ISPRS Journal of Photogrammetry and Remote Sensing 163 (2020) 82–97

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Deep learning-based remote and social sensing data fusion for urban region function recognition

Rui Cao[a,b,c], Wei Tu[a], Cuixin Yang[b], Qing Li[a,b,d], Jun Liu[b], Jiasong Zhu[a], Qian Zhang[c], Qingquan Li[a], Guoping Qiu[b,d,*]

[a] Guangdong Key Laboratory of Urban Informatics & Shenzhen Key Laboratory of Spatial Smart Sensing and Services & MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area, Shenzhen University, Shenzhen 518060, China
[b] College of Electronics and Information Engineering & Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China
[c] International Doctoral Innovation Centre & School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China
[d] School of Computer Science, University of Nottingham, Nottingham NG8 1BB, UK

ARTICLE INFO

Keywords:
Urban function recognition
Multi-modal data fusion
Remote sensing
Social sensing
Deep learning

ABSTRACT

Urban region function recognition is key to rational urban planning and management. Due to the complex socioeconomic nature of functional land use, recognizing urban region function in high-density cities using remote sensing images alone is difficult. The inclusion of social sensing has the potential to improve the function classification performance. However, effectively integrating the multi-source and multi-modal remote and social sensing data remains technically challenging. In this paper, we have proposed a novel end-to-end deep learning-based remote and social sensing data fusion model to address this issue. Two neural network based methods, one based on a 1-dimensional convolutional neural network (CNN) and the other based on a long short-term memory (LSTM) network, have been developed to automatically extract discriminative time-dependent social sensing signature features, which are fused with remote sensing image features extracted via a residual neural network. One of the major difficulties in exploiting social and remote sensing data is that the two data sources are asynchronous. We have developed a deep learning-based strategy to address this missing modality problem by enforcing cross-modal feature consistency (CMFC) and cross-modal triplet (CMT) constraints. We train the model in an end-to-end manner by simultaneously optimizing three costs, including the classification cost, the CMFC cost and the CMT cost. Extensive experiments have been conducted on publicly available datasets to demonstrate the effectiveness of the proposed method in fusing remote and social sensing data for urban region function recognition. The results show that the seemingly unrelated physically sensed image data and social activities sensed signatures can indeed complement each other to help enhance the accuracy of urban region function recognition.

1. Introduction

Nowadays, more than half of the world population resides in cities, which only cover less than 2% of the earth surface. As rapid urbanization is undergoing in Asia and Africa, the urban population is thus still expanding and estimated to grow to 5 billion by 2030 across the whole world (Tu et al., 2018). Therefore, it is of great importance to monitor and manage the limited urban areas for such a huge population.

Urban region function recognition is key to rational urban planning and management. It refers to the inference of the usage purposes of urban regions directly associated with human activities, such as residential, commercial, entertaining, and educational (Zhang et al., 2017; Zhang et al., 2019). It is related to but also different from traditional land use and land cover (LULC) classification; the latter usually stresses on physical characteristics of the earth surface, while the former focuses purely on socioeconomic functional attributes of urban regions. LULC monitoring using remote sensing imagery has been proven to be efficient and effective, since these images can well capture the natural appearance of the land surface. However, region function recognition using remote sensing images alone is not sufficient, especially in high-density cities, such as Shenzhen, London, and New York. This is due to the following facts: (1) urban region functions are of socioeconomic properties and determined by the related human

* Corresponding author at: College of Electronics and Information Engineering & Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen 518060, China.
E-mail address: guoping.qiu@nottingham.ac.uk (G. Qiu).

https://doi.org/10.1016/j.isprsjprs.2020.02.014
Received 1 November 2019; Received in revised form 16 January 2020; Accepted 24 February 2020
Available online 11 March 2020
0924-2716/ © 2020 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). Published by Elsevier B.V. All rights reserved.

activities; (2) shadows of numerous high-rise buildings in high-density cities pose great challenges for remote sensing image processing; (3) mixed urban functions are often clustered in one building or block in east Asian cities.

With the rapid development of information and communication technologies (ICTs), social sensing big data recording human dynamics are becoming increasingly available, such as vehicle GPS trajectories (Tu et al., 2018; Liu et al., 2012), points of interest (POI) (Liu et al., 2017; Hu et al., 2016), mobile phone positioning data (Jia et al., 2018; Tu et al., 2017), social media check-in data (Tu et al., 2017; Gao et al., 2017), and geotagged photos (Cao et al., 2018; Zhu and Newsam, 2015). Different from remote sensing images, these social sensing data are the by-products of human daily life; therefore, they contain rich socioeconomic attributes. When these data meet with remote sensing, the promising trend is to fuse them to recognize urban functions, since the two kinds of data are complementary to each other (Liu et al., 2015). However, remote and social sensing data are significantly different in terms of sources and modalities. Generally, remote sensing images cover the study area. While social sensing data are place-based thus represented by points, polylines, or polygons. Besides, the features of social sensing data may be time-based (Tu et al., 2017), rather than space-based. The fusion of the two multi-source and multi-modal data is no trivial. The key challenge is to alleviate the modality gap and heterogeneity between them.

The emergence of deep learning has advanced many research fields, including image recognition (LeCun et al., 2015), time series classification (Fawaz et al., 2019), and etc. They have also greatly boosted the development of remote sensing (Zhu et al., 2017; Zhang et al., 2016). Significant improvements have been made in many tasks, such as hyperspectral image analysis (Li et al., 2019), image scene classification (Cheng et al., 2017), semantic labeling (Audebert et al., 2018), object detection (Cheng and Han, 2016), and image retrieval (Cao et al., 2020). The major advantages of deep learning approaches are the powerful abilities to automatically learn high-level features from large amount of data, which are vital to bridge the gap between different data modalities at feature level. Therefore, deep learning-based fusion methods are very potential to integrate the multi-source and multi-modal remote and social sensing data.

In this paper, to address the problem of urban region function recognition with cross-modal data sources, we propose an end-to-end deep learning-based multi-modal data fusion method to integrate remote sensing images and social sensing signatures. The two kinds of data are firstly fed into modal-specific encoders of residual convolutional neural network (CNN) and our proposed 1d CNN/LSTM-based network respectively to extract effective features, then those features are fused, and the outputs are finally put into fully connected layers and softmax layer to make predictions. We also propose two ancillary losses to further constrain the network training by drawing the two extracted multi-modal features nearer, which ensures the robustness against missing modalities. Open available datasets are exploited to evaluate our methods, and the results demonstrate their effectiveness and efficiency. In addition, thorough analysis of the methods and results have been conducted to provide insights into fusing the two types of data. Our contributions are summarized as follows:

1. We propose an end-to-end deep multi-modal fusion method to effectively incorporate the multi-modal remote sensing images and social sensing signatures for urban region function recognition.
2. We propose two effective neural networks to extract temporal signature features automatically. One is 1d CNN-based, and the other is LSTM-based, both of which can explicitly take temporal dependencies into account and effectively extract sequential-aware features.
3. To address the data asynchronous problem, we propose two auxiliary losses, i.e. the cross-modal feature consistency (CMFC) loss and the cross-modal triplet (CMT) loss, to make the proposed multi-modal fusion network more robust to missing modalities without significantly impacting the performances.
4. We have conducted extensive experiments on open available datasets to evaluate the effectiveness and efficiency of the proposed methods. We also analyze and discuss the results thoroughly to give insights into fusing the two multi-modal data.

The rest of the paper is organized as follows. In Section 2, we review related works on remote and social sensing for LULC classification and urban function recognition. In Section 3, we illustrate and formulate the problem of integrating remote sensing images and social sensing signatures for urban region function recognition. Section 4 elaborates the proposed methods to extract image and temporal signature features, as well as the multi-modal deep learning fusion for urban region function classification. Section 5 evaluates our methods on publicly available datasets and analyzes the results. In Section 6, we discuss about several important issues concerning the strength and possible improvement of the proposed methods. Finally, we conclude in Section 7.

2. Related work

2.1. Remote sensing for LULC classification

Land use and land cover classification through remote sensing imagery is a fundamental research topic in remote sensing community. Due to the limited spatial resolution of optical remote sensing imagery, pixel-centric spectral-based methods are the mainstream of traditional LULC classification works (Blaschke et al., 2014). However, the rapid development of high spatial resolution remote sensing imagery brings opportunities in digging into more complex spatial patterns, and geographic object-based image analysis (GEOBIA) (Blaschke et al., 2014) has thus become a new paradigm for LULC classification. It firstly divides imagery into segmented objects and then classifies them. Recently, the advance of deep learning techniques empowers the abilities to learn more high-level semantic features from images (LeCun et al., 2015). This also significantly benefits the development of LULC studies. Deep learning-based land use and land cover classification is explored both in pixel-level (Audebert et al., 2018; Marmanis et al., 2018), object-level (Zhang et al., 2019; Zhang et al., 2018), and scene-level (Cheng et al., 2017; Albert et al., 2017).

2.2. Social sensing for urban function studies

Urban function recognition differs from traditional LULC classification in the stress on socioeconomic functional attributes of cities. Traditional methods heavily depend on labor-intensive land survey, which is time-consuming and expensive. To ease the situation, social sensing data-driven methods emerge with the increasing access to ICT big data (Liu et al., 2015). Vehicle trajectories and public transport data are demonstrated to be informative for urban function recognition. Gao et al. (2019) identify urban region function from ride-hailing vehicle trajectories. Du et al. (2019) use multi-modal transportation data to recognize urban functional zones. Pan et al. (2013) use taxi GPS traces to classify functional urban land use. Liu et al. (2012) reveal the intra-urban land use variation from taxi traffic patterns. Social media data are also exploited to characterize urban functions. Yuyun et al. (2017) utilize social media data to generate dynamic functional land use map. Yao et al. (2017) categorize urban land use from POI data. Gao et al. (2017) extract urban functional regions from POIs and social media check-in data. Besides, mobile phone positioning data are effectively utilized for urban analysis. Pei et al. (2014) exploit mobile phone data for urban functional land use classification. Then, mobile phone and social media data are further integrated to categorize urban functional regions and portray urban dynamics (Tu et al., 2019; Tu et al., 2017; Cao et al., 2015). In addition, geotagged social media and street view images are also leveraged for urban function studies, such as functional

land use mapping (Zhu et al., 2019; Srivastava et al., 2018; Li et al., 2017; Zhu and Newsam, 2015; Leung and Newsam, 2012), building function classification (Kang et al., 2018), and urban mobility pattern prediction (Zhang et al., 2019). These works demonstrate that social sensing data can provide important socioeconomic information for urban function recognition.

