Remote Sensing of Environment 297 (2023) 113767

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information

Wen Zhou [a,*], Claudio Persello [a], Mengmeng Li [b], Alfred Stein [a]

[a] Dept. of Earth Observation Science, Faculty of Geo-information Science and Earth Observation (ITC), University of Twente, 7500AE Enschede, the Netherlands
[b] Key Lab of Spatial Data Mining & Information Sharing of Ministry of Education, Academy of Digital China (Fujian), Fuzhou University, Fuzhou, China

ARTICLE INFO

Edited by Marie Weiss

Keywords:
Building use classification
Mixed-use classification
Data fusion
Multimodal deep learning
Transformers
Natural language processing
Remote sensing

ABSTRACT

Assigning detailed use categories to buildings is a challenging and relevant task in urban land use classification with applications in urban planning, digital city modelling and twinning. This study aims to provide the categorisation of buildings with detailed use information by considering the possibilities of mixed-use. Mixed-use combines different use forms, and serves as a new type of use category. We obtain attributive information by combining satellite imagery that reflects spatial information and textual information from publicly available *point-of-interest* data collected by citizens and available on online maps. We propose a multimodal transformer-based building-use classification method to capture and fuse these different data sources within an end-to-end learning workflow. We evaluate the effectiveness of our proposed method on four urban areas in China. Experiments show that the proposed method effectively maps building use according to eight types of fine-grain categories, with a Micro F_1 score equal to 80.9%, and a Macro F_1 score equal to 62% for Wuhan research area. The proposed method is able to harness the relationship between the features obtained from the different data sources and results in higher accuracy than the state-of-the-art fusion-based multimodal integration methods. The proposed method can effectively increase the attributive grain of building use resulting in high classification accuracy.

1. Introduction and related work

Urban land use, as the highest level of human modification (Li et al., 2020; Theobald et al., 2020), reflects socio-economic functions and human activities. It is an important component of urban planning (Srivastava et al., 2019), landscape design, environmental management, health promotion, biodiversity conservation (Chen et al., 2021b) and city digital twins (Akroyd et al., 2022; Xia et al., 2022). Most research currently provides dominated categories for each land use unit and excludes the presence of mixed-use (Chen et al., 2021b; Gong et al., 2020; Häberle et al., 2022; Srivastava et al., 2018b; Zhu et al., 2019). Mixed-use is defined as a blending of multiple uses of a single object in the same space. Mixed-use may occur for different spatial units, in particular individual buildings, street blocks, and neighborhoods (Raman and Roy, 2019). Mixed-use, however, is a critical component in smart growth (Song et al., 2013), public health (McGuire, 2014), quality of life (Urbanism, 2000), compact cities, eco-cities, cycling-friendly cities, and sustainable development (Jiao et al., 2021). For example, The Congress of New Urbanism' Charter argues that: "Neighborhoods should be compact, pedestrian-friendly, and mixed-use"(Urbanism, 2000). Thus, acquiring mixed-use information is the basis for evaluating existing planning and design as well as for planning future urban development strategies. Recent research on mixed-use focuses on its quantification, e.g. using Shannon diversity index (He et al., 2021) of administrative units without considering detailed use information. Therefore, it is important to map land use including its mixed-use.

Due to the population increase and high urbanisation rates, urban land use changes rapidly. Collecting land use information including mixed-use at fine spatial scales is usually laborious and resource intensive, involving numerous field surveys (Liu et al., 2018; Zhan et al., 2014). This makes it imperative to design models that are capable of automating the generation of accurate and up-to-date land use maps including mixed-use.

Remote sensing (RS) images can provide interesting and specific land use information. The methods for Land use classification (LUC) from RS images can be categorized into three types based on the way of using images information. 1) Traditional pixels based classification methods such as support vector machines, fuzzy k-means clustering algorithms

* Corresponding author.
E-mail address: wen.zhou@utwente.nl (W. Zhou).

https://doi.org/10.1016/j.rse.2023.113767
Received 7 February 2023; Received in revised form 9 August 2023; Accepted 13 August 2023
Available online 24 August 2023

(He et al., 2014), maximum likelihood classification (He et al., 2014; Khorram et al., 1987; Rozenstein and Karnieli, 2011), which use the spectral information of individual pixels directly. 2) Object based image segmentation and classification methods (Galletti and Myint, 2014), which consider image spectral information and incorporate geometric and texture information of segmented objects. 3) Deep learning based image classification methods (Bergado et al., 2020; Huang et al., 2018a; Li and Stein, 2020; Zhang et al., 2018a, 2019; Zhou et al., 2020), which automatically learn a large number of deep features from images without manual feature extraction. These last methods usually obtain better performance than OBIA.

While the spatial resolution of RS images and available classification methods have improved, it is still challenging to obtain detailed land use information. Other types of data sources may also provide land use information, such as social media images reflecting building instance classification (Hoffmann et al., 2022; Hoffmann et al., 2019; Zhu et al., 2019), and social text information reflecting land use information (Chen et al., 2020; Häberle et al., 2019; Jendryke et al., 2017; Zhu et al., 2019). Multiple data sources have been combined in the past for detailed urban land use classification (Chen et al., 2021a; Gong et al., 2020; Hu and Wang, 2012; Huang et al., 2018b; Song et al., 2018). Here we consider point of interest (POI) data point data with coordinates and a site name, indicating for instance, use information of a location. POI data has been employed for building use classification (Deng et al., 2022; Lin et al., 2021), urban mixed-use measurement (Liu et al., 2018; Yue et al., 2017), and urban land use mapping (Barlacchi et al., 2021; Zhong et al., 2020). To achieve land use maps with detailed use information, we will leverage multiple data sources, capitalizing on the recently developed possibilities to fuse social media and RS data for geo-information retrieval, following Zhu et al. (2022).

