table VIII, when only propagating local information with gate functions, "Local + Gate" model obtains a G_IoU 82.31%. When incorporating global information, "Local + Global + Gate" model has a better

TABLE VIII

INFLUENCE OF NLM PROPAGATOR AND GATED MESSAGE PROPAGATOR ON THE TESTING SET OF TLCGIS DATASET

Local	Global	Gate	G_IoU (%)
✓			81.88
✓		✓	82.31
✓	✓		82.06
✓	✓	✓	83.10

G_IoU 83.10%, which indicates that the global information is also useful for image + LiDAR-based road extraction. Moreover, by comparing the performance of "Local + Global" model and "Local + Global + Gate" model, we can observe that the gate functions help to make an absolute improvement of 1.04% on G_IoU, which also demonstrates the effectiveness of gated message propagator for image + LiDAR-based road extraction.

VII. CONCLUSION

In this work, we investigate a challenging task for land remote-sensing analysis, i.e., how to robustly extract traffic roads using the complementary information of aerial images and vehicle crowdsourced trajectories. To this end, we introduce a novel CMMPNet, which learns modality-specific features explicitly with two individual autoencoders and enhances these features mutually with a tailor-designed DEM. Specifically, we comprehensively extract and dynamically propagate the complementary information of each modality to enhance the representation of another modality. Extensive experiments conducted on two real-world benchmarks show that the proposed CMMPNet is not only effective for image + trajectory-based road extraction, but also suitable for image + LiDAR-based road extraction.

Nevertheless, there are still several issues worthy of further study. **First**, the connectivity of traffic roads has not been explicitly explored in conventional works. Intuitively, the temporal information of vehicle trajectories could be utilized to distinguish disconnected road regions (e.g., urban roads are usually separated by fences and green belts). However, existing image + trajectory datasets lack the road connectivity annotation. To facilitate the researches in this field, we will construct a large-scale multimodal road extraction with rich connectivity annotation and propose a multimodal spatial-temporal framework to explicitly estimate the road connectivity in future work. **Second**, some elevated roads at different heights are overlapped on aerial images. The height information accessed with GPS devices is relatively coarse. Thus in future work, we will also develop some advanced approaches to effectively recognize the roads at different heights with the coarse height information of crowdsourced trajectories.

REFERENCES

[1] W. Von Engelhardt, J. Zimmermann, and J. Zimmerman, *Theory of Earth Science*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[2] N. R. Council *et al.*, *Basic Research Opportunities in Earth Science*. National Academies Press, 2001.

[3] G. R. Keller and C. Baru, *Cyberinfrastructure for the Solid Earth Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[4] Y. Zhang, Y.-L. Hsueh, W.-C. Lee, and Y.-H. Jhang, "Efficient cache-supported path planning on roads," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 951–964, Apr. 2016.

[5] T. Wang, Y. Zhao, J. Wang, A. K. Somani, and C. Sun, "Attention-based road registration for GPS-denied UAS navigation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 4, pp. 1788–1800, Apr. 2021.

[6] Q. Wang, T. Han, Z. Qin, J. Gao, and X. Li, "Multitask attention network for lane detection and fitting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 8, 2021, doi: 10.1109/TNNLS.2020.3039675.

[7] M. Wang and C. Luo, "Extracting roads based on Gauss Markov random field texture model and support vector machine from high-resolution RS image," *IEEE Trans. Geosci. Remote Sens.*, vol. 9, pp. 271–276, 2005.

[8] S. Movaghati, A. Moghaddamjoo, and A. Tavakoli, "Road extraction from satellite images using particle filtering and extended Kalman filtering," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 7, pp. 2807–2817, Jul. 2010.

[9] W. Shi, Z. Miao, and J. Debayle, "An integrated method for urban main-road centerline extraction from optical remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 6, pp. 3359–3372, Jun. 2014.

[10] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, Jun. 1999.

[11] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. London, U.K.: Springer-Verlag, 2009, ISBN 978-1-84800-279-1. [Online]. Available: https://link.springer.com/book/10.1007/978-4-431-67044-5

[12] 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.

[13] A. Buslaev, S. Seferbekov, V. Iglovikov, and A. Shvets, "Fully convolutional network for automatic road extraction from satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 207–210.

[14] 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.

[15] S. Rogers, P. Langley, and C. Wilson, "Mining GPS data to augment road models," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 1999, pp. 104–113.

[16] S. Schroedl and K. Wagstaff, "Mining GPS traces for map refinement," *Data Min. Knowl. Discovery*, vol. 9, no. 1, pp. 59–87, 2004.