2.3. Fusion of remote and social sensing

Due to different data sources, modalities, and resolutions of remotely sensed data, data fusion has thus long been studied in remote sensing community (Ghamisi et al., 2019). With the emerging of geotagged ICT big data, the fusion of remote and social sensing data has attracted increasingly attention recently (Deng et al., 2019; Qi et al., 2019; Li et al., 2017). Zhang et al. (2019) integrate remote sensing images, mobile phone positioning data, and POIs for functional zone classification using cross-correlations. Tu et al. (2018) use hierarchical clustering to combine satellite imagery and mobile phone positioning data to portray urban functional zones. Jia et al. (2018) extract handcrafted features from satellite imagery and mobile phone positioning data, and then use support vector machine (SVM) for land use classification. Liu et al. (2017) exploit satellite images and social media data to classify land use using probability topic model and SVM. Chen et al. (2018) develop a novel framework to integrate satellite images and POI data for social functional mapping of urban green space. Zhang et al. (2017) integrate satellite images and POI data for urban functional zone mapping using hierarchical semantic cognition. Hu et al. (2016) combine satellite images and POIs to classify urban land parcels by calculating feature similarities. Jendryke et al. (2017) use SAR imagery and social media message to conduct urban land use change analysis. Besides, remote sensing and street-level images are also integrated for land cover classification (Lefèvre et al., 2017; Chi et al., 2017), urban land use classification (Srivastava et al., 2019; Cao et al., 2018; Cao and Qiu, 2018; Workman et al., 2017; Zhang et al., 2017), zoning (Feng et al., 2018), and building type classification (Hoffmann et al., 2019).

Remote and social sensing data are of significantly different sources and modalities, they possess different information about urban land surface and are complementary to each other (Liu et al., 2015; Tu et al., 2018; Zhang et al., 2019). However, it is difficult to fuse them directly due to enormous data modality gap. Thus, it is important but also challenging to develop remote and social sensing data fusion methods to improve urban function recognition. Most existing works use handcrafted features, which require human experts and are laborious. Due to powerful representation learning ability of deep learning, the paper therefore presents an effective deep learning-based multi-modal fusion method to automatically extract features from remote sensing imagery and social sensing signature data, and further fuse them to identify urban region function in an end-to-end manner.

3. Problem statement

The problem of urban region function recognition from multi-source and multi-modal remote sensing imagery and social sensing signature data is illustrated in Fig. 1. It can be formally defined as follows: for a region $\mathcal{R}_i$, given the satellite imagery $\boldsymbol{I}_i$ and the social sensing signature $\boldsymbol{S}_i$ of the region, the class c_i that $\mathcal{R}_i$ belongs to is to be predicted, as formulated in Eq. 1:

$$c_i^* = \arg\max_k p(c_k|\boldsymbol{I}_i, \boldsymbol{S}_i), \tag{1}$$

where $c_i \in C = \{c_k|k = 1, 2, ..., C\}$, C is the number of function types. Common functions include residential, commercial, entertaining, transportation, and etc. Social sensing temporal signature $\boldsymbol{S}_i = \{h(t)|t = 1, 2, 3, ..., n\}$ is time series data and reflects human dynamics over time. Typical examples include hourly aggregated number of mobile phone calls (Pei et al., 2014), human activity strength over time (Tu et al., 2018; Jia et al., 2018), transportation ridership over time (Tu et al., 2018), and etc.

In real-world scenarios, the remote and social sensing data are asynchronous. Therefore, another issue of note is the robustness of model against missing data, since the missing of one kind of data may severely corrupt the established model trained with complete data. For example, the model may be trained with $\boldsymbol{I}$ and $\boldsymbol{S}$, but tested with either $\boldsymbol{I}$ or $\boldsymbol{S}$ in some specific regions. Thus, effective methods against missing modalities are also needed in real world scenarios.

4. Methodology

In this section, we describe the details of the proposed deep multi-modal fusion network, which is capable of integrating remote sensing images and social sensing temporal signature (TS) data for urban region function recognition.

4.1. Overview of the deep multi-modal fusion network

The overall architecture of the proposed deep multi-modal fusion network is presented in Fig. 2. The network Φ is composed of three major parts, i.e. the image encoder ϕ_i, the temporal signature (TS) encoder ϕ_s, and the data fusion module ϕ_f. The network takes satellite images $\boldsymbol{I}$ and temporal signatures $\boldsymbol{S}$ as inputs, and outputs the predicted probability distribution $\boldsymbol{p}$ over all the categories, i.e. $\boldsymbol{p} = \Phi(\boldsymbol{I}, \boldsymbol{S}) = \phi_f(\phi_i(\boldsymbol{I}), \phi_s(\boldsymbol{S}))$. Specifically, the modified ResNet (a) is exploited as the image encoder ϕ_i, and either the proposed 1-d SPP-Net (b) or LSTM-Net (c) is leveraged as the TS encoder ϕ_s. The extracted features are further fused and fed into fully connected (FC) layers and softmax layer for classification. The key of the network is to learn a joint embedding space where image and signature features can well join together to make predictions. Meanwhile, apart from the conventional cross entropy loss for classification, we propose two auxiliary losses, i.e. cross-modal feature consistency (CMFC) loss and cross-modal triplet (CMT) loss, to enforce the proposed network to learn more coordinated multi-modal features, which can make the network more robust to missing modalities. In the following subsections, the image encoder, the temporal signature encoder, the fusion methods, and the loss functions for network training are described in details.

4.2. Image encoder

Residual neural networks (ResNets) (He et al., 2016), specifically ResNet-18 and ResNet-50, are modified as image encoder to extract features from satellite images. The ResNets leverage shortcut connections between every few stacked layers to achieve residual learning, which enables them to go deep and learn more effectively. The architecture of the modified ResNet is shown in Fig. 2a. The input images are firstly fed into a 7×7 convolutional layer with stride of 2 (followed by batch normalization and ReLU activation), then a 3×3 max pooling layer with stride 2 is operated on the output. After that, the output feature maps are fed into four groups of residual blocks (ResBlocks). The output feature maps are further reduced to feature vectors by global average pooling. Finally, a fully connected layer is appended to produce the output, where the original 1000-dimensional fully connected layer is modified to 256-d. This output will be exploited as the final extracted image feature.

4.3. Temporal signature (TS) encoder

The 1-d SPP-Net and LSTM-Net are proposed as TS encoders to extract time-dependent features from time-series signature data. The former is a one-dimensional convolutional neural network enhanced by 1-d spatial pyramid pooling (SPP) (He et al., 2015) and the latter is a stacked bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) based network.

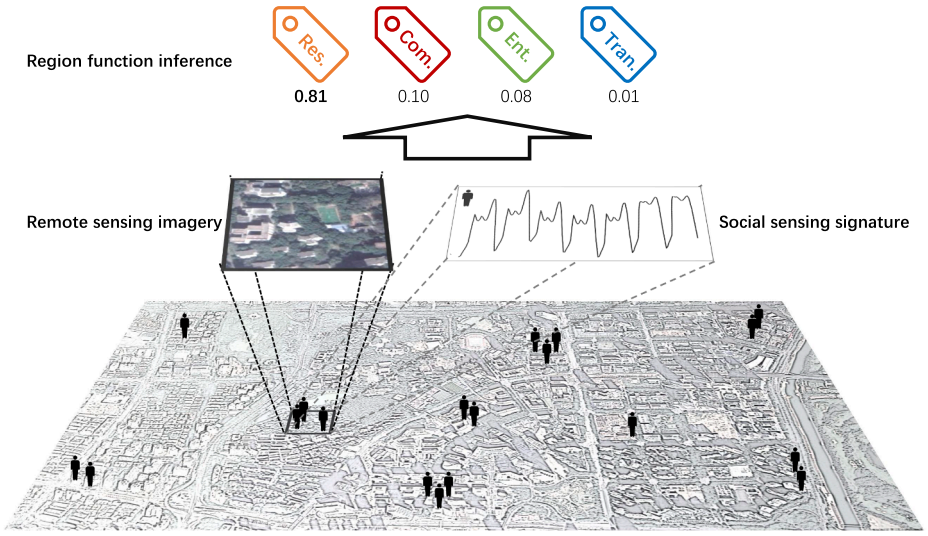

Fig. 1. Illustration of integrating remote sensing imagery and social sensing signature data for urban region function recognition. Common functions include residential (res.), commercial (com.), entertaining (ent.), transportation (tran.), and etc.

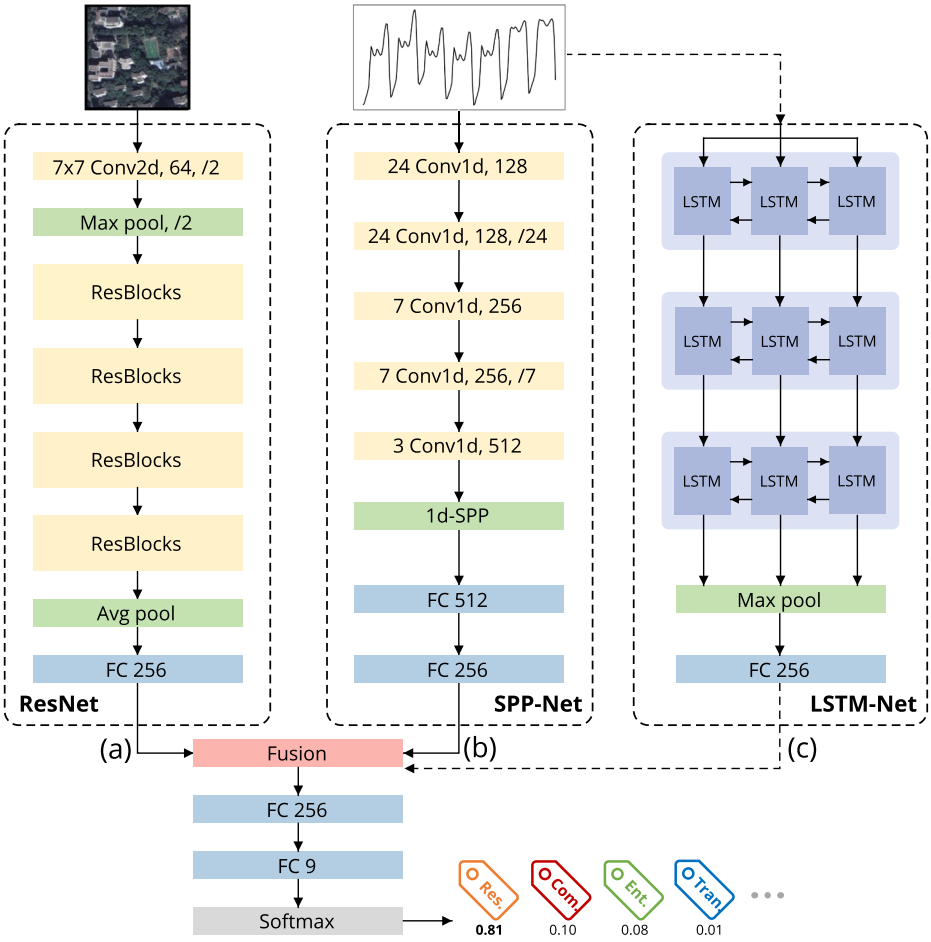

Fig. 2. Overview of the proposed deep multi-modal fusion network, which includes the modified ResNet (a) as the image encoder, and the 1-d SPP-Net (b) or the alternative LSTM-Net (c) as the temporal signature (TS) encoder.

4.3.1. 1-d spatial pyramid pooling network (SPP-Net)

Despite the tremendous success in image recognition, convolutional neural networks have recently also been used in time series analysis and shown superior performances (Fawaz et al., 2019). One-dimensional CNNs can automatically learn distinguishable features from time-series data without human expert knowledge, thus showing promising potential for time series classification. Inspired by the spatial pyramid pooling (SPP) (He et al., 2015) used for image recognition, we propose a one-dimensional SPP enhanced network to extract distinctive features from temporal signatures.