Each data source can be seen as a modality. Multimodal integration refers to the process which integrates information from multiple modalities to create a coherent perception or understanding of the world. Multimodal integration methods can be divided into three types, 1) Data fusion at the early or input level (Khorram et al., 1987), where data that share the same form of media are combined, generating a new type of data. An example is the fusion of panchromatic images with multispectral images to produce a new image with high spatial and spectral resolution (pansharpening). 2) Feature fusion at the intermediate level (Antol et al., 2015; Mroueh et al., 2015; Ouyang et al., 2014; Wu et al., 2014), where features are extracted from different forms of media and fused into one same feature space. For instance, information is extracted from an image and from a text, both are transposed into a vector, and vectors of these two modalities are concatenated into a new vector. 3) Decision fusion at the late level (Cao et al., 2018; Chen et al., 2021a; Gong et al., 2020; Häberle et al., 2022; Workman et al., 2017), where each modality generates one decision, while results of different modalities are combined to generate an overall decision. Feature fusion and decision fusion are most suitable for research with input data consisting of different form of media such as image and text. So far, LUC research has been based primarily upon multimodality decision fusion (Cao et al., 2018; Chen et al., 2021a; Gong et al., 2020; Häberle et al., 2022; Lu et al., 2022; Workman et al., 2017; Zhong et al., 2020), while feature fusion based LUC studies (Srivastava et al., 2019) are rare.

Much effort has been made to develop multimodality land use classification. So far, the following problems have not yet been solved:

1) LUC research is mainly based upon pixels, objects (Häberle et al., 2022; Kang et al., 2018; Srivastava et al., 2019), and scene blocks (Zhang et al., 2018b; Zhou et al., 2020) as basic units. The bigger spatial units, however, may contain several land use categories, while current research usually assigns a single dominant category to each unit, thus neglecting mixed-use from the majority type of land use unit.
2) When fusing imagery information with textual information, most research use the decision fusion multimodal integration, neglecting the relationship between different modality features. For example, Häberle et al. (2022) used Bi-directional LSTM for text classification, several CNN models to classify images, and a single decision fusion method to combine their results. Song et al. (2018) used RS image to obtain building outline and POI data to determine building use categories. Chen et al. (2018b) firstly classified POI, then integrated land use results based upon POI with information from other data sources. Chen et al. (2018b) and Liu et al. (2017) combined the classification of POI with other data features for land use classification. Bao et al. (2020), Feng et al. (2021) and Lu et al. (2022) transformed the classification results of POI data to image and combined these with other image features for land use classification. The above research integrated the classification of textual data with features from other data or classification results but has not effectively used the relations between features extracted from different modalities.

To alleviate the above shortcomings, we realized that buildings may be devoted to more than one human activities. Considering that building use classification is a subset of land use classification, we propose multimodal Transformer-based feature fusion for building use classification based on remote sensing images and POI data. In our study, buildings are the objects with the smallest non-divisible units for land use classification. Rather than the current land use studies giving a dominant category to land use units, we give building use categories considering mixed-use situation. To do so, we utilise the relationships between different modalities by projecting textual features and image features into the same space, and then use a Transformer network (Vaswani et al., 2017) to classify fused features. The contributions of this work are as follows:

1) We increase the spatial and attributive grain of LUC by considering mixed-use of objects (buildings) level land use units. Thus, we diverge from assigning a dominant use category to each land use unit. Instead, we aim to predict the complete set of use categories for each building by considering various combinations of uses as a distinct type of building use category. By doing so, we aim to enrich the semantic information associated with buildings, offering a more comprehensive understanding of their functional attributes. In particular, it allows us to capture the intricate and diverse ways in which buildings are used, and provides a more nuanced representation of urban spaces.
2) We propose multimodal Transformer-based feature fusion, which simultaneously learns textual features, image spatial features and their relationships, and gives different modality features different attention.
3) We investigate the synergy between RS and POI for fine attributive grain building use classification, and the performance of decision fusion based and feature fusion based multimodal integration method for building use classification.

2. Study area and data

We selected four urban study areas from China: Wuhan, Zhengzhou, Xiamen, and Beijing, both in China (Fig. 1). These study areas cover northern China, middle China, southern China, and represent a diverse range of geographic characteristics, including coastal and inland cities. In terms of social and economic factors, Beijing is classified as a first-tier

