

Received October 9, 2019, accepted November 1, 2019, date of publication November 21, 2019, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2954851

# A Latent Feature-Based Multimodality Fusion Method for Theme Classification on Web Map Service

ZELONG YANG<sup>1</sup>, ZHIPENG GUI<sup>1,2,3</sup>, HUAYI WU<sup>1,3</sup>, AND WENWEN LI<sup>4</sup>

<sup>1</sup>State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

<sup>2</sup>School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

<sup>3</sup>Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

<sup>4</sup>School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287-5302, USA

Corresponding author: Huayi Wu (wuhuayi@whu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 41930107 and Grant 41971349.

**ABSTRACT** Massive maps have been shared as Web Map Service (WMS) from various providers, which could be used to facilitate people's daily lives and support space analysis and management. The theme classification of maps could help users efficiently find maps and support theme-related applications. Traditionally, metadata is usually used in analyzing maps content, few papers use maps, especially legends. In fact, people usually considers metadata, maps and legends together to understand what maps tell, however, no study has tried to exploit how to combine them. This paper proposes a method to fuse them with the purpose of classifying map themes, named latent feature based multimodality fusion for theme classification (LFMF-TC). Firstly, a multimodal dataset is created that supports the supervised classification on map themes. Secondly, textual and visual features are designed for metadata, maps, and legends using some advanced techniques. Thirdly, a latent feature based fusion method is proposed to fuse the multimodal features on the feature level. Finally, a neural network classifier is implemented using supervised learning on the multimodal dataset. In addition, a web-based collaboration platform is developed to facilitate users in labeling multimodal samples through an interactive Graphical User Interface (GUI). Extensive experiments are designed and implemented, whose results prove that LFMF-TC could significantly improve the classification accuracy. In theory, the LFMF-TC could be used for other applications with few modifications.

**INDEX TERMS** Cartography, machine learning, multimodality fusion, theme classification, web map service.

## I. INTRODUCTION

Maps enable people to intuitively sense geospatial entities' morphology, distribution and spatial relationships by visualizing them using some creative efforts (e.g. symbolization, generalization), which could serve for people's daily lives, space analysis and management. Web Map Service (WMS), as a popular standard sharing geospatial data as interoperable maps, has been adopted and implemented by many software and authorities [1]. There have been large number of WMS maps shared online across many disciplines [2], [3]. The theme classification of maps could help people efficiently find maps of their interest [3], [4], analyze hot map

themes and their changes [2], adaptively monitor the service quality [5], [6] and compose service chains [7].

Features extraction plays a key role in the theme classification. In the WMS, a map layer consists of metadata, map and legend, and each of them contains some information related to the map theme. Most studies have exploited in extracting textual features from the metadata to describe maps using some natural language process (NLP) techniques [2], [3]. However, few papers tried to understand maps using maps or legends due to the semantic gap between low-level visual features and themes. Recently, as the deep learning (DL) develops, especially the convolutional neural network (CNN), some salient and hierarchical visual features could be automatically learnt from raw images [8]–[10], which have motivated researchers to exploit them in various applications. Besides, the development of optical character recognition (OCR)

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao.

empowers people in extracting accurate textual information from images. All of them could be used to extract more useful features from maps and legends for classifying map themes.

Recently, information fusion also has shown promising performance in lots of applications, which can use multi-source information to compensate for the limitations with single source [11], [12]. Especially, the multimodality fusion can integrate heterogeneous features due to multimodal sources or various feature extractors [13], [14]. Many studies have illustrated its advantages compared to the unimodal fusion [15]–[17]. However, to the best of our knowledge, only few studies tried to fuse the metadata and maps on map analysis, and all of them firstly processed metadata and maps separately, which lost many information before the final fusion.

In this paper, a latent feature based multimodality fusion method for theme classification (LFMF-TC) is proposed, which works with multimodal features from metadata, maps and legends to classify map themes. The contributions of this work are: (1) creating a multimodal dataset for map themes classification and developing an interactive web-based platform to facilitate it; (2) designing textual and visual features for metadata, maps and legends using some NLP, CNN and OCR techniques; (3) proposing and implementing a latent feature based fusion method to integrate multimodal features; (4) designing a series of experiments and applying an in-depth analysis of their results to investigate the effectiveness of LFMF-TC.

The rest of the paper is organized as follows: section 2 presents a review of the available papers related to the map classification. Section 3 describes the architecture and some implementation details of the LFMF-TC. Section 4 presents a series of experiments which were conducted to analyze the effectiveness and configuration of the LFMF-TC. Section 5 furtherly discusses how LFMF-TC takes effect using confusion matrixes and samples. Section 6 summarizes our findings and discusses future research avenues.

## II. LITERATURE REVIEW

In the WMS, some other materials are also published with maps to help users understand and use maps, such as metadata, legends. Based on the material they used, existing map classification methods could be generally divided into two main categories: map-based classification and metadata-based classification.

Map-based classification studies involved the use of visual features from maps to classify map types like themes. Numerous visual features have been proposed to describe maps' visual appearance or infer their semantics, which achieved successful results. For example, [18] extracted distributed color histograms to measure users' preference on the map appearance. Particularly, to handle the semantic gap between visual features and themes, most studies tried to use the cartography standards or summarize some historical conventions on specific entities. Reference [19] summarized some heuristics for roads and rivers using the shape compactness and

color to extract roads and rivers from maps automatically. The development of deep learning, especially deep convolutional neural networks (DCNNs), opens up new research opportunities for image processing, because they can automatically learn salient visual features from raw images [20]. It has motivated researchers to explore its application for map-based classification. Reference [21] distinguished maps from other images using the ResNet-50. Reference [10] compared several popular DCNNs (e.g. AlexNet, VGG Net, ResNet, Inception, Inception-ResNet) in classifying map types defined by themselves. However, with all their benefits, there exist some limitations in map-based classification. Because they can only handle those maps with significant visual features. For those maps with ambiguous symbols, it needs some other descriptive information to infer their themes even for human.