The architecture of the proposed 1-d SPP-Net is presented in Fig. 2b. The network is composed of three major parts, i.e. five 1-d

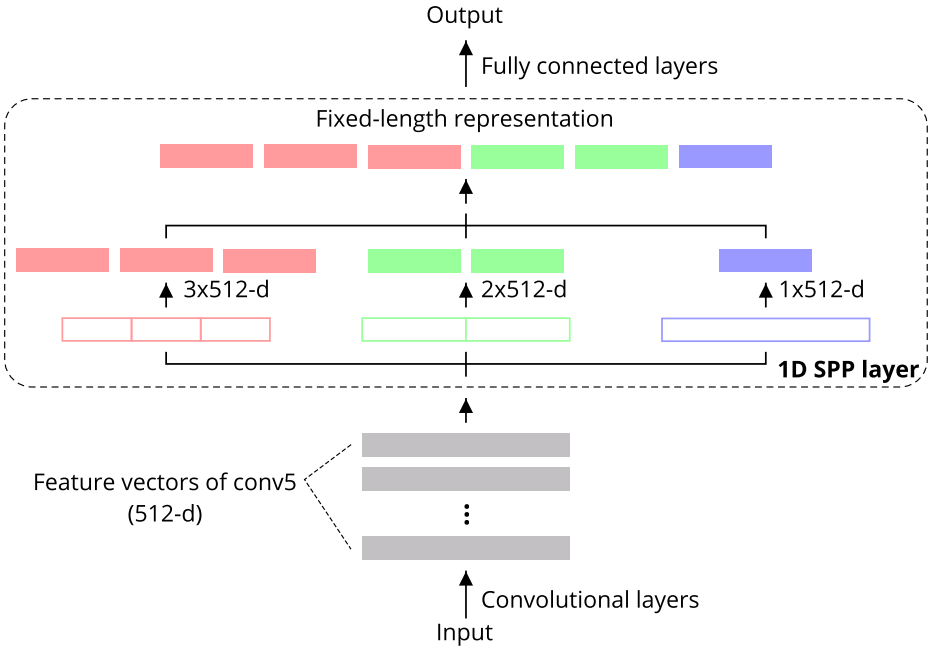

Fig. 3. The structure of 1-d spatial pyramid pooling (SPP) layer.

convolutional layers (each layer is closely followed by batch normalization and ReLU activation), one 1-d SPP layer, and two fully connected layers. The convolutional layers use kernel size of 24, 24, 7, 7, and 3, with stride of 1, 24, 1, 7, and 1, and padding of 12, 12, 3, 3, and 1, respectively. The output of the convolutional layers is then fed into the 1-d SPP layer to obtain fixed-length representation, which is further processed by the fully connected layers. The final output 256-d features are exploited as extracted TS features. The convolutional layers can extract short-term time-series features, by stacking them, the patterns with long-term dependencies can be captured in higher layers. The 1-d SPP layer and fully connected layers can further aggregate and condense the convolutional features to higher level semantic features.

The structure of the proposed 1-d SPP layer is illustrated in Fig. 3. The layer is composed of bins of three levels, with number of 1, 2, and 3, respectively. The bin sizes are proportional to the size of input features. For the output feature vectors of the last convolutional layer (conv5), each channel-wise feature is aggregated by max pooling within the three-level bins. Then, the aggregated features of three levels are concatenated together before further fed into the fully connected layers. Due to the multi-level pooling operation, the 1-d SPP layer can aggregate multi-scale information; besides, it can aggregate extracted convolutional features into fixed-length output features regardless of the input time sequence length. The output feature size is $N_b \times N_k$, only related to the number of bins N_b and the number of kernels N_k of conv5, specifically in our setting, the output size is 6×512.

4.3.2. Stacked bidirectional long short-term memory network (LSTM-Net)

Recurrent neural networks (RNNs) are a kind of neural network designed to deal with sequential data, such as textual data and time-series data. However, conventional RNNs are incapable of capturing long-term dependencies from data due to the vanishing gradient problem. The long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997), a special type of RNN, is proposed to address this issue. Therefore, we propose an LSTM-based network to extract distinctive sequential-aware features from temporal signatures.

The architecture of the proposed LSTM-Net is shown in Fig. 2c. The network consists of three major parts, i.e. three stacked bidirectional LSTM layers, a max pooling layer, and a fully connected layer with 256 neurons. The input sequential data are firstly reshaped and processed by the three stacked bidirectional LSTM layers to extract temporally dependent features. Then, the extracted features are aggregated by the max pooling layer, the output of which is further fed into the fully connected layer to obtain the final extracted feature.

The basic LSTM uses cell and hidden state to store the long and short-term memories of the information of past sequence, respectively. Each element $\boldsymbol{x}_t$ in the input sequence $\{\boldsymbol{x}_t | t = 1, 2, ..., n\}$ will be processed by the recurrent LSTM cell, together with previous cell state $\boldsymbol{c}_{t-1}$ and hidden state $\boldsymbol{h}_{t-1}$, the information will be regulated by three specially designed gates $\boldsymbol{i}_t, \boldsymbol{f}_t$, and $\boldsymbol{o}_t$, i.e. the input, forget, and output gates, to obtain current cell state $\boldsymbol{c}_t$ and output hidden state $\boldsymbol{h}_t$. The process can be formulated as Eq. 2:

$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t + \boldsymbol{W}_{hi}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\
\boldsymbol{f}_t &= \sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t + \boldsymbol{W}_{hf}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\
\boldsymbol{o}_t &= \sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t + \boldsymbol{W}_{ho}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t + \boldsymbol{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t * \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{2}
$$

where σ is the sigmoid function, and $*$ is the element-wise product. $\boldsymbol{W}$s and $\boldsymbol{b}$s are corresponding weights and biases to be learned.

As illustrated in Fig. 4, bidirectional LSTM extends the basic LSTM by enabling the model to process the sequence from two ends. It preserves information from both the forward and backward directions, i.e. concatenating the forward hidden state $\overrightarrow{\boldsymbol{h}_t}$ and the backward $\overleftarrow{\boldsymbol{h}_t}$ as the final output $\boldsymbol{h}_t$ for each time step. This allows for faster relevant information searching and better context-aware feature learning than that of one-directional. Furthermore, stacking the bidirectional LSTM layers can add higher level of abstraction and enhance the model ability of learning more complex patterns. Our experiments also confirm the effectiveness of the inclusion of the bidirectional mechanism and the stacking design.

4.4. Fusion methods

As presented in Fig. 2, the extracted image feature $\boldsymbol{x}^i$ and TS feature $\boldsymbol{x}^s$ from the aforementioned modal-specific encoders ϕ_i and ϕ_s are further fused in feature-level to produce the fused feature $\boldsymbol{x}^f$, where $\boldsymbol{x}^i, \boldsymbol{x}^s \in \mathbb{R}^n (n = 256)$. Specifically, three methods are harnessed to fuse the extracted features of the two different modalities, i.e. concatenation, element-wise sum, and element-wise max pooling. For

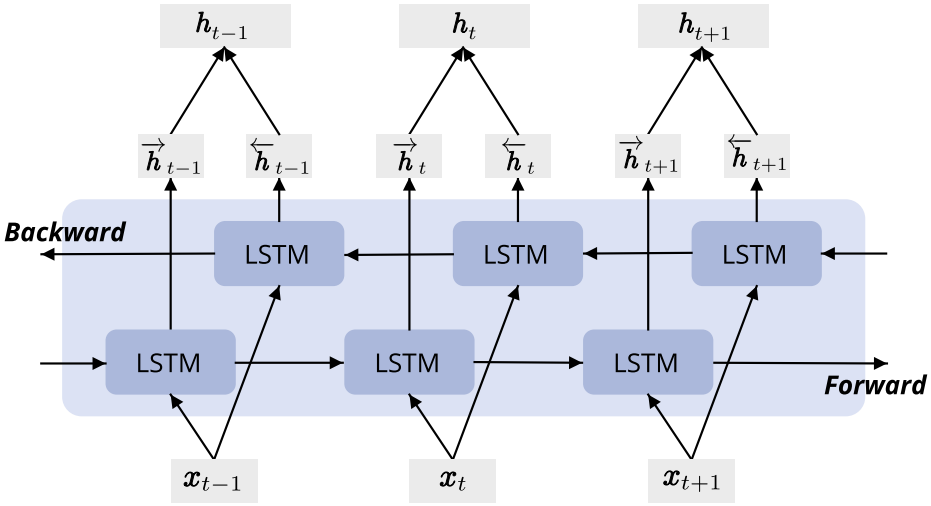

Fig. 4. The structure of bidirectional LSTM layer.

concatenation, $\boldsymbol{x}^f = [\boldsymbol{x}^i, \boldsymbol{x}^s]$, and $\boldsymbol{x}^f \in \mathbb{R}^{2n}$. For element-wise sum, $\boldsymbol{x}^f = \boldsymbol{x}^i + \boldsymbol{x}^s$, and $\boldsymbol{x}^f \in \mathbb{R}^n$. For element-wise max pooling, $\boldsymbol{x}_k^f = max(\boldsymbol{x}_k^i, \boldsymbol{x}_k^s)$, where $k = 1, 2, ..., n$, and $\boldsymbol{x}^f \in \mathbb{R}^n$. The illustration of the three fusion methods are shown in Fig. 5. Furthermore, the fused feature $\boldsymbol{x}^f$ is then fed into fully connected layers and softmax layer to make the final prediction.

4.5. Loss functions

To achieve effective classification as well as make the network more robust to missing modalities, two kinds of losses are proposed to constrain the network training, i.e. the major and auxiliary losses. The major loss is the *cross entropy loss* $\mathcal{L}_{ce}$ for the classification task. The auxiliary losses are used to complement the major loss to increase the model robustness against missing modalities, including the *cross-modal feature consistency (CMFC) loss* $\mathcal{L}_{fc}$ and the *cross-modal triplet (CMT) loss* $\mathcal{L}_{tri}$. The overall loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}, \tag{3}$$

where λ_1 and λ_2 control the relative importance of the losses.

The cross entropy loss is used to regularize the network to learn from category labels, and it is widely used in classification tasks. It can be formulated as follows:

$$\mathcal{L}_{ce} = -\sum_i \sum_{k=1}^{C} y_{i,k} log\,(p_{i,k}), \tag{4}$$

where $y_{i,k}$ and $p_{i,k}$ are the ground truth label and predicted probability value of class k for the i-th sample, C is the total number of classes.

The cross-modal feature consistency (CMFC) loss is computed via cosine distance as Eq. 5:

$$\mathcal{L}_{fc} = \sum_k \left(1 - \frac{\boldsymbol{x}_k^i \cdot \boldsymbol{x}_k^s}{\|\boldsymbol{x}_k^i\| \cdot \|\boldsymbol{x}_k^s\|}\right), \tag{5}$$

where $\boldsymbol{x}_k^i$ and $\boldsymbol{x}_k^s$ are the features of the image and temporal signature data respectively, $\|\boldsymbol{x}\|$ is the length of vector $\boldsymbol{x}$. The cosine distance is equivalent to the cosine similarity, which is widely used to measure document similarity in information retrieval. Analogous to document alignment, since both the image and signature data are indicative of the same urban function properties of the same region, there ought to be correlation between them in spite of different modalities. The CMFC loss enforces the features of image and signature to be consistent and similar with regard to vector orientation.