[17] J. J. Davies, A. R. Beresford, and A. Hopper, "Scalable, distributed, real-time map generation," *IEEE Pervasive Comput.*, vol. 5, no. 4, pp. 47–54, Oct. 2006.

[18] X. Liu, J. Biagioni, J. Eriksson, Y. Wang, G. Forman, and Y. Zhu, "Mining large-scale, sparse GPS traces for map inference: Comparison of approaches," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 669–677.

[19] J. Biagioni and J. Eriksson, "Map inference in the face of noise and disparity," in *Proc. 20th Int. Conf. Adv. Geogr. Inf. Syst. (SIGSPATIAL)*, 2012, pp. 79–88.

[20] Y. Wang, X. Liu, H. Wei, G. Forman, and Y. Zhu, "CrowdAtlas: Self-updating maps for cloud and personal use," in *Proc. 11th Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2013, pp. 27–40.

[21] Z. Shan, H. Wu, W. Sun, and B. Zheng, "COBWEB: A robust map update system using GPS trajectories," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, 2015, pp. 927–937.

[22] S. Karagiorgou, D. Pfoser, and D. Skoutas, "A layered approach for more robust generation of road network maps from vehicle tracking data," *ACM Trans. Spatial Algorithms Syst.*, vol. 3, no. 1, pp. 1–21, May 2017.

[23] W. Shi, S. Shen, and Y. Liu, "Automatic generation of road network map from massive GPS, vehicle trajectories," in *Proc. 12th Int. IEEE Conf. Intell. Transp. Syst.*, Oct. 2009, pp. 1–6.

[24] C. Chen, C. Lu, Q. Huang, Q. Yang, D. Gunopulos, and L. Guibas, "City-scale map creation and updating using GPS collections," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1465–1474.

[25] W. Yang, T. Ai, and W. Lu, "A method for extracting road boundary information from crowdsourcing vehicle GPS trajectories," *Sensors*, vol. 18, no. 4, p. 1261, Apr. 2018.

[26] S. Ruan *et al.*, "Learning to generate maps from trajectories," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 1, pp. 890–897, Apr. 2020.

[27] T. Sun, Z. Di, P. Che, C. Liu, and Y. Wang, "Leveraging crowdsourced GPS data for road extraction from aerial imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7509–7518.

[28] Y. Li, L. Xiang, C. Zhang, and H. Wu, "Fusing taxi trajectories and RS images to build road map via DCNN," *IEEE Access*, vol. 7, pp. 161487–161498, 2019.

[29] H. Wu, H. Zhang, X. Zhang, W. Sun, B. Zheng, and Y. Jiang, "Deepdualmapper: A gated fusion network for automatic map extraction using aerial images and trajectories," in *AAAI Conference on Artificial Intelligence*, 2020.

[30] *Earth Science*. [Online]. Available: https://en.wikipedia.org/wiki/Earth_science

[31] G. A. Fine, *Authors of the Storm: Meteorologists and the Culture of Prediction*. Chicago, IL, USA: Univ. Chicago Press, 2009.

[32] M. Birylo, "The creation of flood risks model using a combination of satellite and meteorological models the first step," *Acta Geodynamica Geomaterialia*, vol. 12, no. 2, pp. 151–156, May 2015.

[33] Y. Y. Kagan and D. D. Jackson, "Probabilistic forecasting of earthquakes," *Geophys. J. Int.*, vol. 143, no. 2, pp. 438–453, Nov. 2000.

[34] R. N. Clark and A. N. Rencz, "Spectroscopy of rocks and minerals, and principles of spectroscopy," *Manual remote Sens.*, vol. 3, pp. 3–58, Jun. 1999.

[35] W. Wang, N. Yang, Y. Zhang, F. Wang, T. Cao, and P. Eklund, "A review of road extraction from remote sensing images," *J. Traffic Transp. Eng.*, vol. 3, no. 3, pp. 271–282, 2016.

[36] H. Zhang, Y. Liao, H. Yang, G. Yang, and L. Zhang, "A local-global dual-stream network for building extraction from very-high-resolution remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 16, 2020, doi: 10.1109/TNNLS.2020.3041646.

[37] Q. Lin, J. Zhao, G. Fu, and Z. Yuan, "CRPN-SFNet: A high-performance object detector on large-scale remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 416–429, Jan. 2022.

[38] S. Hinz and A. Baumgartner, "Automatic extraction of urban road networks from multi-view aerial imagery," *J. Photogramm. Remote Sens.*, vol. 58, nos. 1–2, pp. 83–98, 2003.

[39] P. N. Anil and S. Natarajan, "A novel approach using active contour model for semi-automatic road extraction from high resolution satellite imagery," in *Proc. 2nd Int. Conf. Mach. Learn. Comput.*, Feb. 2010, pp. 263–266.