Metadata, as one complementary descriptive information for maps in WMS, gives textual description about the map content and some other attributes. Compared to maps, there exists no semantic gap in metadata. Moreover, many applications related to maps are directly based on text, such as keyword-based map retrieval. Therefore, many researchers have put their efforts in classifying maps based on metadata. Reference [2] extracted keywords from metadata to conduct a topic survey on their crawled maps. Reference [18] matched users' queries with metadata to search for maps of users' interest. Furthermore, [22] used TFIDF to weight extracted words from metadata in classifying web services. Reference [3] adopted GCMD and SWEET ontology to increase the matching accuracy in classifying maps. However, the textual metadata usually lacks descriptions or has limited descriptive capability on maps' visual features. In addition, some metadata fields are left empty or unrelated to maps.

As stated above, neither maps nor metadata alone can handle varying conditions well, but maps and metadata could offer complementary information for each other. Hence, by fusing data from two complementary sources, the performance of map classification could be improved. Some existing studies utilized the fusion of them to enhance their applications, and their results revealed significant improvement. For example, [18] used maps' similarity on appearance to refine metadata based search results. In addition, legends, as an interpreter for map symbols, tell what are included in the map, which could also provide complementary information for the theme classification. But, to our best knowledge, no study has explored in using legends to classify maps. Therefore, in this research work, we proposed to fuse maps, metadata and legends to classify map themes.

There exist some challenges in fusing maps, metadata and legends, because they are organized using multiple modalities (e.g. image, text), and heterogeneous features will be extracted from them. Recently, many multimodality fusion strategies have been proposed in various applications [16], [17], [23]–[25]. Their fusion operations are mainly performed at the feature-level and decision-level. However, in the map classification, existing studies mostly fuse maps and metadata at the decision-level to avoid handling heterogeneous

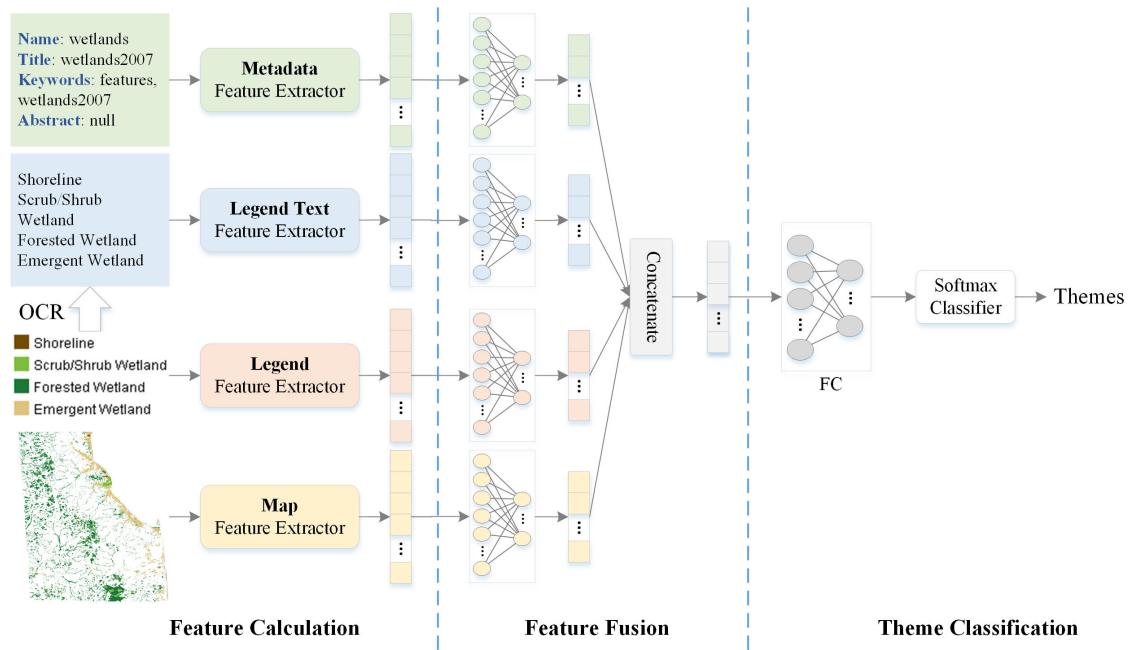


FIGURE 1. Architecture of LFMF-TC.

features together. They firstly conducted the classification on maps and metadata separately, then some rules were used to fuse the stand-alone classification results [18]. The major drawback of them is the stand-alone result to each individual information source. Firstly, it requires to train many classifiers for each information source and needs another classifier or rules to fuse all results. Moreover, the acquisition of concurrent features from all sources may be necessary to collect sufficient information to make an improved classification result, but stand-alone results will lose some information compared to the raw features. In contrast, the feature-level fusion could collect concurrent features and integrate them to provide sufficient information for making an improved decision. Therefore, in this study, we proposed a latent feature based multimodality fusion method to fuse maps, metadata and legends on the feature-level.

### III. METHODOLOGY

In this section, we firstly describe the architecture of LFMF-TC. Then, some details are described on how to calculate the multimodal features and fuse them.

#### A. ARCHITECTURE

The architecture of LFMF-TC is presented in FIGURE 1, which mainly consists of three functional parts: feature calculation, feature fusion and theme classification. The feature calculation aims to extract multimodal features from metadata, maps, legends and legend text, where the legend text is recognized from legends using the Google Vision API and manual rectifications. The details about four feature extractors are described in the III-B. Once multimodal features extracted, the feature fusion module fuses them into a single

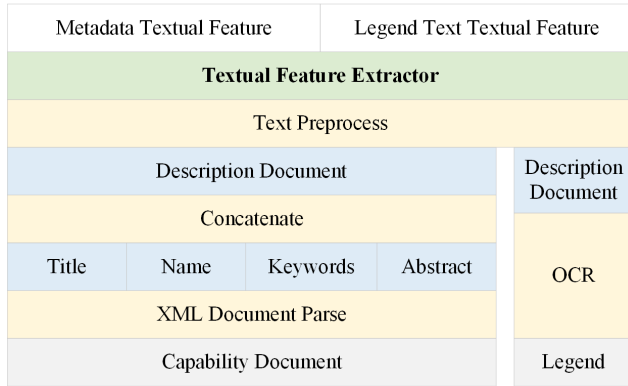
feature vector. In this module, a latent feature based fusion method is proposed to handle their heterogeneous structures, whose details are described in the III-C. Finally, the theme classification module maps the fused feature vectors to pre-defined themes. It contains of a fully connected (fc) layer and a softmax classifier. They work together to generate a probability distribution for each map over all themes. To measure and train the overall framework, the categorical cross entropy loss function is chosen, which will give a high penalty when the predicted theme diverges from its ground truth.