The cross-modal triplet (CMT) loss further utilizes the category information and tries to draw cross-modal features of the same class nearer, while push features of different classes far away. It can be formulated as Eq. 6:

$$\mathcal{L}_{tri} = \sum_{(u,v)} \sum_{\substack{a,\,p,\,n \\ l_a = l_p \neq l_n}} [d(\boldsymbol{x}_a^u, \boldsymbol{x}_p^v) - d(\boldsymbol{x}_a^u, \boldsymbol{x}_n^v) + m]_+, \tag{6}$$

where $[x]_+$ represents $max(x, 0)$;$\boldsymbol{x}_a$, $\boldsymbol{x}_p$, and $\boldsymbol{x}_n$ are the features of the anchor, positive, and negative samples, and l_a, l_p, l_n are their corresponding labels; $d(\boldsymbol{x}_a, \boldsymbol{x}_p)$ and $d(\boldsymbol{x}_a, \boldsymbol{x}_n)$ are the cosine distances between the anchor-positive and the anchor-negative pairs, respectively; m denotes the margin that enforces a distance between similar and dissimilar pairs. $(u, v) \in \{(i, s), (s, i)\}$, where i and s represent the image and temporal signature modality, respectively.

5. Experiments

5.1. Datasets

In this paper, the Urban Region Function Classification (URFC) datasets [1] are used to evaluate the proposed methods. The datasets are collected from urban areas in China. Two kinds of data are provided, i.e. satellite images and user visit data, with labels of 9 categories, i.e. residential area (res.), school (sch.), industrial park (ind.), railway station (rail.), airport (air.), park (park), shopping area (shop.), administrative district (adm.), and hospital (hosp.). For each region, a high-resolution satellite image patch and a corresponding user visit file are provided.

There are two subsets of URFC datasets, and we denote them as URFC-A and URFC-B in this paper. The URFC-A dataset comprises 50,000 pairs of data, of which only 40,000 pairs are with public ground truth labels. While the URFC-B dataset includes 10 times the size of URFC-A, and 400,000 pairs are with public labels. In our experiments, only the data with public ground truth labels of URFC-A and URFC-B are utilized so that we can use the labels for evaluation. The data distributions of each category for the two datasets are shown in Table 1. It can be seen that URFC-A and URFC-B vary in data distributions, and the data samples are distributed unevenly across different categories.

5.1.1. Satellite image data

The satellite image patches are of the size of 100×100 pixels. Five images of each category from URFC-A and URFC-B datasets are presented in Fig. 6. As can be noticed, the satellite images are of high

[1] https://dianshi.baidu.com/competition/30/data.

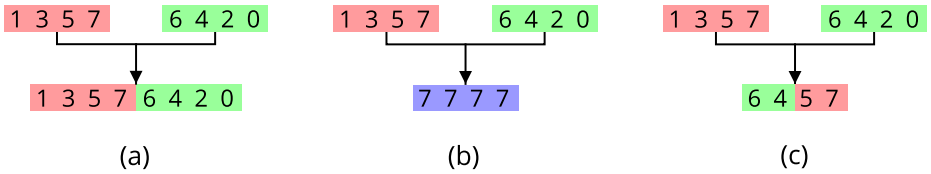

Fig. 5. Illustration of fusion methods for multi-modal features, i.e. (a) concatenation, (b) element-wise sum, (c) element-wise max pooling.

Table 1
The categorical data distribution of the URFC-A and URFC-B datasets.

	res.	sch.	ind.	rail.	air.	park	shop.	adm.	hosp.	Sum
URFC-A	9,542	7,538	3,590	1,358	3,464	5,507	3,517	2,617	2,867	40,000
URFC-B	120,370	91,053	51,015	6,588	16,494	62,684	21,135	13,181	17,480	400,000

Fig. 6. Overview of the satellite image data for the nine categories. The images of the left and right five columns are from URFC-A and URFC-B respectively.

resolution, however, the small size of the image patches limits the recognizable contents of the images. In addition, there are many low-quality images as shown in the latter two patches for each category. It can also be seen that the visual contents of a same class vary significantly, and there are very similar contents, such as high-rise buildings and green vegetation, across different categories. These make the satellite image datasets of high intra-class variation and inter-class similarity, which implies the difficulty in distinguishing the images from different categories.

5.1.2. User visit data

In this paper, the univariate temporal signatures of user visit, i.e. time-series sequences of hourly aggregated number of user visit, are used for experimental evaluation. The original user visit data record the occurrence time of users within regions. The range of visit time is from October 1st in 2018 to March 31st in 2019, lasting for half a year (182 days, 26 weeks). The user visit data of each region are recorded in a file separately. Each user visit data file corresponds to a satellite image patch of the same region. In the experiments, the user visit data are aggregated to obtain the hourly user visit count, i.e. the visit numbers over 4,368 h, which are harnessed as the temporal signatures to measure the strength of user activities. It should be noticed that the individual user information is not exploited in order to maintain the generality of the data.

Fig. 7 presents the average user visit number over hours of week for the nine categories of URFC-A and URFC-B. It can be seen that, in general, the temporal signature patterns of the two datasets are similar across most categories, only with different magnitudes and scales in some classes. We can also see that the average visit numbers vary dramatically across different categories. The patterns of visit numbers are significantly different between workdays (from Monday to Friday) and weekends for *residential area*, *school*, *industrial park*, *administrative district*, and *hospital*; while the patterns are similar for categories like *park* and *shopping area*. The temporal distributions of *railway station* and *airport* are similar for both workdays and weekends, only with smaller numbers on Saturday and Sunday. For workdays, the *residential area*

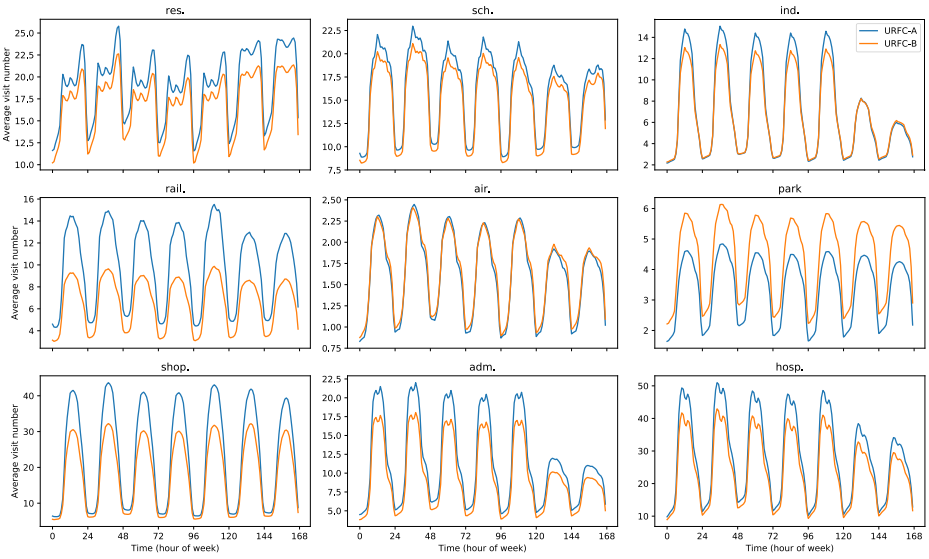

Fig. 7. Overview of the aggregated user visit data in terms of average visit number over hours of week for the nine categories of the URFC-A and URFC-B datasets.

category exhibits an obvious three-peak pattern; the *administrative district* and *hospital* both show a two-peak pattern but with different maximum peaks; while the *school* and *industrial park* categories present a noticeable one-peak pattern. Therefore, as can be seen, the temporal signatures of user visit can help distinguish region functions.

5.2. Experimental setup

All the networks in the experiments are implemented using the PyTorch (Paszke et al., 2019) framework. Uni-modal network classifiers are composed of single encoders and two fully connected layers with 256 and 9 neurons respectively, which are denoted as ResNets, SPP-Net, and LSTM-Net respectively, according to the encoder shown in Fig. 2. The multi-layer perceptron (MLP) consists of four layers, with 1024, 256, 256, and 9 neurons, respectively. The random forest (RF) and support vector machine (SVM) classifiers are implemented by the scikit-learn library. For data fusion comparison, the decision-level (late) fusion method is also examined as baseline, which includes two major steps, i.e. uni-modal classification and decision fusion. The satellite image and temporal signature data are firstly exploited separately to learn two classifiers, and the decision scores of the two uni-modal classifiers are fused by summation afterwards to make the final decision.

We have conducted two groups of experiments, and the data settings are presented in Table 2. The first group splits URFC-A as training, validation, and testing sets, while URFC-B is used for generalization testing. The other group divides URFC-B into training, validation, and testing sets, whereas takes URFC-A as generalization testing set. For both groups, the training, validation, and testing sets account for 80%, 10%, and 10% of the original datasets respectively, while the generalization testing set includes the whole datasets.

5.3. Evaluation metrics

To evaluate the classification results, we adopt overall accuracy, Kappa coefficient, F1 score, and time cost as evaluation metrics. Let $x_{i,j}$ denotes the element of i-th row and j-th column in the confusion matrix, i.e. the number of samples of class i that are predicted to be class j;n is the number of classes, and N is the total number of all the samples. Then the evaluation metrics can be formulated as follows:

(1) *Overall accuracy*: $p_0 = \sum_{i=1}^{n} x_{ii}/N$.

(2) *Kappa coefficient*:

$$K = \frac{p_0 - p_e}{1 - p_e}, \tag{7}$$

where $p_e = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} x_{i,j} \sum_{j=1}^{n} x_{j,i} \right) / N^2$.

(3) *F1 score*:

$$F1_i = \frac{2 p_i r_i}{p_i + r_i}, \tag{8}$$

where p_i and r_i are the precision and recall score of class i respectively, $p_i = x_{ii} / \sum_{j=1}^{n} x_{ij}$, $r_i = x_{ii} / \sum_{j=1}^{n} x_{ji}$. $F1_i$ measures the classification result of a certain class i. While the *average F1 score* ($\overline{F1}$) is the average of all the F1 scores of different categories and can measure the overall classification results of all the n classes:

$$\overline{F1} = \frac{1}{n} \sum_{i=1}^{n} F1_i. \tag{9}$$

(4) *Time cost*: The time cost (*minutes*) of network training is reported to emphasize the importance of model efficiency.

Table 2
Experiment settings of data split.

Setting	Training	Validation	Testing	Generalization
A => B	32,000	4,000	4,000	400,000
B => A	320,000	40,000	40,000	40,000

5.4. Network training

Adam optimizer is used to train the networks, with a learning rate of 0.0003, which is decayed by a factor of 0.1 for each epoch after 10 epochs. The training batch size is 192. The maximum training iteration is set to 20 epochs. Cross entropy loss is exploited as optimization objective. In addition, the parameters of convolutional layers of the ResNets are initialized by weights pretrained on ImageNet (Deng et al., 2009). The images are resized to 128×128. Random horizontal and vertical flips, random rotation of 90, 180, 270 degrees, and random cropping and padding are used for image data augmentation while training.