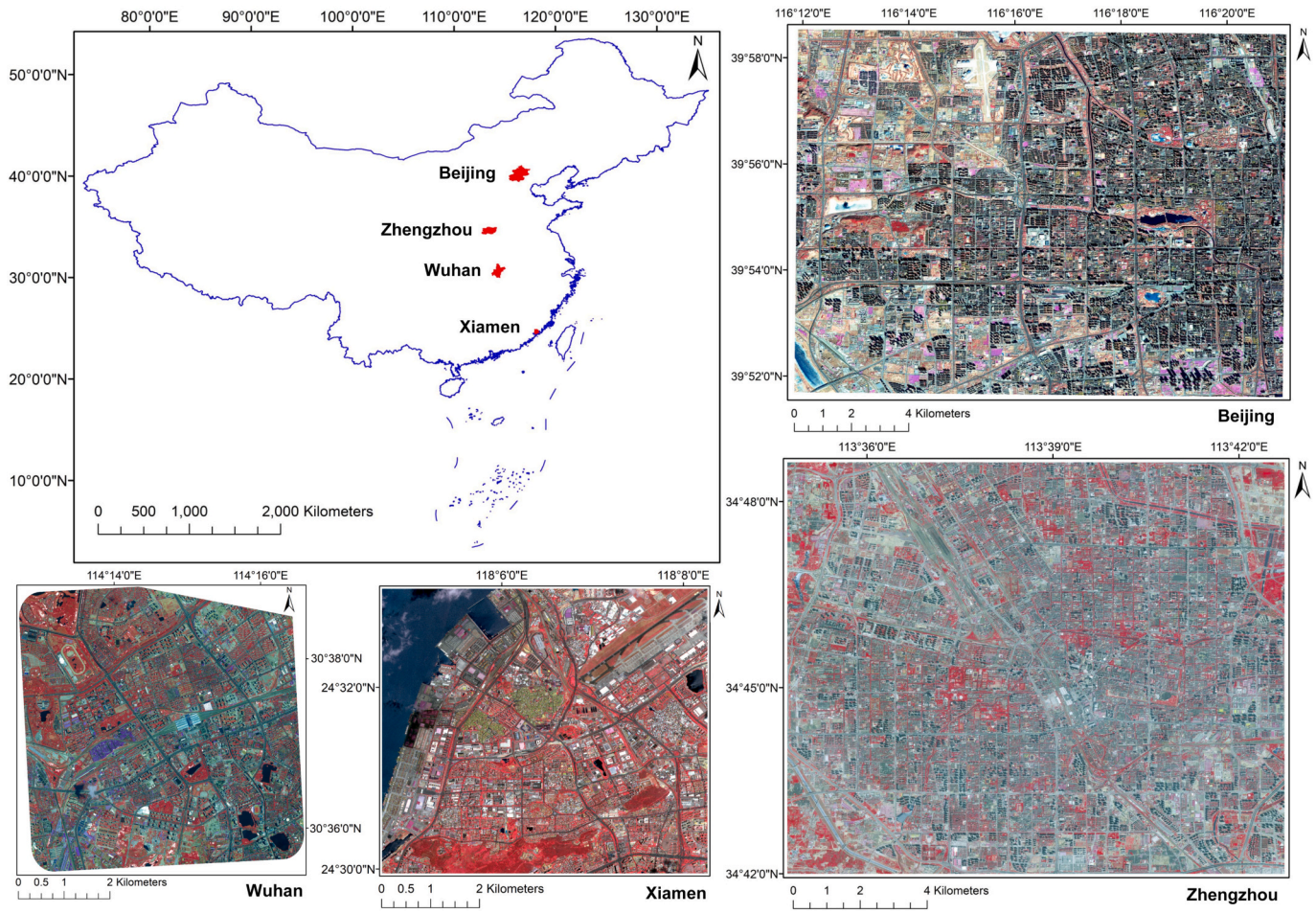


Fig. 1. Locations of four research area and their RS image. Beijing is a first-tier city, Zhengzhou and Wuhan are new first-tier cities, and Xiamen is a second-tier city.

city due to its high level of development and significant economic influence. Zhengzhou and Wuhan have recently been designated as first-tier cities, and Xiamen is considered a second-tier city. The choice of these four research areas is deliberate and idiosyncratic. They serve as unique test cases to evaluate the transferability of our proposed model. By including cities with different levels of development and economic profiles, we can assess how well our model adapts to varying urban landscapes and verify its effectiveness in diverse contexts.

The Wuhan study area covers most area of the Jianghan district, and parts of the Dongxihu, Qiaokou, and Jiangan district. The Zhengzhou study area covers parts of the Zhongyuan, Huiji, Jinshui, Guangchenghuizu, and Erqi districts. The Xiamen study area mainly lies in the Huli district and includes parts area of Siming district. Finally, the Beijing study area fully covers the Shijingshan district, and parts of the Mentougou, Haidian, Xicheng, Fengtai districts. Satellite images of the first three cities were obtained from the SuperView-1 satellite, all with a consistent spatial resolution of 0.5 m, while the image of Beijing was acquired from the GF-2 satellite, with the same spatial resolution. Images of Wuhan, Zhengzhou, Xiamen, Beijing is from 2019, 2016, 2020, and 2022, respectively. POI data of the four cities were acquired from "Amap" (https://lbs.amap.com/) from the same years same with their satellite images. Building footprints of the four cities are downloaded from the "Baidu Map" acquired in the year corresponding to their satellite images. These contain 6566,40,487, 7179, and 54,445 building footprints, respectively.

3. Methodology

Building is the basic spatial unit of this research; to determine its use category, we need three key processes (Fig. 2). We first capture the RS image and POI data corresponding to the same building, and use a file to align these two types of data. Second, we generate building use classification data sets by manually labelling their category and specifying their uncertainty. Third, we use this data set to train the multimodal deep learning method, and used this trained model to classify unlabelled buildings.

3.1. Data sets generation

We merged adjacent building polygons if adjacent polygons belong to the same building but were divided into several parts. Next, we used building polygons to capture aerial images. As shown in Fig. 2, we generated the centre point of every building polygon, and then captured RS image patches by using the centre point as the captured patch's centre and setting a suitable patch size. This progress can guarantee that the corresponding building lies in the centre of the captured RS image patch with its surrounding around the patch. In this research, the size of extracted patches is 224 × 224 pixels. This size is large enough to involve most buildings' own and their surroundings' information.

Matching between POI data and building polygons is done according to their spatial relationship. Most POI lie within the building's polygon,

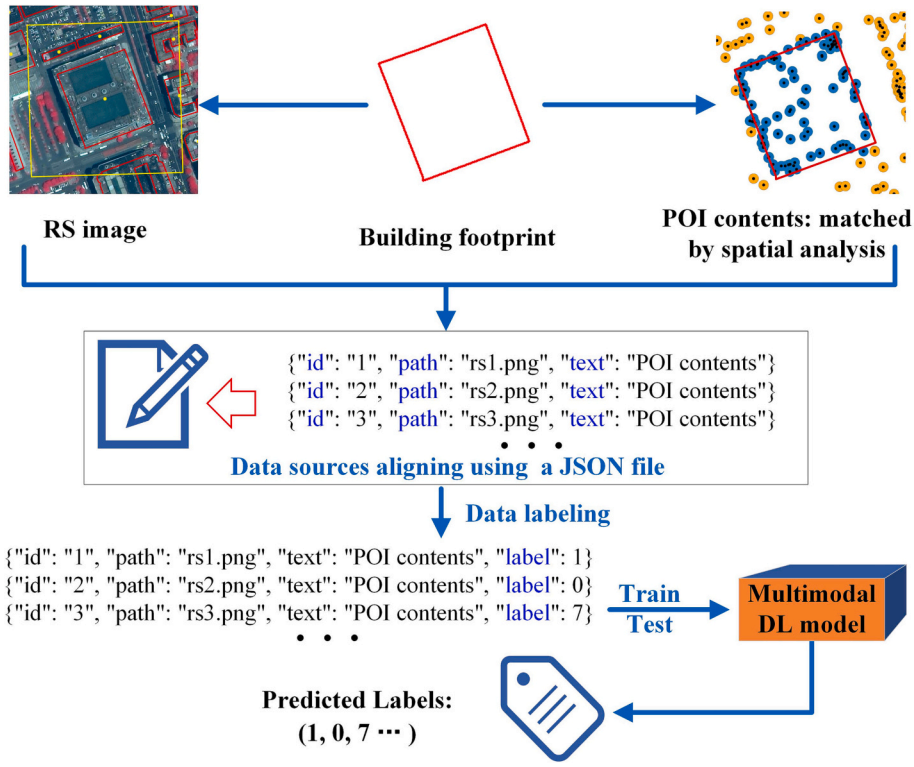


Fig. 2. Workflow of the proposed method.

while some POIs that describe the information of buildings may be outside the building footprint, e.g., some of the POIs of the blue circle in Fig. 2. Therefore, we matched every POI with its corresponding building by searching the nearest building of that POI within a 5 m radius.