#### B. FEATURE CALCULATION

This section describes how we extract features from multimodal information sources. Based on their modalities, the methods are organized as two parts: textual features (i.e. metadata, legend text) and visual features (i.e. map, legend).

##### 1) TEXTUAL FEATURES

The textual description about a map comes from the metadata and legend text. Metadata contains many fields which describe the map from many perspectives, such as providers, spatial range, content, temporal information. The legend text mainly tells users what symbols represent in the map, which could provide more detailed information on entity types compared to the metadata. As noted in the literature review, some methods have been developed to extract and weight keywords from the metadata. In this study, we only want to testify the effectiveness of fusing multimodal information in classifying map themes. Therefore, two typical word weighting schemas are selected: term frequency (TF) and term frequency inversed document frequency (TF-IDF).



**FIGURE 2.** Structures of the textual feature extractor on metadata and legend text.

Their algorithms are shown as the equation (1) and (2), where  $tf_{t,d}$  means the frequency of the  $t$ -th word in a document;  $n$  means the total number of documents;  $n_{t,d}$  means the number of documents containing the  $t$ -th word.

$$TF = tf_{t,d} \quad (1)$$

$$TF-IDF = tf_{t,d} \left( \log \frac{n}{n_{t,d}} + 1 \right) \quad (2)$$

FIGURE 2 shows how we calculate TF and TF-IDF for the metadata and legend text. In metadata, there exist four fields closely related to the map theme: title, name, keywords, and abstract, which are used to calculate the textual features. Firstly, a (Extensible Markup Language) XML document parser is developed to extract the content of four fields from the WMS capability document. Because the sentence structure is not a concern for both TF and TF-IDF, their contents are simply concatenated as a document named description document. To remove the non-meaningful words or punctuations which will affect the quality of textual features, some text preprocess techniques are applied to it like filtering (punctuation, stop words), lowercase, lemmatization. Finally, the metadata textual features (i.e. TF, TF-IDF) are calculated by the textual feature extractor using above equations. Same procedures are conducted in calculating textual features for the legend text, with an exception that the description

document is from the legend using the Google Vision API assisted with manual rectifications.

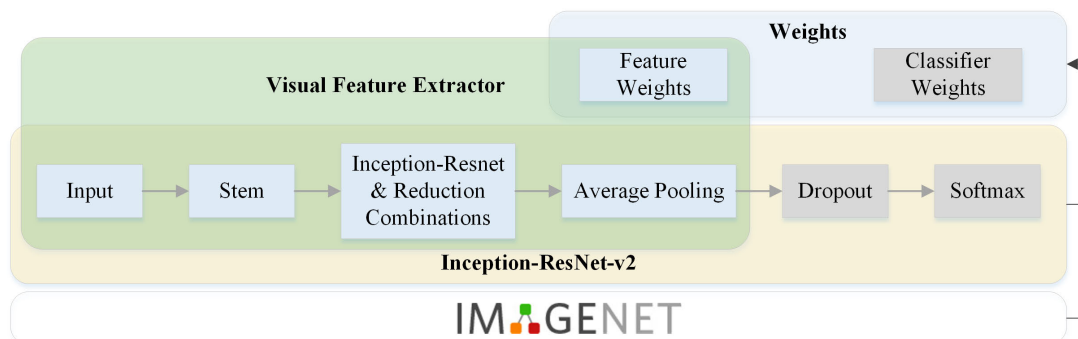
## 2) VISUAL FEATURES

Maps use symbols to represent the morphology and spatial distributions of geospatial entities. Legends consist of symbols and their descriptions, whose symbols and layout may be related to map themes. As noted in the literature review, DCNNs have become the popular candidate in the image processing. Inception-Resnet-v2, as a distinguish one among them, integrates two unique modules to enhance its capability: inception and residual [26]. The inception module can extract more details by using multiple convolution branches linked to the feature map, and the residual connection resolves the problem of gradient propagation through adding direct connections between the input and output [27]. But the Inception-Resnet-v2 is used for classification, a key aspect here is to identify a splitting point used to accomplish feature extraction for maps and legends. As shown in FIGURE 3, the output of the average pooling is selected as the feature map for maps and legends.

As you know, massive labelled images are required to train the Inception-Resnet-v2 because of millions of weight parameters to learn. Recently, transfer learning has been proved to be an effective method to apply DCNNs on small datasets, which can leverage the knowledge from other big datasets [28]. FIGURE 3 shows how we combine them to implement our visual feature extractors. Firstly, the Inception-Resnet-v2 network is trained on the ImageNet dataset which contains millions of annotated images [29], and there have been well trained model and parameters shared online. Then, based on the splitting point in the previous paragraph, the weight parameters could be divided into two groups, named feature weights and classifier weights. Finally, those modules and feature weights in the green rectangle (FIGURE 3) work together as the visual feature extractor for maps and legends in our study.