The learning curves of training loss and validation accuracy on URFC-A and URFC-B are presented in Fig. 8. It can be seen that the training loss is decreasing for each epoch, and there is a significant loss

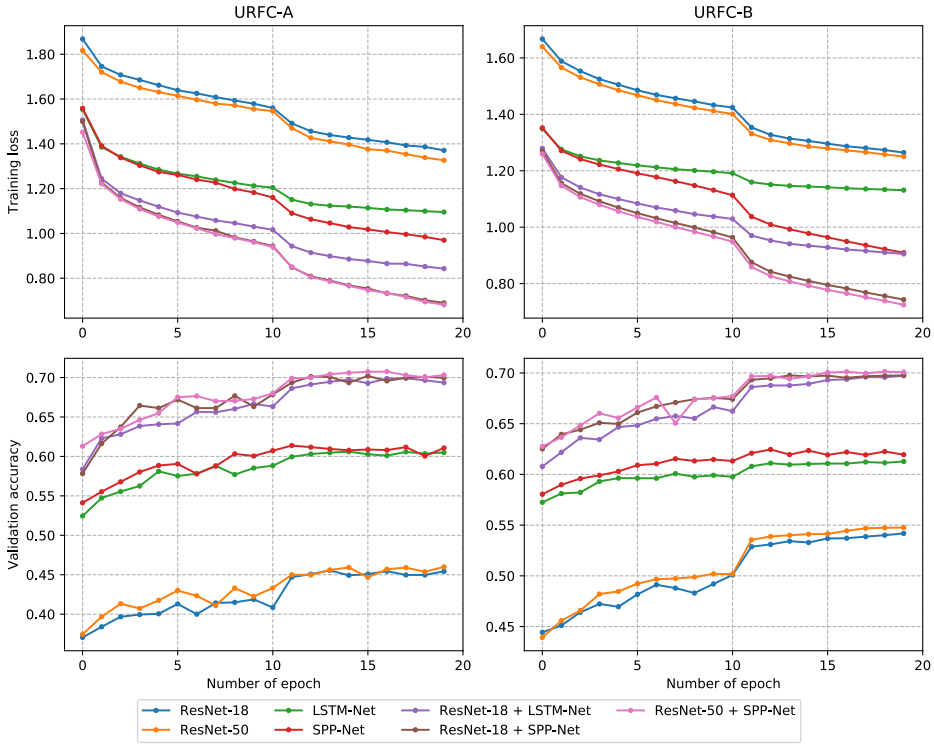

Fig. 8. The learning curves of different methods on URFC-A and URFC-B. The upper rows are curves of training cross entropy loss, and the lower rows are curves of validation accuracy.

drop after 10 epochs due to learning rate decay. On the contrary, the validation accuracy increases with the training going and tends to stabilize after 10 epochs. The training loss curves of the proposed fusion networks are significantly lower than that of networks using single kind of data, which demonstrates the effectiveness of the proposed data fusion method.

5.5. Results and analysis of multi-modal data fusion

5.5.1. Overall results

The overall classification results of URFC-A and URFC-B are presented in Tables 3 and 4, respectively. Therein lie the uni-modal classifiers as baselines for single data classification. Specifically, ResNet-18 and ResNet-50 are exploited for satellite image classification, while the proposed 1-d SPP-Net and LSTM-Net are used for temporal signature classification. Besides, the decision-level fusion network and two-stage classification method are leveraged as baselines for data fusion. For the two-stage method, random forest (RF) classifier is used, and the input features are composed of original temporal sequences and extracted deep features from images. Specifically, two different deep image features are harnessed, i.e. features extracted by ResNet-18 and ResNet-50, which are trained on the target datasets. The corresponding results are denoted as RF (ResNet-18) and RF (ResNet-50), respectively. The reported time costs of RF results include the training time of both image network and RF classifier on the target datasets. The proposed multi-modal fusion networks exploit ResNets and SPP-Net/LSTM-Net as image and temporal signature encoder, respectively.

As can be seen from Table 3, for using image data alone, both networks achieve testing accuracies slightly over 45%. Besides, using deeper model (ResNet-50) doesn't improve the classification results significantly than using its shallower counterpart (ResNet-18), with average F1 score increases around 1%, while overall accuracy decreases slightly and time cost is more than doubled. For using temporal signature data only, the proposed SPP-Net and LSTM-Net can achieve an testing accuracy of more than 60%, significantly improving the classification results compared with using single satellite images, which

Table 3
Overall classification results using models trained on the URFC-A dataset. The best results are highlighted in bold. (I: image, S: signature; late: late fusion).

Data	Method	Validation			Testing			Generalization			Time (min)
		Accuracy	Kappa	Avg.F1	Accuracy	Kappa	Avg.F1	Accuracy	Kappa	Avg.F1	
I	ResNet-18	45.05%	0.34	39.92%	**45.30%**	0.34	39.62%	**37.20%**	0.23	27.64%	**9**
	ResNet-50	**46.13%**	**0.36**	**41.71%**	45.28%	**0.35**	**40.65%**	36.75%	**0.23**	**28.30%**	22
S	LSTM-Net	60.63%	0.54	56.71%	60.05%	0.53	56.38%	56.05%	0.46	47.18%	**3**
	SPP-Net	**61.33%**	**0.54**	**57.46%**	**62.40%**	**0.56**	**59.21%**	**57.94%**	**0.48**	**49.59%**	5
I + S	RF (ResNet-18)	66.28%	0.60	63.67%	65.73%	0.59	63.20%	55.81%	0.45	47.67%	11
	RF (ResNet-50)	65.85%	0.60	62.61%	65.93%	0.60	63.84%	58.42%	0.48	51.28%	24
	ResNet-18 + SPP-Net (late)	63.05%	0.56	55.66%	63.88%	0.57	56.27%	56.36%	0.46	44.83%	13
	ResNet-18 + LSTM-Net	69.73%	0.64	68.08%	68.63%	0.63	66.88%	59.02%	0.50	51.19%	**11**
	ResNet-18 + SPP-Net	69.95%	0.65	68.60%	**70.20%**	**0.65**	68.45%	**60.59%**	**0.52**	**53.02%**	13
	ResNet-50 + SPP-Net	**71.00%**	**0.66**	**69.82%**	70.05%	0.65	**68.86%**	60.04%	0.51	52.62%	27

Table 4
Overall classification results using models trained on the URFC-B dataset. The best results are highlighted in bold. (I: image, S: signature; late: late fusion).

Data	Method	Validation			Testing			Generalization			Time (min)
		Accuracy	Kappa	Avg. F1	Accuracy	Kappa	Avg. F1	Accuracy	Kappa	Avg. F1	
I	ResNet-18	54.20%	0.41	44.97%	54.29%	0.41	44.36%	50.03%	0.39	43.97%	**90**
	ResNet-50	**54.78%**	**0.42**	**46.23%**	**55.02%**	**0.42**	**45.60%**	**50.97%**	**0.40**	**45.08%**	220
S	LSTM-Net	61.30%	0.51	50.21%	61.94%	0.52	50.62%	62.10%	0.55	57.40%	**30**
	SPP-Net	**62.61%**	**0.53**	**53.00%**	**62.97%**	**0.53**	**52.92%**	**65.90%**	**0.59**	**63.22%**	50
I + S	RF (ResNet-18)	67.71%	0.59	61.24%	67.95%	0.59	61.06%	60.77%	0.53	55.19%	118
	RF (ResNet-50)	68.04%	0.60	62.01%	68.37%	0.60	61.80%	64.42%	0.57	60.00%	248
	ResNet-18 + SPP-Net (late)	61.88%	0.51	44.93%	62.65%	0.52	45.41%	59.32%	0.51	47.90%	125
	ResNet-18 + LSTM-Net	69.53%	0.62	64.39%	69.35%	0.62	63.86%	71.69%	0.66	70.49%	**106**
	ResNet-18 + SPP-Net	70.17%	0.63	65.35%	**70.23%**	**0.63**	64.88%	74.63%	0.70	74.07%	127
	ResNet-50 + SPP-Net	**70.31%**	**0.63**	**65.61%**	70.14%	0.63	**64.89%**	**75.13%**	**0.71**	**74.84%**	255

further implies the importance of social sensing data in region function recognition applications. Meanwhile, the training time costs of neural networks are both less than 5 *min*, indicating the efficiency of the proposed networks. When using both data sources, the proposed networks (ResNet-18 + SPP-Net, ResNet-50 + SPP-Net) can dramatically improve the testing results by around 8% compared with the best results of using single source of data (SPP-Net). The results are also significantly better than that of the RF data fusion baselines, with more than 4% increase for all the metrics. For generalization testing, the overall results drop as expected, however, the relative performances of different models are in consistent with that of the testing results.

We can see from Table 4 that the general results are similar with that of Table 3, as the results of integrating both data sources significantly outperform that of using single data source, and the results of proposed fusion networks are better than the RF baselines. Besides, the results using image alone on URFC-B are noticeably increased by more than 8% than that of URFC-A, which indicates that sufficient image data can significantly improve the performances of deep models, helping them find more important spatial patterns existing in the satellite image patches.

It should also be noticed that, for both URFC-A and URFC-B, the classification results using images alone are significantly lower than that of using temporal signatures alone. There are two probable reasons: 1) the satellite image patches of small size can only provide limited useful region function information due to their limited spatial coverage extent as well as high intra-class variety and inter-class similarity; 2) the satellite images reflect more of the physical attributes of the land, while the temporal signature data can reflect human dynamics directly which are more relevant with semantic region functions. This further reinforces the need for incorporating remote sensing imagery with social sensing data to complement each other to improve urban region function recognition performance.

5.5.2. Per-class results

The per-class F1 score results of URFC-A and URFC-B using different sources of data are shown in Tables 5 and 6, respectively. For different input data, the results of ResNet-50, SPP-Net, and ResNet-50 + SPP-Net are reported. We can see that the results vary across different categories. This is due to two main reasons: 1) certain classes are very similar in terms of visual content or temporal signature, which results in different distinguishing difficulties; 2) the data samples are distributed unevenly for different categories. It can be seen that the image data perform better in the classes of *railway*, *airport*, and *park*, and they are especially effective in classifying *airport*, with the F1 score of more than 80%. While the temporal signature data are more distinguishable for all the other six classes. The fusion of the two data sources is effective for all the function categories, since all the results of fusing data are better than using any single source of data alone. The results demonstrate that the image and temporal signature data are complementary to each other, and proper fusion of the two data sources can help boost the classification results significantly across all the classes.

To further investigate which classes are easily confused with each other, the confusion matrices of testing results on URFC-A are presented in Fig. 9. The upper row presents the confusion matrices of only using satellite images, temporal signatures, and using both data respectively, while the bottom row shows the corresponding normalized matrices.

As is shown in Fig. 9, using satellite image and temporal signature data alone, the classification results display different patterns. For using satellite images alone, the *airport* samples can be well classified, with recall rate of 91%. The recall rate of *residential area* reaches 65%, however, it is at the cost of many wrongly classified samples as shown in the first column of sub-figure (d), since the samples of *residential area* dominate the whole dataset with about one fourth of the total samples. Besides, the *railway station* and *park* regions can achieve recall rate of over 50%. While the regions of *administrative area*, *hospital*, and *shopping area* are severely misclassified, with recall rates less than 30%, and most of them (more than 37%) are incorrectly assigned to *residential area*.