We labelled the building use category based upon two modalities according to the land use classification system proposed by the Ministry of Housing and Urban-Rural Development of the People's republic of China (CAUPD, 2018), see appendix. According to this classification system, building use is classified into six main types (Table 1). Currently building use classification considering mixed-use is rare (Srivastava et al., 2018a). In our research we have considered buildings' mixed-use by assigning class labels that combine multiple categories.

We selected the Wuhan study area for generating the labelled data set, containing 6566 building footprints. We labelled building category by combining a visual inspection of RS image and doing a reading analysis of the POI data. For example, some residential, industrial, buildings can be interpreted from RS image only, and POI data can tell if other usage occurs involved in these buildings. Some buildings are hard to interpret from RS image, and are labelled according to their POI. And some buildings lack POI and are hard to interpret from RS image; those assigned the category "Unknown". We found that 34.8% buildings lack POI data, while 45.7% among these are hard to label from only their corresponding RS images. After labeling, building use was classified int 23 categories. As shown in Table 1, the number of samples for the different categories is too low to train the multimodal deep learning model. Therefore, we selected 5451 buildings in 8 categories. i.e. the bold categories in Table 1, for our experimental analysis. We randomly selected 60% of the labelled buildings as training samples, 20% as validation samples, and 20% as test samples.

Table 2 shows the number of labelled samples in each category in different data sets, and the proportion of samples including POI data. All samples of the category of "RBA", "BA", and "RA" contain POI data, counting 9.8% of the total, where labelling into these categories relies on POI data. Also, 98.9% in category "RB", 96.5% in category "B", and 84.7% in category "A" contain POI data, being 57.2% of the total. The reason is that most of the labels are determined according to two types of data, especially the POI data, and only a few samples can be assigned labels according to the labels of the surrounding similar buildings. 88.5% of "I" (industrial use buildings) lack POI data, but these can be well recognised from the satellite image. Finally, 46.7% of "R" have no POI data, which means these residential buildings are labelled according to satellite imagery only.

Table 1
Considered use categories and the statistic of the generated data set.

Category	Number	Category	Number	Category	Number
Unknown	1044	**R A**	**60**	B A S	1
R	**2414**	R S	2	B I	11
B	**1024**	**R B A**	**164**	B I W	3
A	**196**	R B A S	1	B S	5
I	**87**	R B I	6	B W	16
W	4	R B S	5	A I	2
S	5	**B A**	**186**	Being build	9
R B	**1320**	B A I	1	Total	5451/6566

R: Residential use; B: Business-related and commercial service facilities use.
W: Logistics and warehousing use; A: Public management and public service facilities use.
S: Roads and transportation facilities; I: Industrial use.

Table 2
The statistics of labelled building samples in different sub-sets of Wuhan.

Category	Training	Validation	Testing	Sum	Samples with POI data	
R	1462	471	481	2414	1287	53.3%
B	600	216	208	1024	988	96.5%
A	109	42	45	196	166	84.7%
I	54	17	16	87	10	11.5%
R B A	107	31	26	164	164	100%
B A	120	30	36	186	186	100%
R B	794	261	265	1320	1305	98.9%
R A	31	13	16	60	60	100%
Total	3277	1081	1093	5451	4166	76.4%

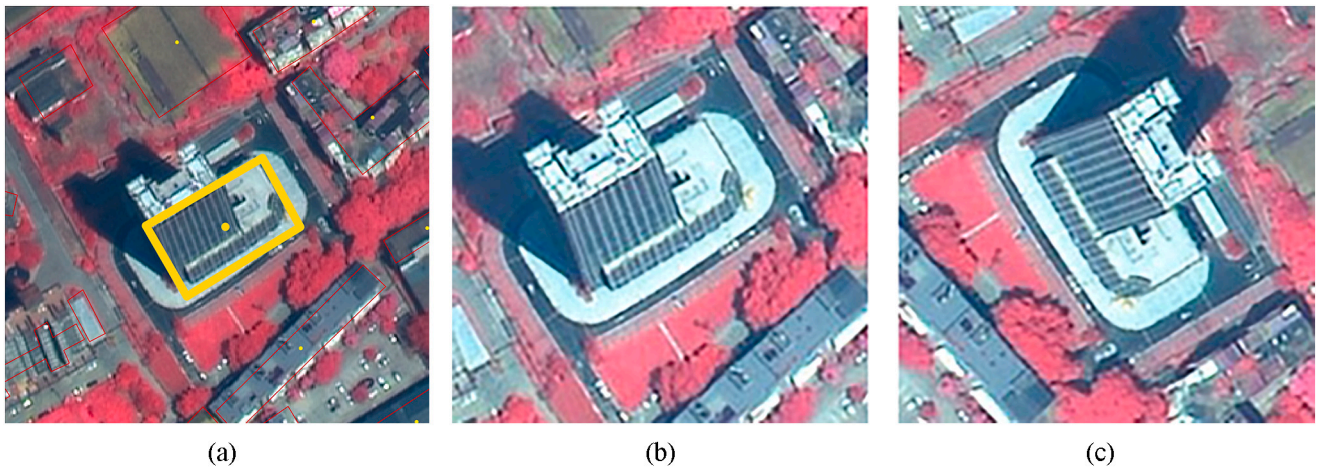


Fig. 3. Examples of the augmented image data sets. (a) Building's outline and its centre point. (b) Captured satellite image. (c) Captured satellite image rotated 90^0 to the right.

3.2. Data augmentation

The maximum length of the input text sequence in the deep learning method is usually fixed. If the length of the input text sequence exceeds the max length, then it is truncated, while if it is shorter, it is padded up to the max length using zero values. In this paper, the max length of input text sequence has been set to 300 characters. To adequately learn the data features, we augmented the training and validation data set of Wuhan study area by adjusting the orientation of satellite images and the sequence of the POI data contents. Fig. 3 shows a sample of a building's captured satellite image, and its augmented result.

Among the 5451 selected building samples, 54.5% have more than one corresponding POI data and we have gathered each building's POI data in two different turns. We next adjusted the sequence of the POI data contents to augment the data set. Each downloaded POI data has a unique ID, which we used to adjust their sequence. We combined the captured satellite images without orientation change like Fig. 3 (b) with POI contents ($P_1, P_2 \ldots P_n$), ordered according to increasing size, and the captured satellite images with orientation change like Fig. 3 (c) with the reversely ordered POI contents ($P_n \ldots P_2, P_1$). The POI content for buildings without POI data was set to "Null".