## C. LATENT FEATURE BASED MULTIMODALITY FUSION

After extracting above multimodal features, we need to fuse them for the further classification. As stated in the



**FIGURE 3.** Structures of the visual feature extractor.

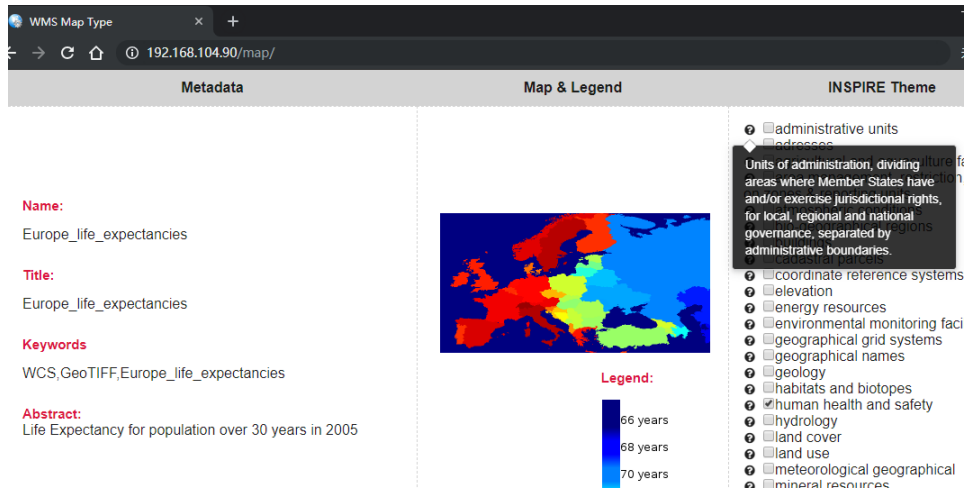


FIGURE 4. GUI for labelling multimodal map samples.

literature review, the feature-level fusion can leverage the sufficient information from all features to enhance the classification result. Therefore, in this study, a new high dimensional feature vector is constructed by concatenating above feature vectors. Although all textual and visual features have been represented as numeric vectors, their length and sparsity are significantly different due to heterogeneous sources and algorithms. For example, the visual features are with fixed length, while the length of textual features varies as samples change. Besides, the textual feature vectors are high dimensional and sparse, because both metadata and legend text are usually short and organized with free description from various publishers. On the one hand, the significant difference on their feature length will degrade the fusion purpose, in other words, the classifier maybe put more focus on those features represented as longer vectors [25]. On the other hand, the feature sparsity will lead to huge computation cost. In this paper, latent features are constructed to keep important information from raw features and decrease their sparsity. In this study, as a compromise on avoiding intensive manual work in exploiting the ideal size for different features, a fixed size is used for all latent features. The neural network with one hidden layer is used to construct latent features. It can not only deal with raw features with variant length, and it can also work together with the classification module to optimize latent features.

#### IV. EXPERIMENTS

##### A. DATASET AND THEMES

The theme schema comes from the Infrastructure for Spatial Information in Europe (INSPIRE) directive [30]. It consists of 34 fine-granule themes which are grouped into 9 clusters. In this paper, the coarse theme clusters were used in experiments, because it is a labor intensive work to manually label multimodal samples for map themes. Besides, to reduce the difficulty, we restricted a map only related to a most related theme.

TABLE 1. The number of samples in each theme cluster.

ID	Theme Cluster Name	Number
1	Statistics and Health	145
2	Marine and Atmosphere	146
3	Earth Science	184
4	Land Use and Land Cover	111
5	Elevation, Ortho-imagery, Grids	210
6	Environmental Monitoring and Observations	110
7	Biodiversity and Management Areas	122
8	Facilities, Utilities and Public Services	124
9	Topo and Cadastre, Reference Data	325
Total number		1477

A web based Graphical User Interface (GUI) was developed to facilitate the multimodal samples labelling (FIGURE 4). It consists of three panels named metadata, map & legend, and INSPIRE theme. The metadata panel shows the content of selected metadata fields (name, title, keywords, and abstract). The map & legend panel presents the layer's map and legend. All of them were parsed and cached using the standard operations defined by WMS. In the theme panel, the fine-granule themes are listed, because they are easier for users to judge than the coarse-granule themes, and the latter could be easily got using the mapping relationship maintained in the INSPIRE directive. When labelling samples, users only need to check or uncheck corresponding themes for maps, and they can also view the theme definition during labelling by moving their mouse over the question icon. Table 1 summarizes the samples number. There exist at least 100 samples in each theme cluster. Although imbalanced samples could lead classifiers to predict samples as the type with most samples, it can only achieve up to 23% accuracy on our samples when no useful features used.

##### B. EXPERIMENTS AND ANALYSIS

A series of experiments were conducted to verify the effectiveness of LFMF-TC and investigate the impact of some parameters. Experiment 1 tests and analyzes the performance



of textual features in classifying map themes using several popular classifiers. Experiment 2 tests and analyzes the performance of visual features and their combination using the same classifiers. Experiment 3 demonstrates improvements made by our proposed LFMF-TC. Experiment 4 further investigates the impact of the latent feature size in LFMF-TC. To measure their performance, the k-fold cross validation was adopted, and its parameter  $k$  was set to five. It firstly divided labeled samples into five subsets, then samples in each subset were iteratively used for testing while the others for training the classifier. Therefore, five independent accuracy results could be calculated for each classifier. The average value of them was used as the overall accuracy to measure classifiers. Traditionally, samples are randomly divided into  $k$  subsets in the  $k$  fold cross validation. In this study, to guarantee the class balance and sample representativeness on the fine-granule themes across subsets, a hierarchical partition strategy was proposed in dividing samples with hierarchical themes. Experiment 5 investigates its advantages by comparison to the traditional random partition strategy.

### 1) TEXTUAL FEATURES BASED THEME CLASSIFICATION

In experiment 1, several popular supervised classification algorithms were used to analyze whether textual features were useful in classifying map themes, including K-NearestNeigobor (KNN), NaiveBayes, Decision Tree, Random Forest, Logistic Regression, Linear SVM, and Neural Network (NN). Except the NN, all others were from the Python Sklearn library. In experiments, the parameter  $k$  was set to 1 in the KNN after some initial experiments on our dataset. The NN classifier contained a hidden layer containing 64 neurons and a softmax classifier, and the RELU activation function and the categorical cross entropy loss function were used in training it. The NN classifiers were trained by running 500 iterations. All other parameters were set to their default value.

Their results are shown in FIGURE 5. Compared to the extreme situation stated in IV-A where no useful information contained in features, both metadata and legend text achieved significant improved accuracy than 23% across all selected classifiers. Hence, both of them contained some useful information for classifying map themes. In addition, the metadata achieved higher accuracy than the legend text universally, which means metadata contained more useful information. On the one hand, there were more maps with empty legend text than maps with empty metadata. Although the keywords and abstract in metadata are optional, the name and title are required for layers containing maps. But legends are optional, and part of them does not contain legend text. On the other hand, four fields in metadata usually contain more attributes and words than the legend context. About the feature type, no significant advantage was found on the TF or TF-IDF across classifiers or information sources, but the TF-IDF showed a little better than the TF in the NN classifier which achieved the best accuracy on both metadata and legend text.