For using temporal signature data alone, the *residential area* category can achieve a recall rate of 82%; the recall rates of *hospital*, *school*, and *shopping area* all exceed 60%; while the regions of *railway station* are poorly categorized, with recall rate of only 30%. Interestingly, the *airport* and *park* regions are easily confused with each other through temporal signatures only. This is consistent with the temporal variation patterns of the average visit number shown in Fig. 7.

When combining the two kinds of data, the classification results are further enhanced, with recall rates of all the categories achieving more than 50%. Besides, the recall rates of almost all the categories are improved compared with only using one kind of data except for *residential area*. The misclassification of tangled classes is also noticeably alleviated. These further confirm that the image and temporal signature

Table 5
Per-class F1 score of testing results on URFC-A using different data sources. The best results are highlighted in bold. (I: image, S: signature).

	res.	sch.	ind.	rail.	air.	park	shop.	adm.	hosp.	Avg.F1
I	49.46%	39.53%	38.55%	50.00%	84.01%	52.39%	29.97%	2.95%	18.99%	40.65%
S	71.64%	71.27%	57.81%	42.05%	51.70%	43.30%	68.51%	54.39%	72.19%	59.21%
I + S	**72.46%**	**72.60%**	**64.52%**	**64.31%**	**85.76%**	**62.91%**	**69.30%**	**55.25%**	**72.63%**	**68.86%**

Table 6
Per-class F1 score of testing results on URFC-B using different data sources. The best results are highlighted in bold. (I: image, S: signature).

	res.	sch.	ind.	rail.	air.	park	shop.	adm.	hosp.	Avg.F1
I	59.83%	51.20%	59.50%	44.58%	81.62%	57.73%	31.78%	10.16%	13.96%	45.60%
S	72.10%	70.61%	62.30%	30.42%	30.80%	49.23%	56.27%	44.27%	60.29%	52.92%
I + S	**73.43%**	**74.06%**	**72.46%**	**50.96%**	**82.35%**	**63.85%**	**59.82%**	**46.19%**	**60.86%**	**64.89%**

data can complement to each other, and effectively fusing the two data sources can significantly improve the classification results.

5.5.3. Qualitative results

It is of note from Fig. 9 that some function categories are significantly misclassified, such as *administrative area*, *hospital*, and *shopping area* for only using image data; *railway station* when only using temporal signature data; some are very easily confused with each other, such as *airport* and *park* when only using temporal signatures. To go deeper and investigate into more details of the classification results, eight typical examples are presented in Fig. 10, and the corresponding classification results using image data alone, temporal signatures alone, and using both data are presented in Table 7.

For some cases, the satellite image data are more useful than the temporal signature data. For instance, Fig. 9e shows that the regions of *airport* and *park* are easily confused with each other when only using temporal signatures. Fig. 10a and b present a very good example of this kind; we can see that the temporal signatures are extremely sparse which indicates rare human dynamics in this region, so that it is very hard to figure out the function of the region via temporal signatures alone. However, in these cases, the image data can contribute more valuable information than the signatures: for Fig. 10a, there is an airplane in the image which implies that the region is part of an airport; for Fig. 10b, the landscape of forest and lake indicates the high probability of the region being part of a park. For Fig. 10c, the region is mistaken for *park* when using temporal signature data alone, however, the closely spaced metal roofs in the image indicate that the region is highly likely to be part of a railway station.

Nevertheless, the temporal signatures are far more distinguishable than the satellite image patches in some more scenarios. For Fig. 10d, e, f, and g, the regions of different functions are all mistaken for *residential area* through images alone, which is understandable since barely much useful information is presented in the small image patches. In contrast, the temporal signatures provide more direct features about human dynamics, which help to make the correct predictions. As expected, the temporal signatures also present similar patterns as that shown in Fig. 7 (*adm., hosp., shop., sch.*). Fig. 9d shows that *residential area* and *school* are easily confused when only using image data. This is just the case of Fig. 10g and h. It is hard to tell whether the region belongs to *residential area* or *school* via the satellite image of Fig. 10g, while the temporal signature demonstrates a higher probability of school pattern. Interestingly, for Fig. 10h, the track-and-field ground is easily recognized in the image patch which can thus confuse us to think the region as part of a school; however, the temporal signature tells a different story, showing the place to be *residential*. The contradiction between the image and signature data may be due to the reason that the track-and-field ground is near the residential buildings or part of a large residential community. In these cases, the temporal signatures can provide us with more useful information about human dynamics in terms of region functions.

As we can see, the satellite images and temporal signatures are complementary to each other, they can provide informative clues of region functions from very different perspectives, and the combination of both information helps to improve the classification results significantly.

5.6. Ablation study of proposed losses against missing modalities

Due to real-world data quality issues, there are situations when multiple sources of data are ready in the training phase, while only unimodal data are available in the testing phase. On the one hand, we want to make full use of all the available data when training; on the other hand, we also want to ensure the trained model robust to the missing of some modalities in the application stage. Therefore, the cross-modal feature consistency (CMFC) loss and cross-modal triplet (CMT) loss are proposed to constrain the training of the proposed network. The former loss can enforce the networks to learn cross-modal features with maximized cosine similarity between paired data, while the latter enforces ranking constrains from categorical labels to the learned features. Intuitively, the auxiliary losses enforce the networks to learn a common

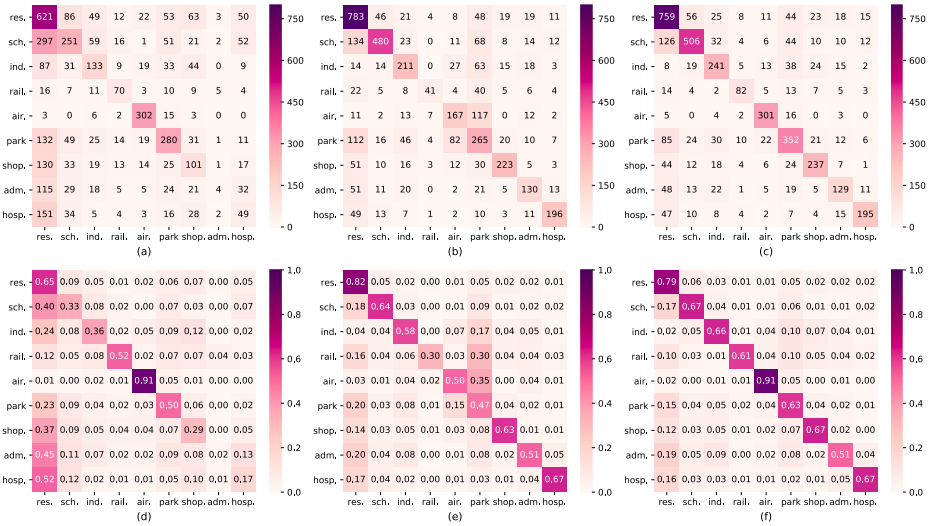

Fig. 9. Confusion matrices of testing results on URFC-A using different data sources, i.e. using satellite images (a, d), temporal signatures (b, e), and both data (c, f), respectively. The upper row (a, b, c) presents confusion matrices, while the bottom row (d, e, f) shows corresponding normalized confusion matrices.

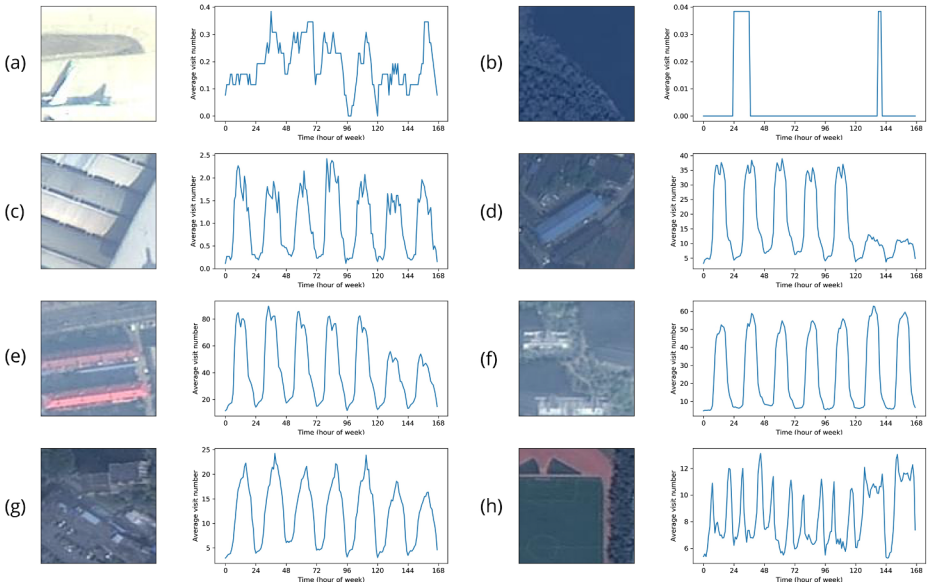

Fig. 10. Case study of eight representative examples: (a) airport, (b) park, (c) railway station, (d) administrative district, (e) hospital, (f) shopping area, (g) school, and (h) residential area. For each case, the left is the satellite image patch of the region, and the right is the corresponding temporal signature of average visit numbers over hours of week.

Table 7
Classification results of case study examples using different input data, corresponding to Fig. 10. (I: image, S: signature).

Case	Label	I	S	I + S	Dominant
(a)	air.	air.	park	air.	image
(b)	park	park	air.	park	image
(c)	rail.	rail.	park	rail	image
(d)	adm.	res.	adm.	adm.	visit
(e)	hosp.	res.	hosp.	hosp.	visit
(f)	shop.	res.	shop.	shop.	visit
(g)	sch.	res.	sch.	sch.	visit
(h)	res.	sch.	res.	res.	visit

embedding space where data of the same categories but different modalities can be mapped as similar features. Hence, when one modality is missing, the features of the missing modality can be replaced by their counterpart features from the other modality to reduce the negative effects of data missing as much as possible.

Experiments were conducted to validate the effectiveness of the proposed CMFC and CMT losses. In the experiments, all the fusion networks are trained with complete modalities of data, while tested in three different scenarios: 1) with single images (I), 2) with single temporal signatures (S), and 3) with both modalities of data (I + S). These scenarios can simulate the extreme situations when all the data of a modality are missing during testing. The uni-modal classifiers (ResNet-18 and SPP-Net) and both the early and late fusion networks trained under cross entropy loss alone are used as baselines. The margin m of the CMT loss is empirically set to 0.1. The loss weights λ_1 and λ_2 are empirically set to 0.1 and 1, respectively. The testing results on URFC-A and URFC-B are presented in Tables 8 and 9, respectively. In general, as can be seen from both tables, for the data missing situations, the testing performances of fusion networks are lower than that of modal-specific classifiers, i.e. ResNet-18 and SPP-Net, which is understandable since the uni-modal classifiers are trained and tested on the same uni-modal data. However, the constraint of the proposed losses can help alleviate the gap significantly.

Specifically, for using single image data on URFC-A (as shown in Table 8), the testing results of the fusion network drop dramatically than that of ResNet-18 classifier, with a decrease of about 9% for both the overall accuracy (from 45.30% to 36.30%) and average F1 score (from 39.62% to 30.59%). While for the fusion network with CMFC loss constraint, the drop is significantly alleviated, with an increase of 4.88% in accuracy (41.18%) and 1.62% in average F1 score (32.21%). The additional constraint of CMT loss further improves the results, with overall accuracy increases to 41.60% and average F1 score to 33.66%. Similarly, for URFC-B (as shown in Table 9), the testing results of the fusion network decline noticeably from 54.29% to 43.09% in overall accuracy and from 44.36% to 34.80% in average F1 score; however, the additional constraints of CMFC and CMT loss terms significantly relieve the drop, with increases of 3.66% and 1.40% for accuracy and F1 score respectively. Though the late-fusion network achieves higher accuracy than its counterpart early-fusion network, the average F1 score is significantly lower on both datasets. In summary, the proposed network with the constraint of proposed losses significantly outperforms all the baselines when only using image data.