3.3. Data set uncertainty

The use categories of buildings are manually labelled based upon the information as reflected by RS and POI. The first uncertainty is associate with the ability of the two modalities to sufficiently reflect the actual use of the building. We did field checking work to evaluate this uncertainty by comparing the labels given according to the two modalities with the in-situ results. The second uncertainty concerns human errors in manually labelling the training samples. We assessed this uncertainty component of Wuhan by randomly sampling 10% of the data set and

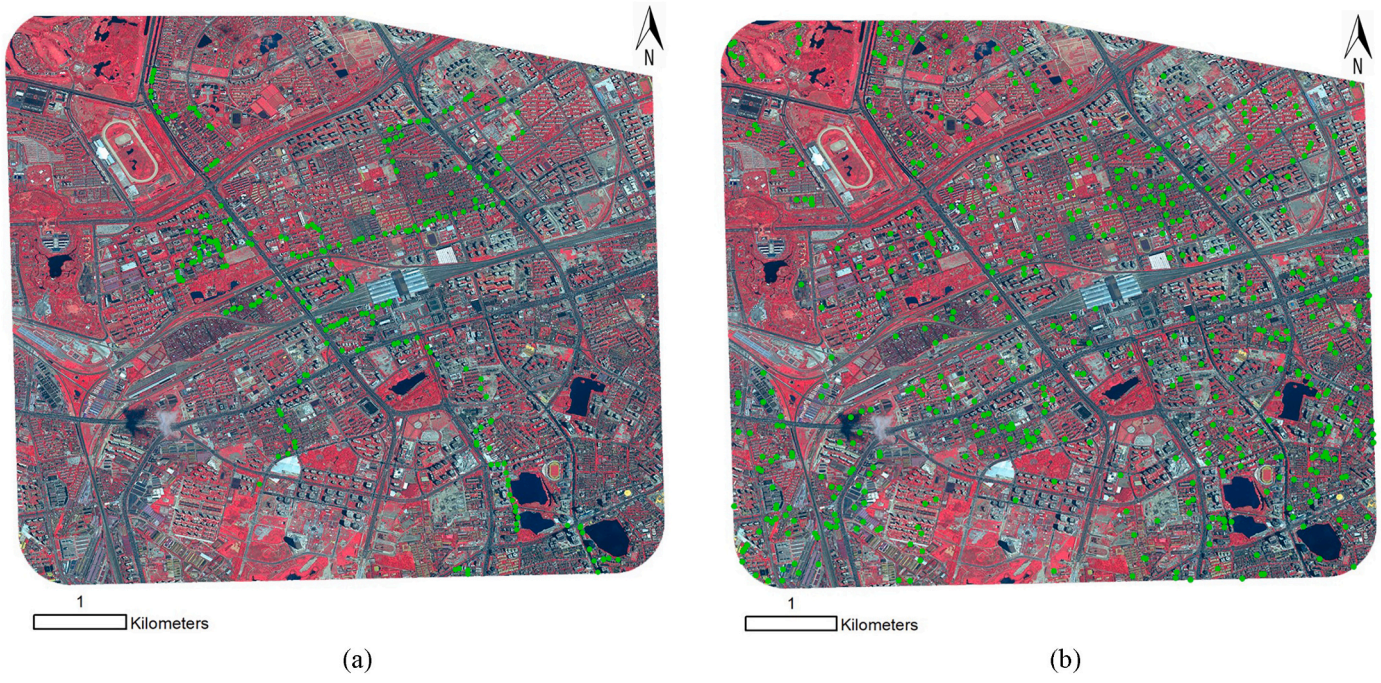


Fig. 4. Samples for data set uncertainty checking of Wuhan. (a) Field checking. (b) Relabeling samples.

Table 3
Number of labelled samples at the Zhengzhou, Xiamen, and Beijing study areas.

	Category	R	B	A	I	RBA	BA	RB	RA	Total
Zhengzhou	Number	435	162	98	232	50	43	281	31	1332
	Samples with POI (%)	65.5	85.2	61.2	5	100	95.3	97.5	93.5	66.7
Xiamen	Number	154	270	15	7	67	32	488	7	1040
	Samples with POI (%)	17.5	78.5	73.3	0	100	100	99.2	100	80.8
Beijing	Number	304	283	151	192	104	81	216	83	1411
	Samples with POI (%)	18.1	89	78.8	10.9	100	100	98.6	97.6	65.4

relabelling them based upon the two modalities, followed by comparing the relabelled result with their original labels.

For buildings with multiple use labels, we used the multi-label evaluation method for the first type of uncertainty, using the Accuracy (A, Eq. (1)) and F_1 score (F_1, Eq. (2)). We represented the field audit label as Y_i, and the assigned label as Z_i. The second type of uncertainty was evaluated by randomly selecting and relabelling building samples and calculating the correspondence between the relabelled results and the original results. Fig. 4 shows the field checking and relabelling samples for the first and second types of uncertainty evaluation.

$$A = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|} \tag{1}$$

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|Y_i \cap Z_i|}{|Y_i| + |Z_i|} \tag{2}$$

where n is the total number of field checking samples.

To evaluate the uncertainty of classification results, 20% of the samples of Wuhan have been selected as test data. For other research areas, we randomly selected and manually labelled 1332, 1040, and 1411 samples respectively as test data for the Zhengzhou, Xiamen, and Beijing study area. The detailed number of different category and the POI containing ratio is shown in Table 3.

3.4. Transformer based multimodal deep learning model

The proposed method involves two types of modalities containing both image and textual features, respectively. Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP), as they have better performance in terms of both accuracy and efficiency (Chen et al., 2018a; Vaswani et al., 2017).

Convolutional Neural Networks (CNNs) have been the preferred architecture in computer vision tasks for a long time. The typical architectures are LeNet-5 (Lecun et al., 1998), AlexNet (Krizhevsky et al., 2012), VGG (Karen and Andrew, 2014; Szegedy et al., 2015), GoogleLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), ResNeXt (Xie et al., 2017), DenseNet (Huang et al., 2017), SENet (Hu et al., 2018). Inspired by the successes of Transformers in NLP, Dosovitskiy et al. (2020) applied a standard Transformer directly to image classification, with the fewest possible modifications; splitting an image into patches and providing the sequence of linear embeddings of these patches as an input to a Transformer. When models are trained on larger datasets, they attained excellent results equal to or better than CNN (Bhojanapalli et al., 2021; Dosovitskiy et al., 2020).