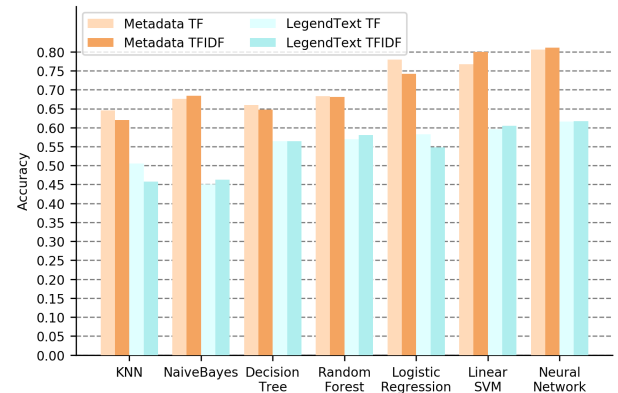


FIGURE 5. The classification accuracy based on textual features.

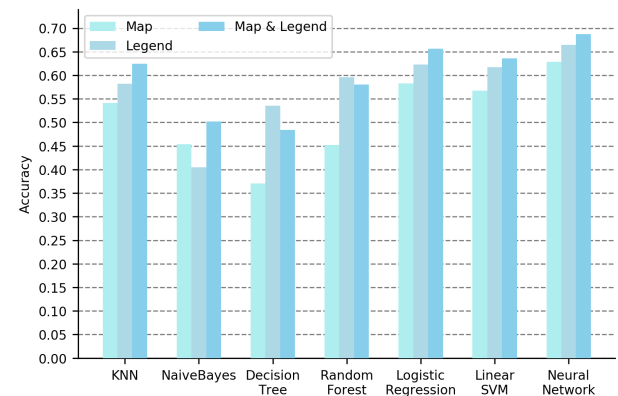


FIGURE 6. The classification accuracy based on visual features.

Therefore, the TF-IDF was used in our method for metadata and legend text.

### 2) VISUAL FEATURES BASED THEME CLASSIFICATION

In experiment 2, same classifiers were used to analyze the usefulness of visual features in classifying map themes. Beside the stand-alone map features and legend features, their combination was also tested by concatenating them into a single vector, because they were with the same length. FIGURE 6 shows their results. Although maps were used in nearly all existing studies to conduct visual analysis, it was surprising that legends achieved better performance with most classifiers in classifying map themes, whose reason will be investigated in our future research. In addition, intuitively, the combined feature contains more information and should be with a better performance. However, the legend feature outperformed it with the decision tree and random forest classifiers. The possible reason may be that these two classifiers did not make full use the information from features with their default parameters. It could be somehow proved by that they achieved significantly lower accuracy compared to other classifiers with higher accuracy. The lower accuracy of the naïveBayes classifier may be due to that it was hard to conduct statistical analysis on high-dimensional features with small number of samples. But all classifiers got improved accuracy than 23% with all visual features, which means

**TABLE 2.** The classification accuracy based on multimodality fusion.

Method	Accuracy
LFMF-TC (vision)	0.7387
Majority Voting	0.7508
Feature Concatenation	0.8368
LFMF-TC	0.8842

that all visual features contained some useful information for classifying map theme.

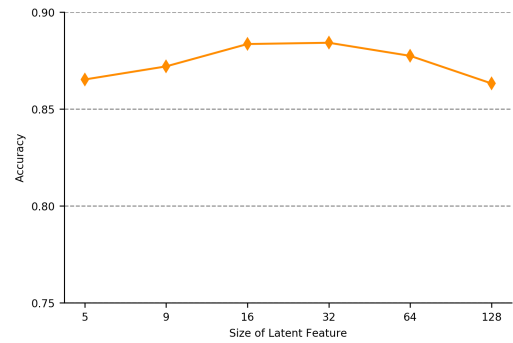
### 3) MULTIMODALITY FUSION BASED THEME CLASSIFICATION

Drawn from above analysis, all information sources (metadata, maps, legends) could provide some useful information for classifying map themes. Among them, the metadata yielded best accuracy (81.11%) with the NN classifier using its TF-IDF feature. In this below, it will be used to verify the effectiveness of our method.

In the previous section, it has shown promising improvement by fusing homogeneous and unimodal features from maps and legends. In this section, we firstly investigate the effectiveness of LFMF-TC by fusing all information (maps, legends, legend text) from visual sources named LFMF-TC (vision). At the same time, three typical fusion methods were selected to fuse metadata, maps and legends: majority voting, feature concatenation, and LFMF-TC. In the feature concatenation and LFMF-TC, a neural network classifier was used to classify their fused features, which shared the same structure as those in previous experiments. The size of latent features was set to 32 in LFMF-TC. Table 2 lists their results accuracy. The LFMF-TC (vision) could further improve the classification accuracy from 68.72% to 73.87% by fusing the legend text. By fusing all information sources, all of three classifiers achieved better accuracy than the LFMF-TC (vision). Among them, LFMF-TC achieved the best accuracy (88.42%) which was also higher than accuracies on any individual information source or fusing part of them. In comparison, the majority voting got a lower accuracy even than the metadata TF-IDF based NN, because it lost much information before fusing decisions. Although the feature concatenation conserved more information than the LFMF-TC, the variant length of multimodal features and limited training iterations degraded its fusion performance.