[40] D. Chaudhuri, N. K. Kushwaha, and A. Samal, "Semi-automated road detection from high resolution satellite images by directional morphological enhancement and segmentation techniques," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 5, pp. 1538–1544, Oct. 2012.

[41] S. Leninisha and K. Vani, "Water flow based geometric active deformable model for road network," *ISPRS J. Photogramm. Remote Sens.*, vol. 102, pp. 140–147, Apr. 2015.

[42] Y. Bengio, I. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2016.

[43] 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.

[44] 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 (CVPRW)*, Jun. 2018, pp. 182–186.

[45] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[46] G. Fu, C. Liu, R. Zhou, T. Sun, and Q. Zhang, "Classification for high resolution remote sensing imagery using a fully convolutional network," *Remote Sens.*, vol. 9, no. 6, p. 498, 2017.

[47] S. Edelkamp and S. Schrödl, "Route planning and map inference with global positioning traces," in *Computer Science in Perspective*. Berlin, Germany: Springer, 2003, pp. 128–151.

[48] R. Stanojevic, S. Abbar, S. Thirumuruganathan, S. Chawla, F. Filali, and A. Aleimat, "Robust road map inference through network alignment of trajectories," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 135–143.

[49] L. Cao and J. Krumm, "From GPS traces to a routable road map," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst. (GIS)*, 2009, pp. 3–12.

[50] B. Niehoefer, R. Burda, C. Wietfeld, F. Bauer, and O. Lueert, "GPS community map generation for enhanced routing methods based on trace-collection by mobile phones," in *Proc. 1st Int. Conf. Adv. Satell. Space Commun.*, Jul. 2009, pp. 156–161.

[51] S. Wang, Y. Wang, and Y. Li, "Efficient map reconstruction and augmentation via topological methods," in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2015, pp. 1–10.

[52] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *Ann. Statist.*, vol. 20, no. 3, pp. 1236–1265, Sep. 1992.

[53] S. Clode, P. J. Kootsookos, and F. Rottensteiner, "The automatic extraction of roads from LiDAR data," *Proc. ISPRS*, Istanbul, Turkey, Jul. 2004.

[54] Z. Hui, Y. Hu, S. Jin, and Y. Z. Yevenyo, "Road centerline extraction from airborne LiDAR point cloud based on hierarchical fusion and optimization," *ISPRS J. Photogramm. Remote Sens.*, vol. 118, pp. 22–36, Aug. 2016.

[55] W. Zhang, "LiDAR-based road and road-edge detection," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2010, pp. 845–848.

[56] X. Hu, Y. Li, J. Shan, J. Zhang, and Y. Zhang, "Road centerline extraction in complex urban scenes from LiDAR data based on multiple features," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 11, pp. 7448–7456, Nov. 2014.

[57] Y. Li *et al.*, "Deep learning for LiDAR point clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2021.

[58] X. Hu, C. V. Tao, and Y. Hu, "Automatic road extraction from dense urban area by integrated processing of high resolution imagery and LiDAR data," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 35, no. B3, pp. 288–292, 2004.

[59] B. Parajuli, P. Kumar, T. Mukherjee, E. Pasiliao, and S. Jambawalikar, "Fusion of aerial LiDAR and images for road segmentation with deep CNN," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2018, pp. 548–551.

[60] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[62] S. Ross, D. Munoz, M. Hebert, and J. A. Bagnell, "Learning message-passing inference machines for structured prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 2737–2744.

[63] J. Winn and C. M. Bishop, "Variational message passing," *J. Mach. Learn. Res.*, vol. 6, pp. 661–694, Apr. 2005.

[64] D. Wang, W. Ouyang, W. Li, and D. Xu, "Dividing and aggregating network for multi-view action recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 451–467.

[65] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1774–1783.

[66] L. Zhang, D. Xu, A. Arnab, and P. H. S. Torr, "Dynamic graph message passing networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3726–3735.

[67] Z. Zhong, C.-T. Li, and J. Pang, "Hierarchical message-passing graph neural networks," 2020, *arXiv:2009.03717*.

[68] L. Liu, J. Chen, H. Wu, J. Zhen, G. Li, and L. Lin, "Physical-virtual collaboration modeling for intra- and inter-station metro ridership prediction," *IEEE Trans. Intell. Transp. Syst.*, early access, Nov. 24, 2020, doi: 10.1109/TITS.2020.3036057.

[69] I. Spinelli, S. Scardapane, and A. Uncini, "Adaptive propagation graph convolutional network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 10, pp. 4755–4760, Oct. 2021.

[70] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, early access, Nov. 30, 2021, doi: 10.1109/TPAMI.2021.3131222.