For using single temporal signature data on URFC-A (as shown in Table 8), the proposed network without additional losses achieves the best results, while the extra constraints of the proposed losses decrease the results slightly, with overall accuracy and average F1 score decline less than 1%. On the contrary, for URFC-B (as shown in Table 9), the additional constraints of the proposed losses increase the results, with an increase of about 1% for both overall accuracy and average F1 score. This indicates that the imbalanced contribution of the two kinds of data will affect the performance of the additional constraints imposed by the proposed losses. This is understandable since the proposed losses aim to maximize the similarity of the features from the two data sources. In the experiment, the contributions of the satellite image and temporal signature data vary significantly, with 17.10% of accuracy gap between uni-modal networks of ResNet-18 (45.30%) and SPP-Net (62.40%) for URFC-A and 8.68% for URFC-B. These constraints may therefore negatively influence the accuracy of the data with higher contribution when the imbalance is significant, like the case of URFC-A for temporal signature data.

Moreover, for using both sources of data, the addition of the two proposed losses can even improve the testing results on URFC-A (as shown in Table 8), with an increase of 0.80% in accuracy and 1.52% in average F1 score. For URFC-B (as shown in Table 9), the performances of the network trained with extra proposed losses only decrease slightly for both accuracy and F1 score. These results on both datasets demonstrate that the proposed CMFC and CMT loss are effective to train the proposed network against the issue of missing modalities, with very limited negative effects and even slight boost in performance.

6. Discussion

6.1. Capacity for sequential-aware modeling

Temporal dependency is one of the most important properties of the temporal signature data, which indicates the vital role of the sequential

Table 8
Comparison of testing results on URFC-A under situations of missing data. The best results are highlighted in bold and the second best underlined. (I: image, S: signature; late: late fusion. $\lambda_1 = 0.1$, $\lambda_2 = 1.0$).

Testing	Method	Loss	Metrics		
			Accuracy	Kappa	Avg.F1
I	ResNet-18	$\mathcal{L}_{ce}$	45.30%	0.34	39.62%
S	SPP-Net	$\mathcal{L}_{ce}$	62.40%	0.56	59.21%
I	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	40.33%	0.27	25.95%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	36.30%	0.26	30.59%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	41.18%	0.29	32.21%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	**41.60%**	**0.29**	**33.66%**
S	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	58.45%	0.51	49.88%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	**60.63%**	**0.54**	**56.77%**
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	59.95%	0.53	53.27%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	59.90%	0.53	56.31%
I + S	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	63.88%	0.57	56.27%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	70.20%	0.65	68.45%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	70.03%	0.65	68.48%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	**71.00%**	**0.66**	**69.97%**

order of the data. It is significantly important to take full advantage of the information for time-series data classification. In order to demonstrate the capability of sequential-aware modeling of the proposed 1-d SPP-Net and LSTM-Net, extra experiments have been conducted. The sequential order of the input temporal signatures is randomly shuffled, and then the classification results with shuffled input data are compared with that of original signatures. The testing results of URFC-A are presented in Table 10. Random forest (RF), support vector machine (SVM), and multi-layer perceptron (MLP) are utilized as baseline classifiers to compare with the proposed networks.

It can be seen that, for MLP, SVM, and RF, the resulting metrics of using shuffled data are similar or almost the same to that of using original data, with all the evaluation results less than 1% difference, which implies that the three classifiers do not account for the sequential order of the input data. On the contrary, it is different for both LSTM-Net and SPP-Net. When the input data are randomly shuffled, the values of metrics drop significantly, with accuracy, Kappa coefficient, average F1 score decrease 2.12%, 0.03, 2.52% respectively for LSTM-Net, and 3.57%, 0.05, 3.58% respectively for SPP-Net. These results show that the sequential order of input signatures is of great significance for the LSTM-Net and SPP-Net, which suggests that both networks possess the ability of modeling temporal dependencies. The 1-d SPP-Net encodes the short-term sequence within a sliding window via local convolution operation, and the long-term dependency can be captured by stacking multiple convolutional layers which increases the receptive fields of the network. While the LSTM-Net captures the sequential information by feeding the elements of a sequence recurrently into bidirectional LSTM unit, which produces outputs by fusing the current input (new information) with previous cell and hidden state (i.e. the long and short-term memories). Although the two methods function in different

Table 10
Comparison of testing results on original and shuffled input temporal signature data. The better results are highlighted in bold.

Method	Original			Shuffled			Time (min)
	Accuracy	Kappa	Avg.F1	Accuracy	Kappa	Avg.F1	
MLP	**52.75%**	**0.43**	**44.02%**	52.58%	0.43	43.31%	3
SVM	52.90%	0.43	43.12%	52.90%	0.43	43.12%	140
RF	57.28%	0.49	51.83%	**57.33%**	**0.49**	**52.06%**	2
LSTM-Net	**60.05%**	**0.53**	**56.38%**	57.93%	0.50	53.86%	3
SPP-Net	**62.40%**	**0.56**	**59.21%**	58.83%	0.51	55.63%	5

Table 9
Comparison of testing results on URFC-B under situations of missing data. The best results are highlighted in bold and the second best underlined. (I: image, S: signature; late: late fusion. $\lambda_1 = 0.1$, $\lambda_2 = 1.0$).

Testing	Method	Loss	Metrics		
			Accuracy	Kappa	Avg. F1
I	ResNet-18	$\mathcal{L}_{ce}$	54.29%	0.41	44.36%
S	SPP-Net	$\mathcal{L}_{ce}$	62.97%	0.53	52.92%
I	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	44.07%	0.27	20.56%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	43.09%	0.30	34.80%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	44.85%	0.31	34.98%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	**46.75%**	**0.34**	**36.20%**
S	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	61.35%	0.51	45.37%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	60.28%	0.51	51.28%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	**62.21%**	**0.53**	49.68%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	61.44%	0.52	**52.26%**
I + S	ResNet-18 + SPP-Net (late)	$\mathcal{L}_{ce}$	62.65%	0.52	45.41%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce}$	**70.23%**	**0.63**	**64.88%**
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc}$	70.00%	0.62	64.65%
	ResNet-18 + SPP-Net	$\mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{fc} + \lambda_2 \mathcal{L}_{tri}$	70.06%	0.62	64.79%

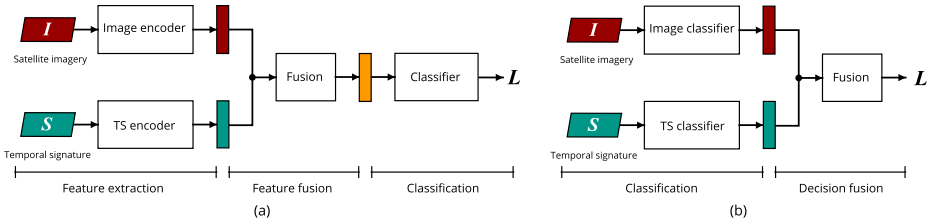

Fig. 11. Illustration of the workflow of (a) feature-level (early) fusion, (b) decision-level (late) fusion.

Table 11
Comparison of testing results with different fusion methods. The best results are highlighted in bold.

Method	Fusion method	Metrics		
		Accuracy	Kappa	Avg. F1
ResNet-18 + LSTM-Net	concat	**69.45%**	**0.64**	**67.97%**
	sum	68.88%	0.63	67.36%
	max	68.63%	0.63	66.88%
ResNet-18 + SPP-Net	concat	69.85%	0.64	68.59%
	sum	70.15%	0.65	**68.93%**
	max	**70.20%**	**0.65**	68.45%

mechanisms, they can both capture temporal dependencies, and therefore are suitable for time-series data classification. Moreover, with the help of graphics processing unit (GPU), the training time costs of the neural networks are significantly lower than traditional classifiers like SVM, which demonstrates the efficiency of the proposed networks.

6.2. Comparison between feature and decision-level fusion

The proposed feature-level (early) fusion and baseline decision-level (late) fusion methods are presented in Fig. 11a and 11b, respectively. As we can see from Tables 3 and 4, the classification results of late fusion are significantly lower than that of early fusion, with more than 6% drop in validation and testing accuracy on both the URFC-A and URFC-B datasets. This may be due to the reason that the temporal signatures are much more distinguishable in region function classification than the satellite images for the experimental datasets.

The late fusion is conducted after separately training of the two uni-modal classifiers, of which the one trained on temporal signatures is more powerful than that trained on satellite images; however, the fusion averages the accuracy of the two classifiers. While early fusion is conducted in the feature-level, which means that the features extracted from two different data sources are fused before the training of the final classifier. This indicates that the classifier is able to find a way to maximize useful information from both features during training. Therefore, the early fusion strategy is better than late fusion when there is significant performance gap between different modalities.

Nevertheless, the late-fusion method also has its advantage. Compared with early fusion, the late-fusion method is much easier to interpret, since the prediction scores of uni-modal classifiers can be extracted before decision fusion easily, and thus the contribution of different input data can be measured directly.

6.3. Comparison of different fusion methods

In our experiments, we compare the results of using different fusion methods, i.e. concatenation, element-wise sum, and element-wise max pooling, the experimental results of URFC-A are presented in Table 11. As can be seen, there is no significant difference in testing results between the three fusion methods, with almost all the variation of the resulting metrics less than 1%. This suggests that the proposed deep multi-modal fusion network is insensitive to the choice of fusion methods for the experimental datasets.

7. Conclusions

Remote and social sensing data are complementary to each other as they possess their own unique characteristics. The integration of them has the potential to improve the accuracy of urban region function recognition. The key challenge is effective fusion of the two kinds of data. In this paper, we propose an end-to-end deep multi-modal fusion network to effectively fuse satellite imagery and social sensing signature data. The two data sources are put into modal-specific encoders of residual CNN and our proposed 1-dimensional CNN/LSTM-based network respectively to extract features, which are further fused and then fed into fully connected layers and softmax layer to make final predictions. The two proposed 1-dimensional neural networks can extract discriminative features from temporal signatures which explicitly take temporal dependencies into account. To address the asynchronous problem of remote and social sensing data, we propose two auxiliary losses (cross-modal feature consistency loss and cross-modal triplet loss) to make the trained network more robust to missing modalities. We have conducted extensive experiments on publicly available datasets, and the results demonstrate the effectiveness and efficiency of our proposed methods. The paper has shown the power of deep learning in integrating remote and social sensing data for urban region function recognition, and thus providing an effective way for related urban studies. In the future, we plan to pay more attention to the interpretation of the black-box results, and further apply the method to more real-world scenarios.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The author acknowledges the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, and the University of Nottingham. This work was also supported by the UK Engineering and Physical Sciences Research Council [Grant No. EP/L015463/1], the National Natural Science Foundation of China (No. 41871329, 71961137003), the Shenzhen Scientific Research and Development Funding Program (No. JCYJ20170818092931604, JCYJ20180305125113883).