### 4) EFFECT OF SIZE OF LATENT FEATURE SET TO CLASSIFICATION ACCURACY

The influence of latent feature size could be analyzed from two perspectives: information volume and computation cost. A latent feature with bigger size could conserve more information for further analysis, but it also requires huger computation cost to make full use of those information. In this experiment, we changed the latent feature size in the LFMF-TC to investigate its effect on the classification accuracy. All classifiers were trained by running 500 iterations. Their results are shown in FIGURE 7. As the size increased, the result accuracy firstly increased then

**FIGURE 7.** The classification accuracy of the LFMF-TC with different latent feature size.

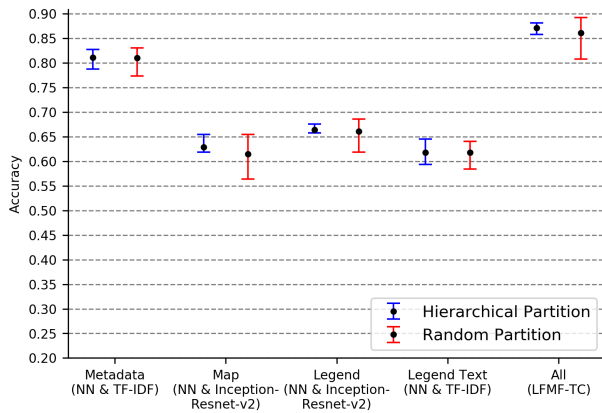
decreased. As stated above, a latent feature with small size only conserved little information, which will limit the performance improvement. As the size increases, more information could be conserved to improve the accuracy, therefore, the result accuracy firstly increased. However, bigger size requires huger computation cost to make full use of conserved information. Therefore, the result accuracy went down after a size (32) due to inadequate training with limited iterations.

### 5) INFLUENCE OF PARTITION STRATEGIES

In this experiment, we investigated the effect of partition strategies on the classification accuracy, where the random and hierarchical partition strategies were compared. Beside comparing two strategies on multimodal features using LFMF-TC, they were also compared on the single information source. Therefore, other four classifiers were selected, which achieved the best accuracy on each single information source. The five classifiers were named Metadata (NN & TF-IDF), Map (NN & Inception-Resnet-v2), Legend (NN & Inception-Resnet-v2), Legend Text (NN & TF-IDF) and All (LFMF-TC), where NN means the neural network classifier. All of them were repeated on the subsets divided using the traditional random partition strategy. In experiments, beside the average accuracy, we also recorded five accuracies taking each subset for validation. FIGURE 8 shows their results, where dot points mean the average accuracy and cap lines mean the minimum and maximum accuracy. Results showed that the hierarchical partition strategy cannot only achieved slightly higher average accuracy, it was also with smaller variance of accuracies across all classifiers. Because the hierarchical partition strategy guaranteed the class balance and samples representativeness across its subsets.

## V. DISCUSSION

The effectiveness of LFMF-TC has been verified by above experiments, however, how it works is still somehow confusing. In this section, we try to analyze its mechanism from more perspectives using confusion matrixes and some example results.



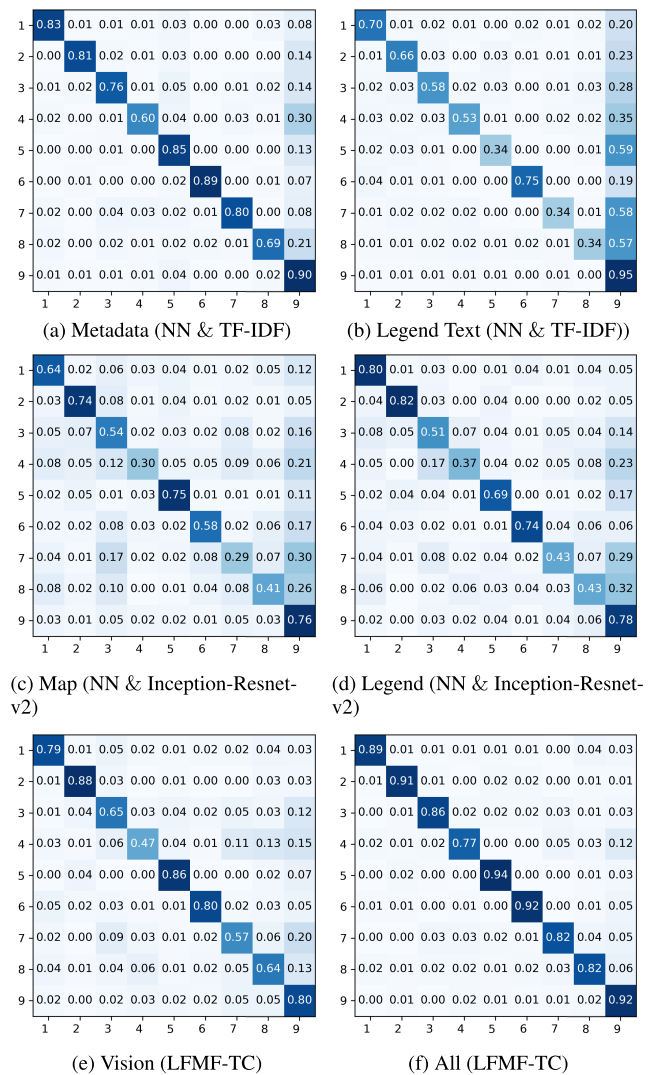
**FIGURE 8.** The classification accuracy on different sources with hierarchical and random partition strategies (NN represents the neural network classifier).

### A. CONFUSION MATRIXES

As the name suggests, the confusion matrix (CM) could help people to see if a model confuses any two classes, where the class level accuracy and predictions' distribution are presented. In FIGURE 9, each row of them represents maps in an actual theme cluster while each column represents maps in a predicted theme cluster. The theme cluster for each number could be found in Table 1. Generally, most misclassified maps accumulated in the ninth theme cluster across all themes, especially in the single source based CMs. Because our samples were unbalanced, and the ninth theme cluster contained largest number of samples. When insufficient information found in source data, the classifier will predict a map as the ninth theme cluster, which was proved by feeding some classifiers with empty features. The single source data usually contained less information compared to their fusion, therefore, the accumulation phenomenon was more serious on them. Particularly, there existed deepest colors on the ninth theme cluster in the legend text CM, because there were many maps with empty legend text. In addition, there existed more error predictions outside the ninth theme cluster in CMs on maps and legends. Although the legend text shared similar average accuracy (around 65%) as maps and legends. Although the legend text shared similar average accuracy (around 65%) as maps and legends, few visual features for maps and legends are empty. As more information embraced in the vision (LFMF-TC) and All (LFMF-TC), more useful information could be extracted. As a result, the accumulation phenomena became weaker. Specially, the fourth theme cluster (land use and land cover) got a relative low accuracy across all classifiers. The possible reasons include insufficient description in metadata and arbitrary cartography units and colors in their maps and legends.