[71] Z. Wang *et al.*, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5764–5773.

[72] L. Liu, Z. Qiu, G. Li, Q. Wang, W. Ouyang, and L. Lin, "Contextualized spatial–temporal network for taxi origin-destination demand prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3875–3887, May 2019.

[73] J. Lou, Y. Jiang, Q. Shen, R. Wang, and Z. Li, "Probabilistic regularized extreme learning for robust modeling of traffic flow forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 16, 2020, doi: 10.1109/TNNLS.2020.3027822.

[74] L. Liu *et al.*, "Dynamic spatial-temporal representation learning for traffic flow prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 11, pp. 7169–7183, Nov. 2021.

[75] *Gaussian Blur*. [Online]. Available: https://en.wikipedia.org/wiki/Gaussian_blur

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[77] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[78] X. Wang, R. B. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.

[79] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[80] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[81] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[82] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[83] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[84] G. Lin, C. Shen, I. Reid, and A. van den Hengel, “Deeply learning the messages in message passing inference,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 361–369.

[85] M. T. Teichmann and R. Cipolla, “Convolutional CRFs for semantic segmentation,” in *Proc. Brit. Mach. Vis. Conf.*, 2019, pp. 1–12.

[86] A. M. Liaw ja Wiener, “Classification and regression by randomForest,” *R News*, vol. 2, pp. 18–22, Dec. 2007.

[87] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture,” in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 213–228.

Lingbo Liu received the Ph.D. degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2020.

From March 2018 to May 2019, he was a Research Assistant with The University of Sydney, Sydney, NSW, Australia. He is currently a Post-Doctoral Fellow with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. He has authored or coauthored more than 15 articles in top-tier academic journals and conferences. His current research interests include machine learning and urban computing.

Dr. Liu has been serving as a Reviewer for numerous academic journals and conferences such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, CVPR, ICCV, and International Joint Conference on Artificial Intelligence.

Zewei Yang received the B.E. degree from the School of Mathematics, Sun Yat-sen University, Guangzhou, China, in 2019, where she is currently pursuing the master’s degree in applied mathematics.

Her current research interests include machine learning and data mining.

Guanbin Li received the Ph.D. degree from The University of Hong Kong, Hong Kong, in 2016.

He is currently an Associate Professor with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. He has authored or coauthored more than 60 articles in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning.

Dr. Li was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as an Associate Editor for journal of *The Visual Computer*, the Area Chair for the conference of VISAPP. He has been serving as a Reviewer for numerous academic journals and conferences such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and Conference on Neural Information Processing Systems.

Kuo Wang (Member, IEEE) received the B.E. degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2020, where he is currently pursuing the Ph.D. degree in computer science.

His current research interests include deep learning and recommended system.

Tianshui Chen received the B.E. degree from the School of Information and Science Technology, and the Ph.D. degree in computer science from the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China, in 2013 and 2018, respectively.

He is currently a Lecturer with the Guangdong University of Technology, Guangzhou. He has authored or coauthored approximately 20 articles published in top-tier academic journals and conferences. His current research interests include computer vision and machine learning.

Dr. Chen was a recipient of the Best Paper Diamond Award at IEEE ICME 2017. He has served as a Reviewer for numerous academic journals and conferences, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, CVPR, ICCV, ECCV, Association for the Advancement of Artificial Intelligence, and International Joint Conference on Artificial Intelligence.

Liang Lin (Senior Member, IEEE) was a Post-Doctoral Fellow with University of California, Los Angeles, Los Angeles, CA, USA, from 2008 to 2010. From 2014 to 2015, he was with The Hong Kong Polytechnic University, Hong Kong, and The Chinese University of Hong Kong, Hong Kong, as a Senior Visiting Scholar. From 2017 to 2018, he leaded the SenseTime R&D Teams, Beijing, China, to develop cutting-edges and deliverable solutions on computer vision, data analysis and mining, and intelligent robotic systems. He is a Full Professor with Sun Yat-sen University, Guangzhou, China. He is the Excellent Young Scientist of the National Natural Science Foundation of China, Beijing. He has authored or coauthored more than 100 articles in top-tier academic journals and conferences.

Dr. Lin is a fellow of IET. He was a recipient of the Best Paper Runners-Up Award in ACM NPAR 2010, the Google Faculty Award in 2012, the Hong Kong Scholars Award in 2014, the Best Paper Diamond Award in IEEE ICME 2017, and the Best Paper Nomination Award in ICCV 2019. He served as the Area/Session Chairs for numerous conferences such as ICME, ACCV, ICMR. He has been serving as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, *The Visual Computer*, and *Neurocomputing*.