Several studies on vision-language representation learning focus on modelling the interactions between image and text features with Transformer based multimodal encoders (Huang et al., 2021; Lu et al., 2020; Su et al., 2020; Zhang et al., 2021b). Therefore, Transformer is an excellent choice for textual and image feature extraction and fusion. The Supervised Multimodal Bitransformers (MMBT) (Kiela et al., 2019) model was proposed in 2020, performing better than state of the art methods. The MMBT model is built based on the Bidirectional Encoder Representation from Transformers (BERT) model (Jacob Devlin et al., 2019), which only utilises the encoder representation of Transformers.

Unlike BERT, the MMBT model includes a new module to extract image features and fuse these with textual features as the input of the model. It can employ self-attention over both modalities simultaneously, providing earlier and more fine-grained multimodal fusion. The MMBT model proposed by Kiela et al. (2019) used ResNet (He et al., 2016) to capture image features. In this research, we have replaced Resnet with DenseNet (Huang et al., 2017). Because the identity shortcut of ResNet stabilises its training but limits its representation capacity, while DenseNet has a higher capacity with multi-layer feature concatenation, although it requires high GPU memory and more training time (Zhang et al., 2021a).

Fig. 5 shows the architecture of the revised MMBT model. As shown in Fig. 5, the pretrained BERT and DenseNet are used as the backbone to extract textual features and image feature, and these two networks will then be finetuned in the proposed network. For each building, according to the Json file, its POI data's text content will be input into network. The pre-trained WordPiece tokeniser used in BERT was used to split a word into subwords and characters, and the pre-trained BERT vocabulary was used to transpose tokens from tokens to token sequences. The pre-trained BERT model is “bert-base-chinese” which was trained on the Chinese version of Wikipedia. Then token sequences were transposed to token embedding through the pre-trained token embedding layer.

For each building, its corresponding RS image is input into model, according to its image patch's name and folder. Then the pre-trained DenseNet (trained on ImageNet) is used to extract image features, capturing the output features after the final pooling operation. Next, these features are transposed to vectors with the same dimension as text token embeddings. Segmentation embeddings are used to distinguish text token embedding and image embedding by assigning different segment embeddings to them. 0-indexed positional coding is used for each segment to record token positions, i.e., start counting from 0 for each segment. Token, segmentation, and position embeddings were then fused and are input into the Transformer encoder for the classification task.

3.5. Classification result evaluation

For the multi-class classification task, two F_1 scores are commonly used as accuracy evaluation indexes, i.e. the Macro F_1 score (*MAFS*) and the Micro F_1 score (*MIFS*) (Santos et al., 2011). Compared to *MAFS*, *MIFS* considers the uneven number of samples of the different classes, which is more suitable for our research. In fine tuning the MMBT model, we used validation data set to evaluate model performance and adjust parameters.

For each class, precision (P_i) and recall (R_i) were obtained according to Eqs. (3) and (4), respectively. The F_1 score (F_{1i}) is obtained as the geometric mean of precision and recall (Eq. (5)) and the *MAFS* following Eq. (6), being the average of each category's F_1 score.

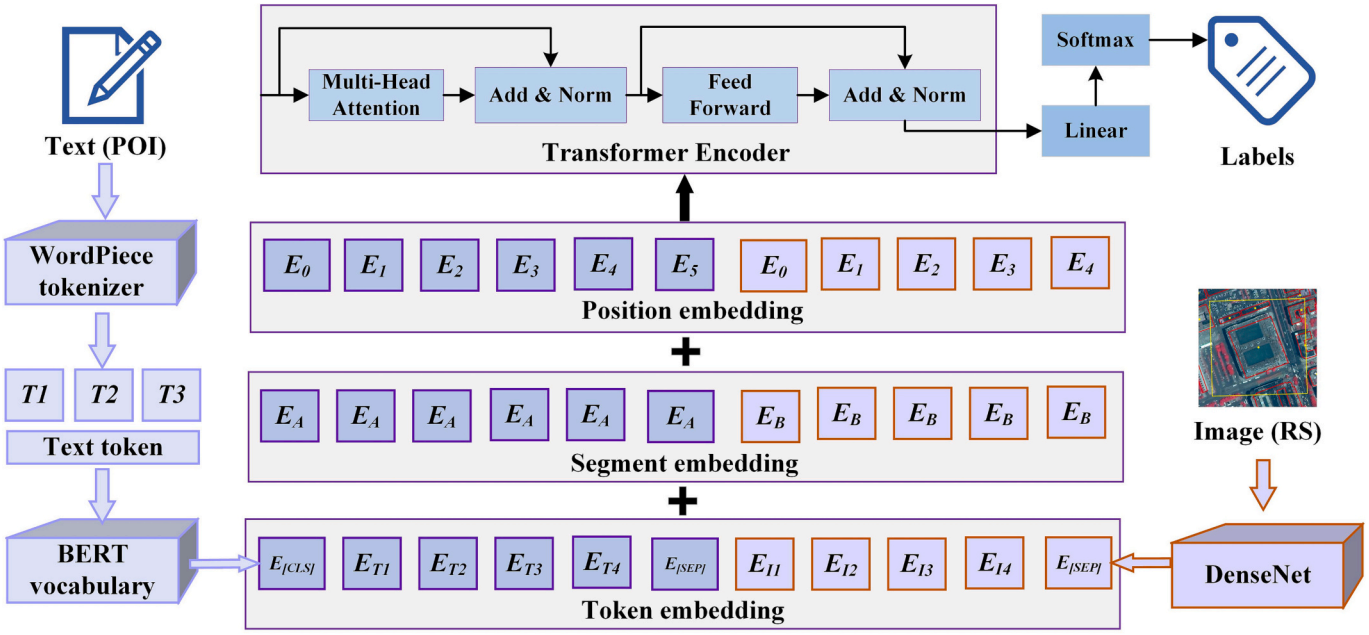


Fig. 5. Illustration of the revised MMBT model.

$$P_i = TP_i/(TP_i + FP_i) \tag{3}$$

$$R_i = TP_i/(TP_i + FN_i) \tag{4}$$

$$F_{1i} = (2P_i \cdot R_i)/(P_i + R_i) \tag{5}$$

$$MAFS = \left(\sum_{i=1}^{n} F_{1i}\right) \Big/ n \tag{6}$$

TP_i :true positives; FP_i: false positives; FN_i: false negatives; n: the number of class types.