### B. EXAMPLES RESULTS

In this section, some misclassified examples were presented in Table 3. For the maps with id 1 and 2, no useful information could be extracted from their metadata even for human. As a result, the metadata based classifier predicted them as the

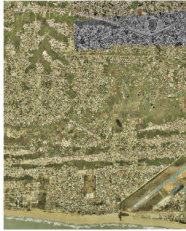

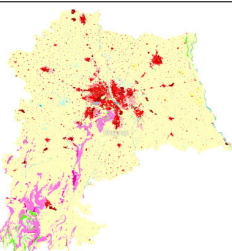
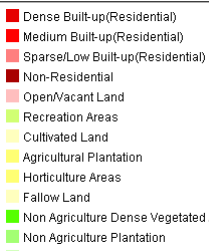


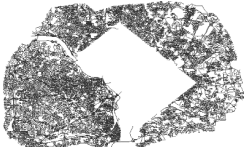


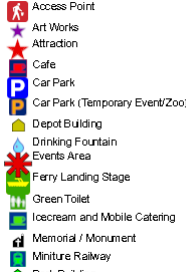


**FIGURE 9.** Confusion matrixes for different sources (NN represents the neural network classifier; rows represent ground truth and columns represent predictions).

ninth theme cluster. But the first map contained significant visual feature as remote sensing images, and the legend text in the second map contained useful information. All of them could be used by the LFMF-TC to make a true prediction. In the third map, California, as a place name, was too general for identifying its theme. In contrast, the legend text was more clear. However, the LFMF-TC also faced some challenges in fusing different information sources. For example, both 4 and 5 contained meaningful words in their metadata, such as streets and user facilities. As a result, the metadata based classifier could make true predictions for them. However, the LFMF-TC misclassified them, because it may be influenced by the visual features. For example, the fourth map looked like contours, and the fifth map legend was similar to land cover and land use maps. Although the legend text in the fifth map could be used to rectify errors, the samples



**TABLE 3.** Example classification results using LFMF-TC.

ID	Metadata	Map	Legend	Theme
1	Name: psn:13c Title: 13c Abstract: Keywords: WCS, GeoTIFF, 13c			Labeled Theme: Elevation, orthoimagery, grids  Metadata based Prediction: Topo&cadastre, reference data  LFMF-TC Prediction: Elevation, orthoimagery, grids
2	Name: ncr:NCR_LU_12_I Title: NCR_LU_12_I Abstract: Keywords:			Labeled Theme: Land use and land cover  Metadata based Prediction: Topo&cadastre, reference data  LFMF-TC based Prediction: Land use and land cover
3	Name: California Title: California Abstract: California Keywords:			Labeled Theme: Topo&cadastre, reference data  Metadata based Prediction: Marine and atmosphere  LFMF-TC based Prediction: Topo&cadastre, reference data
4	Name: 11 Title: Regional Streets (Emergency) Abstract: Keywords:			Labeled Theme: Topo&cadastre, reference data  Metadata based Prediction: Topo&cadastre, reference data  LFMF-TC based Prediction: Elevation, orthoimagery, grids
5	Name: 13 Title: Park User Facilities Abstract: Keywords:			Labeled Theme: Facilities, utilities and public services  Metadata based Prediction: Facilities, utilities and public services  LFMF-TC based Prediction: Land use and land cover

representativeness in the training set may be insufficient to capture it.

## VI. CONCLUSION AND FUTURE RESEARCH

This paper introduces a novel multimodality fusion method to investigate how to fuse metadata, maps and legends together in classifying map themes. In this paper, we addressed challenges in: 1) the lack of labeled multimodal samples for maps; 2) difficulty in capturing discriminative features from metadata, maps and legends; 3) the lack of studies trying to fuse metadata, maps and legends on the feature level to infer their themes; 4) the lack of thorough understanding of how

multimodality fusion works in classifying map themes. The success of LFMF-TC is summarized by integrating following strategies: 1) creating a multimodal sample dataset for maps using a web-based collaborative platform; 2) designing feature extractors for each map information source using some NLP, CNN and OCR techniques; 3) proposing and implementing a fusion method to fuse multimodal features and integrate their discriminative power; 4) designing a series of experiments and analyzing their fusion affects. To our best knowledge, this study represents one of the first attempts in understanding map contents by fusing metadata, maps, especially legends.

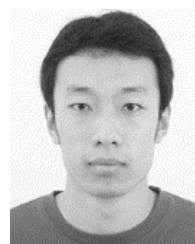
In the future, some features will be optimized by considering the characteristics of map attributes. For example, we plan to use some augmentation strategies to enlarge map samples and fine-tune the Inception-Resnet-v2 in extracting visual features. Besides, some correlation analysis methods could be used to investigate the relationship among features, whose results can be used to optimize feature extractors and the fusion method. In addition, the effect of imbalanced samples will be considered by labeling more examples or designing number-related loss function to measure the model performance.