References

Albert, A., Kaur, J., Gonzalez, M.C., 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017, pp. 1357–1366. https://doi.org/10.1145/3097983.3098070.

Audebert, N., Saux, B.L., Lefèvre, S., 2018. Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks. ISPRS J. Photogramm. Remote Sens. 140, 20–32. https://doi.org/10.1016/j.isprsjprs.2017.11.011.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis – towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014.

Cao, R., Qiu, G., 2018. Urban land use classification based on aerial and ground images.

In: Proceedings of the 16th International Conference on Content-Based Multimedia Indexing, CBMI 2018, La Rochelle, France, September 4–6, pp. 1–6. https://doi.org/10.1109/CBMI.2018.8516552.
Cao, J., Tu, W., Li, Q., Zhou, M., Cao, R., 2015. Exploring the distribution and dynamics of functional regions using mobile phone data and social media data. In: Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management, Boston, MA, USA, July 10, 2015, Boston, MA, USA, pp. 264:1–264:16.
Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q., Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. Remote Sens. 10 (10), 1553. https://doi.org/10.3390/rs10101553.
Cao, R., Zhang, Q., Zhu, J., Li, Q., Li, Q., Liu, B., Qiu, G., 2020. Enhancing remote sensing image retrieval using a triplet deep metric learning network. Int. J. Remote Sens. 41 (2), 740–751. https://doi.org/10.1080/2150704X.2019.1647368.
Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., Yang, Z., 2018. Social functional mapping of urban green space using remote sensing and social sensing data. ISPRS J. Photogramm. Remote Sens. 146, 436–452. https://doi.org/10.1016/j.isprsjprs.2018.10.010.
Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS J. Photogramm. Remote Sens. 117, 11–28. https://doi.org/10.1016/j.isprsjprs.2016.03.014.
Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE 105 (10), 1865–1883.
Chi, M., Sun, Z., Qin, Y., Shen, J., Benediktsson, J.A., 2017. A novel methodology to label urban remote sensing images based on location-based social media photos. Proc. IEEE 105 (10), 1926–1936. https://doi.org/10.1109/JPROC.2017.2730585.
Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, Florida, USA, June 20–25, 2009, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
Deng, X., Liu, P., Liu, X., Wang, R., Zhang, Y., He, J., Yao, Y., 2019. Geospatial big data: new paradigm of remote sensing applications. IEEE J. Sel. Top. Appl. Earth Obser. Remote Sens. 12 (10), 3841–3851. https://doi.org/10.1109/JSTARS.2019.2944952.
Du, Z., Zhang, X., Li, W., Zhang, F., Liu, R., 2019. A multi-modal transportation data-driven approach to identify urban functional zones: An exploration based on Hangzhou City, China. Trans. GIS. https://doi.org/10.1111/tgis.12591.
Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.-A., 2019. Deep learning for time series classification: a review. Data Min. Knowl. Disc. 33 (4), 917–963. https://doi.org/10.1007/s10618-019-00619-1.
Feng, T., Truong, Q.-T., Thanh Nguyen, D., Yu Koh, J., Yu, L.-F., Binder, A., Yeung, S.-K., 2018. Urban zoning using higher-order markov random fields on multi-view imagery data. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 614–630.
Gao, S., Janowicz, K., Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans. GIS 21 (3). https://doi.org/10.1111/tgis.12289.
Gao, Q., Fu, J., Yu, Y., Tang, X., 2019. Identification of urban regions' functions in Chengdu, China, based on vehicle trajectory data. PLOS One 14 (4), e0215656. https://doi.org/10.1371/journal.pone.0215656.
Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P.M., Benediktsson, J.A., 2019. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. IEEE Geosci. Remote Sens. Mag. 7 (1), 6–39. https://doi.org/10.1109/MGRS.2018.2890023.
He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 27-30, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
Hoffmann, E.J., Wang, Y., Werner, M., Kang, J., Zhu, X.X., 2019. Model fusion for building type classification from aerial and street view images. Remote Sens. 11 (11), 1259. https://doi.org/10.3390/rs11111259.
Hu, T., Yang, J., Li, X., Gong, P., 2016. Mapping urban land use by using landsat images and open social data. Remote Sens. 8 (2), 151. https://doi.org/10.3390/rs8020151.
Jendryke, M., Balz, T., McClure, S.C., Liao, M., 2017. Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai. Comput. Environ. Urban Syst. 62, 99–112. https://doi.org/10.1016/j.compenvurbsys.2016.10.004.
Jia, Y., Ge, Y., Ling, F., Guo, X., Wang, J., Wang, L., Chen, Y., Li, X., 2018. Urban land use mapping by combining remote sensing imagery and mobile phone positioning data. Remote Sens. 10 (3), 446. https://doi.org/10.3390/rs10030446.
Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. (in press). https://doi.org/10.1016/j.isprsjprs.2018.02.006.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.
Lefèvre, S., Tuia, D., Wegner, J.D., Produit, T., Nassaar, A.S., 2017. Toward seamless multiview scene analysis from satellite to street level. Proc. IEEE 105 (10), 1884–1899. https://doi.org/10.1109/JPROC.2017.2684300.
Leung, D., Newsam, S., 2012. Exploring geotagged images for land-use classification. In: Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and Its Applications in Multimedia. ACM, pp. 3–8. https://doi.org/10.1145/2390790.2390794.
Li, X., Zhang, C., Li, W., 2017. Building block level urban land-use information retrieval based on Google Street View images. GIScience & Remote Sens. 54 (6), 819–835. https://doi.org/10.1080/15481603.2017.1338389.
Li, J., Benediktsson, J.A., Zhang, B., Yang, T., Plaza, A., 2017. Spatial technology and social media in remote sensing: a survey. Proc. IEEE 105 (10), 1855–1864. https://doi.org/10.1109/JPROC.2017.2729890.
Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep learning for hyperspectral image classification: an overview. IEEE Trans. Geosci. Remote Sens. 57 (9), 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932.
Liu, Y., Wang, F., Xiao, Y., Gao, S., 2012. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape Urban Plann. 106 (1), 73–87. https://doi.org/10.1016/j.landurbplan.2012.02.012.
Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: a new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105 (3), 512–530. https://doi.org/10.1080/00045608.2015.1018773.
Liu, X., He, J., Yao, Y., Zhang, J., Liang, H., Wang, H., Hong, Y., 2017. Classifying urban land use by integrating remote sensing and social media data. Int. J. Geograph. Informat. Sci. 31 (8), 1675–1696. https://doi.org/10.1080/13658816.2017.1324976.
Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.
Pan, G., Qi, G., Wu, Z., Zhang, D., Li, S., 2013. Land-use classification using taxi GPS traces. IEEE Trans. Intell. Transp. Syst. 14 (1), 113–123. https://doi.org/10.1109/TITS.2012.2209201.
Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Informat. Process. Syst. 32, 8024–8035.
Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., Zhou, C., 2014. A new insight into land use classification based on aggregated mobile phone data. Int. J. Geograph. Informat. Sci. 28 (9), 1988–2007. https://doi.org/10.1080/13658816.2014.913794.
Qi, L., Li, J., Wang, Y., Gao, X., 2019. Urban observation: integration of remote sensing and social media data. IEEE J. Sel. Top. Appl. Earth Obser. Remote Sens. 12 (11), 4252–4264. https://doi.org/10.1109/JSTARS.2019.2908515.
Srivastava, S., Vargas Muñoz, J.E., Lobry, S., Tuia, D., 2018. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. Int. J. Geograph. Informat. Sci. 1–20.
Srivastava, S., Vargas-Muñoz, J.E., Tuia, D., 2019. Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. Remote Sens. Environ. 228, 129–143. https://doi.org/10.1016/j.rse.2019.04.014.
Tu, W., Cao, J., Yue, Y., Shaw, S.-L., Zhou, M., Wang, Z., Chang, X., Xu, Y., Li, Q., 2017. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. Int. J. Geograph. Informat. Sci. 31 (12), 2331–2358. https://doi.org/10.1080/13658816.2017.1356464.
Tu, W., Hu, Z., Li, L., Cao, J., Jiang, J., Li, Q., Li, Q., 2018. Portraying urban functional zones by coupling remote sensing imagery and human sensing data. Remote Sens. 10 (1), 141. https://doi.org/10.3390/rs10010141.
Tu, W., Cao, R., Yue, Y., Zhou, B., Li, Q., Li, Q., 2018. Spatial variations in urban public ridership derived from GPS trajectories and smart card data. J. Transp. Geogr. 69, 45–57. https://doi.org/10.1016/j.jtrangeo.2018.04.013.
Tu, W., Zhu, T., Xia, J., Zhou, Y., Lai, Y., Jiang, J., Li, Q., 2019. Portraying the spatial dynamics of urban vibrancy using multisource urban big data. Comput., Environ. Urban Syst. 101428.
Workman, S., Zhai, M., Crandall, D.J., Jacobs, N., 2017. A unified model for near and remote sensing. In: Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, October 22–29, 2017, pp. 2707–2716. https://doi.org/10.1109/ICCV.2017.293.
Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2vec model. Int. J. Geograph. Informat. Sci. 31 (4), 825–848. https://doi.org/10.1080/13658816.2016.1244608.
Yuyun, Akhmad Nuzir, F., Julien Dewancker, B., 2017. Dynamic land-use map based on twitter data. Sustainability 9 (12), 2158. https://doi.org/10.3390/su9122158.
Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4 (2), 22–40. https://doi.org/10.1109/MGRS.2016.2540798.
Zhang, X., Du, S., Wang, Q., 2017. Hierarchical semantic cognition for urban functional zones with VHR satellite images and POI data. ISPRS J. Photogramm. Remote Sens. 132, 170–184. https://doi.org/10.1016/j.isprsjprs.2017.09.007.
Zhang, W., Li, W., Zhang, C., Hanink, D.M., Li, X., Wang, W., 2017. Parcel-based urban land use classification in megacity using airborne LiDAR, high resolution orthoimagery, and Google Street View. Comput. Environ. Urban Syst. 64, 215–228. https://doi.org/10.1016/j.compenvurbsys.2017.03.001.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70. https://doi.org/10.1016/j.rse.2018.06.034.
Zhang, Y., Li, Q., Tu, W., Mai, K., Yao, Y., Chen, Y., 2019. Functional urban land use recognition integrating multi-source geospatial data and cross-correlations. Comput. Environ. Urban Syst. 78, 101374. https://doi.org/10.1016/j.compenvurbsys.2019.101374.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019. Joint deep learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187. https://doi.org/10.1016/j.rse.2018.11.014.

Zhang, F., Wu, L., Zhu, D., Liu, Y., 2019. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogramm. Remote Sens. 153, 48–58. https://doi.org/10.1016/j.isprsjprs.2019.04.017.

Zhu, Y., Newsam, S., 2015. Land use classification using convolutional neural networks applied to ground-level images. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, Washington, USA, November 3–6, 2015, ACM, New York, NY, USA, pp. 61:1–61:4. https://doi.org/10.1145/2820783.2820851.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5 (4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307.

Zhu, Y., Deng, X., Newsam, S., 2019. Fine-grained land use classification at the city scale using ground-level images. IEEE Trans. Multimedia 21 (7), 1825–1838. https://doi.org/10.1109/TMM.2019.2891999.