For obtaining the *MIFS* we used the average precision (*AP*, Eq. (7)) and average recall (*AR*, Eq. (8)) of all samples with different classes, resulting in Eq. (9).

$$AP = \sum_{i=1}^{n} TP_i \Big/ \left(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i\right) \tag{7}$$

$$AR = \sum_{i=1}^{n} TP_i \Big/ \left(\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i\right) \tag{8}$$

$$MIFS = (2 \cdot AP \cdot AR)/(AP + AR) \tag{9}$$

4. Experimental analysis and results

4.1. Building use classification based upon two modalities (Wuhan study area)

For the two types of uncertainty evaluation of the Wuhan data set, after the field audit, the accuracy of our generated data set equals 88.5%, and the F_1 score 91.9%, indicating RS and POI can effectively reflect actual building use. For the 10% relabelled samples, 96.2% are correctly labelled, resulting in the correctness of the generated data set equal to 96.15%, reflecting dataset of Wuhan is reliable.

We used the AdamW optimiser to train the revised MMBT with a learning rate of 0.00003 and the batch size of 16. The default training epoch for MMBT is equal to 3, which in this research was set to 8 to ensure not miss the optimal trained model. Other hyper-parameters have been set according to default values. The training progress of using two modalities for classification is shown in Fig. 6. We used validation data set to evaluate model performance during training, and

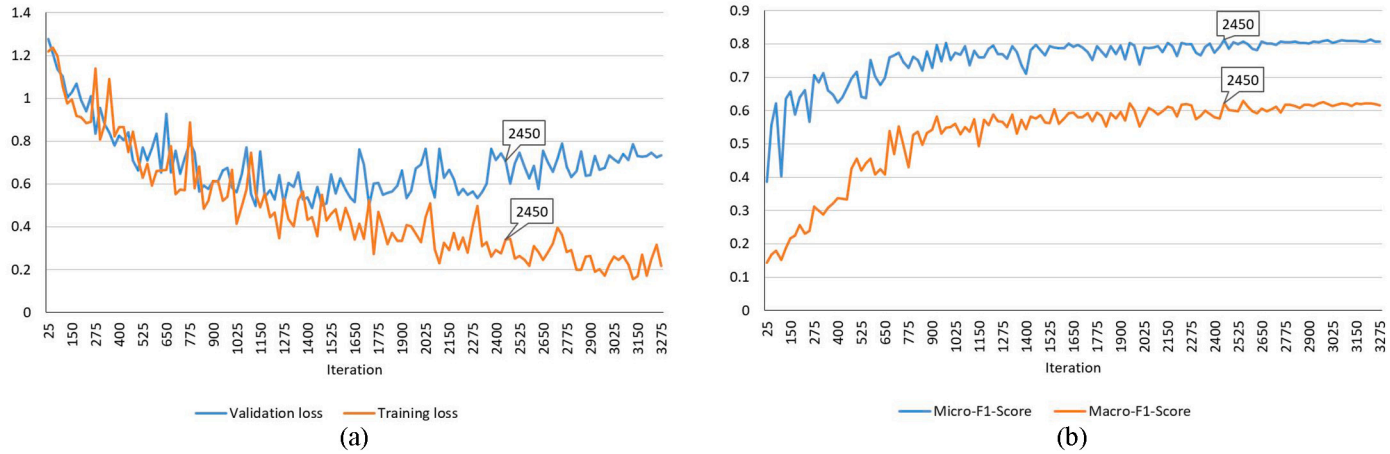


Fig. 6. Training progress when using two modalities for building use classification. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

Table 4
Building use classification results of Wuhan based upon two modalities using samples of Table 2.

Data set	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
F_1-score	0.930	0.738	0.723	0.486	0.321	0.515	0.788	0.483	0.809	0.623

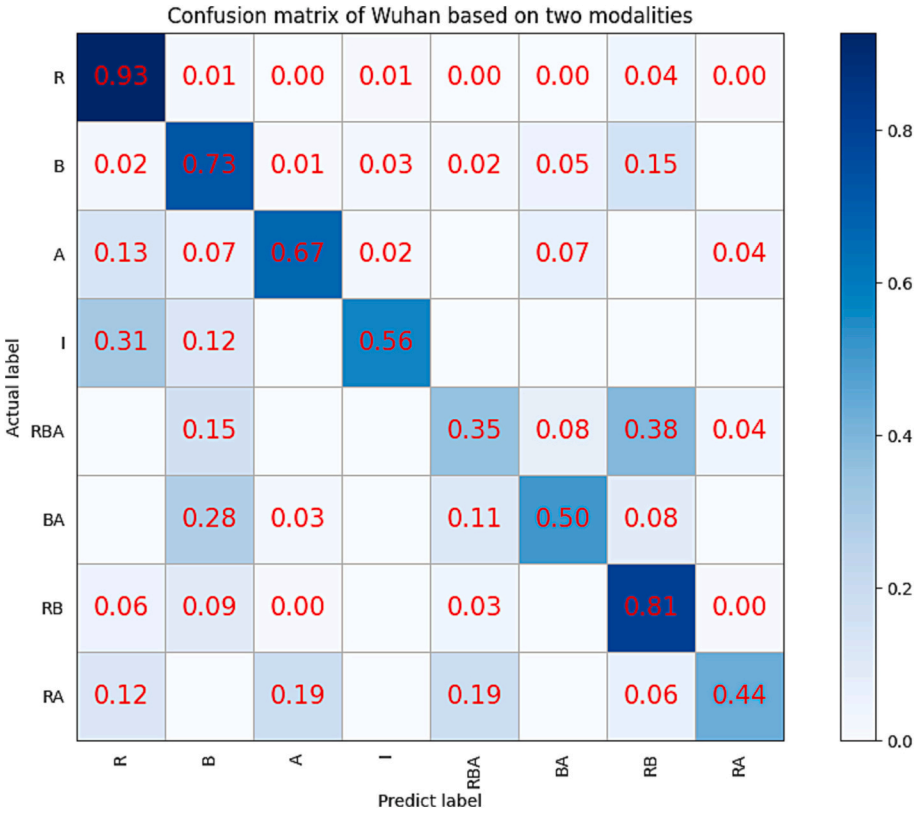


Fig. 7. Confusion matrix when using two modalities for classification based on MMBT model. We divided the number of each cell's samples by the total number of its corresponding row.

the model selection was based on the *MIFS* value on the validation set, i.e., that the model with the highest *MIFS* score after the 2450 iterations and in the sixth epoch has been selected.