## REFERENCES

- [1] J. de La Beaujardiere, *Opengis Web Map Server Implementation Specification*. Wayland, MA, USA: Open Geospatial Consortium Inc., 2006, pp. 6–42.
- [2] Z. Gui, J. Cao, X. Liu, X. Cheng, and H. Wu, “Global-scale resource survey and performance monitoring of public OGC Web map services,” *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 6, p. 88, Jun. 2016.
- [3] W. Li, “Lowering the barriers for accessing distributed geospatial big data to advance spatial data science: The PolarHub solution,” *Ann. Amer. Assoc. Geogr.*, vol. 108, no. 3, pp. 773–793, May 2018.
- [4] J. Gong, H. Wu, T. Zhang, Z. Gui, Z. Li, L. You, S. Shen, J. Zheng, J. Geng, K. Qi, W. Yang, Z. Li, and J. Yu, “Geospatial Service Web: Towards integrated cyberinfrastructure for GIScience,” *Geo-Spatial Inf. Sci.*, vol. 15, no. 2, pp. 73–84, Jun. 2012.
- [5] H. Zhang, M. Hu, and H. Wu, “QoGIS supported OWS framework extension,” *Sci. Surveying Mapping*, vol. 36, no. 4, pp. 148–150, Jul. 2011.
- [6] S. Shen, T. Zhang, H. Wu, and Z. Liu, “A catalogue service for Internet GIS services supporting active service evaluation and real-time quality monitoring,” *Trans. GIS*, vol. 16, no. 6, pp. 745–761, 2012.
- [7] H. Wu, Z. Li, H. Zhang, C. Yang, and S. Shen, “Monitoring and evaluating the quality of Web Map Service resources for optimizing map composition over the Internet to support decision making,” *Comput. Geosci.*, vol. 37, no. 4, pp. 485–494, Apr. 2011.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, p. 436, 2015.
- [9] W. Li and C.-Y. Hsu, “Automated terrain feature identification from remote sensing imagery: A deep learning approach,” *Int. J. Geograph. Inf. Sci.*, pp. 1–24, Nov. 2018.
- [10] X. Zhou, W. Li, S. T. Arundel, and J. Liu, “Deep convolutional neural networks for map-type classification,” 2018, *arXiv:1805.10402*. [Online]. Available: <https://arxiv.org/abs/1805.10402>
- [11] V. Torra and Y. Narukawa, “Introduction,” in *Modeling Decisions: Information Fusion and Aggregation Operators*. Berlin, Germany: Springer, 2007, pp. 1–17.
- [12] X. Jiang, F. Wu, Y. Zhang, S. Tang, W. Lu, and Y. Zhuang, “The classification of multi-modal data with hidden conditional random field,” *Pattern Recognit. Lett.*, vol. 51, pp. 63–69, Jan. 2015.
- [13] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, “Multi-modal fusion for multimedia analysis: A survey,” *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.
- [14] Y. Zheng, “Methodologies for cross-domain data fusion: An overview,” *IEEE Trans. Big Data*, vol. 1, no. 1, pp. 16–34, Mar. 2015.
- [15] C. Wang, H. Yang, and C. Meinel, “A deep semantic framework for multimodal representation learning,” *Multimed Tools Appl.*, vol. 75, no. 15, pp. 9255–9276, Aug. 2016.
- [16] R. Bahmanyar, D. Espinoza-Molina, and M. Datcu, “Multisensor earth observation image classification based on a multimodal latent Dirichlet allocation model,” *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 3, pp. 459–463, Mar. 2018.
- [17] S. Liu, M. Li, Z. Zhang, B. Xiao, and X. Cao, “Multimodal ground-based cloud classification using joint fusion convolutional neural network,” *Remote Sens.*, vol. 10, no. 6, p. 822, May 2018.
- [18] K. Hu, Z. Gui, X. Cheng, K. Qi, J. Zheng, L. You, and H. Wu, “Content-based discovery for Web map service using support vector machine and user relevance feedback,” *PLoS ONE*, vol. 11, no. 11, Nov. 2016, Art. no. e0166098.
- [19] N. Wang, “Research on the identification and extraction methods of map roads and rivers,” M.S. thesis, Guizhou Univ., Guiyang, China, 2007.
- [20] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: An astounding baseline for recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.
- [21] T. Cui, J. Liu, and A. Luo, “Intelligent identification method of network map images based on convolutional neural network,” *Sci. Surveying Mapping*, vol. 44, no. 1, pp. 118–123, Jan. 2019.
- [22] J. Yang and X. Zhou, “Semi-automatic Web service classification using machine learning,” *Int. J. u e-Service, Sci. Technol.*, vol. 8, no. 4, pp. 339–348, Apr. 2015.
- [23] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing*, vol. 174, pp. 50–59, Jan. 2016.
- [24] S. Qian, T. Zhang, C. Xu, and J. Shao, “Multi-modal event topic model for social event analysis,” *IEEE Trans. Multimedia*, vol. 18, no. 2, pp. 233–246, Feb. 2016.
- [25] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, “Robust human activity recognition using multimodal feature-level fusion,” *IEEE Access*, vol. 7, pp. 60736–60751, 2019.
- [26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-V4, inception-resnet and the impact of residual connections on learning,” in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 4278–4284.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [28] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [30] *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 Establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*, Apr. 2007.



deep learning (DL), multimodality fusion and analysis, and cloud computing.



**ZELONG YANG** was born in 1990. He received the bachelor's degree in cartography and geographic information system from Wuhan University, Wuhan, China, in 2014, where he is currently pursuing the Ph.D. degree in cartography and geographic information system with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing. His current research interest is on the geospatial information analysis, using natural language process (NLP), deep learning (DL), multimodality fusion and analysis, and cloud computing.

**ZHIPENG GUI** is currently an Associate Professor of geographic information science with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests are geospatial service chaining, high-performance spatiotemporal data mining, and geovisual analytics. He serves as the Co-Chair for the International Society for Photogrammetry and Remote Sensing (ISPRS) Working Group V/4-Web-based Resource Sharing for Education and Research.



**HUAYI WU** was born in 1966. He received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1999. He was a Postdoctoral Fellow with the Geospace Information and Communication Technology Laboratory, York University, Toronto, Canada, in 2002. He is currently a Professor and the Deputy Director with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University.

His research interests include deep learning, data mining, and distributed computing. He is also a member and the Secretary General of the GIS Theory and Method Committee of China. He is also the Chairman of the IV/2 Working Group, International Society for Photogrammetry and Remote Sensing.



**WENWEN LI** received the Ph.D. degree in earth system and geoinformation science from George Mason University, USA, in 2010. From 2010 to 2012, she worked with the Center for Spatial Studies, University of California at Santa Barbara, USA. She is currently an Associate Professor with the School of Geographical Sciences and Urban Planning, Arizona State University, USA. Her research interest is geographic information science, with a focus on cyberinfrastructure, big data, semantic interoperability, spatial information retrieval, and distributed geospatial information processing. She is also the 2015 NSF CAREER Award Winner. She was the Chair of the Association of American Geographers' Cyber-Infrastructure Specialty Group, from 2013 to 2014.

...