The F_1 -score and the confusion matrix of the buildings use classification results based on two modalities when using samples shown in Table 2 are shown in Table 4 and Fig. 7. Three quarters single-use category's accuracy above 60% were obtained, while a quarter for mixed-use. For instance, the sample number of "A", "BA", and "RBA" are similar, while their accuracy gradually decreases. Categories with a larger number of samples also have a higher accuracy than others, for example, categories of "R", "B", and "RB" have more samples than the other categories, and their accuracies are also much higher than these. Misclassified categories also partly reflect their true labels. For example, 38% "RBA" has been misclassified into "RB" and 15% "RBA" to "B". Also, 28% "BA" have been classified into "B", 6% "RB" have been classified into "R" and 9% "RB" into "B". Hence, also misclassified building mixed-use labels reflect part of buildings' uses.

4.2. Contribution of different data sources

We compared the building use classification results on the test data of Table 2 using both modalities against that based upon one modality only. We used a pretrained Transformer to classify POI data. For the classification of RS images, we used Transformer to classify image features extracted from pretrained DenseNet. The samples used for the experiments based on "RS&POI" are identical to those based on RS alone, while the samples used for POI only experiments are fewer due to the presence of POI data in only 76.4% of the samples, as indicated in Table 2. As shown in Table 5, the *MIF*S based on two modalities, POI, and RS reached 80.9%, 76.4%, and 53.9%, respectively, and *MAFS*

Table 5
F_1 scores of different categories and their overall results trained by different data sources using samples of Table 2 based on the Transformer network.

Data set	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
RS&POI	0.930	0.738	0.723	0.486	0.321	0.515	0.788	0.483	0.809	0.623
POI	0.915	0.711	0.704	0	**0.37**	0.418	0.761	**0.529**	0.764	0.551
RS	0.690	0.578	0.100	0.412	0.108	0	0.255	0	0.539	0.267

Table 6
F_1 -score of four cities' building use classification results.

Study area	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
Wuhan	0.930	0.738	0.723	0.486	0.321	0.515	0.788	0.483	0.809	0.623
Zhengzhou	0.737	0.557	0.547	0.498	**0.370**	**0.550**	0.537	**0.576**	0.606	0.547
Xiamen	0.714	0.539	0.435	**0.833**	**0.413**	0.298	0.775	0.316	0.672	0.540
Beijing	0.724	0.546	0.565	0.409	**0.502**	0.424	0.533	0.494	0.566	0.525

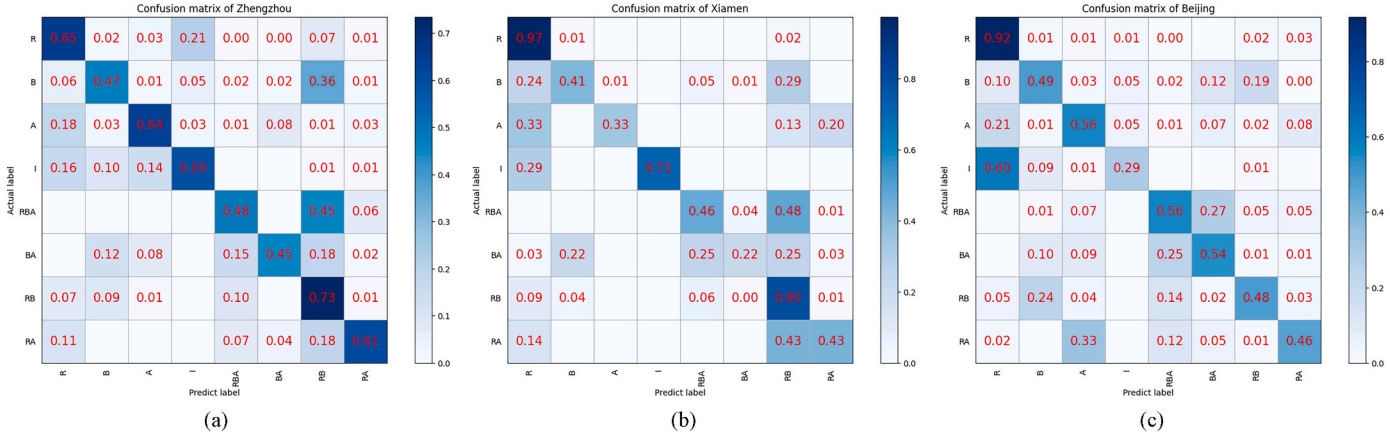


Fig. 8. The confusion matrix of three cities based on two modalities. (a) Result of Zhengzhou. (b) Result of Xiamen. (c) Result of Beijing.

values equal to 62.3%, 55.1%, and 26.7%, respectively. The confusion matrices of results based on POI only and RS only are reported in appendix. The ablation experiments show that POI data contributes more to detailed land use information than RS images.

Except for the categories of "RBA", and "RA" where using only POI data yields higher results compared to using two modalities, the remaining five categories demonstrate improved F_1 scores when both modalities are integrated. Table 2 shows that all samples in the "RBA" and "RA" categories contain POI data, and that the F_1 scores based solely on RS are notably low. This suggests a high reliance on POI data for determining these categories, while the inclusion of RS data introduces more noise than useful information. For the "I" category, RS imagery plays a crucial role since using POI data alone cannot accurately identify this category due to its limited presence. Table 2 shows that only 11.5% or 10 samples contain POI data. By incorporating RS data, the F_1 score for the "I" category increases by 18%, indicating the valuable information conveyed by the absence of POI data. A similar pattern is observed in the "BA" category also. This indicates that integrating RS and POI data effectively enhances building use classification.

4.3. Building use classification of the Zhengzhou, Xiamen, and Beijing study area

We analyzed the generalisation of the proposed method by applying the model trained in Wuhan to the Zhengzhou, Xiamen, and Beijing study area. We used the model trained in Wuhan based upon two modalities to classify buildings in these three cities using two modalities. We evaluated the classification results by comparing the classification results with the manually labelled reference data.

Table 6 presents the F_1 score for the classification result of four study areas. The *MIFS* for Zhengzhou equals 60.6%, for Xiamen 67.2%, and for Beijing 56.6%. The corresponding *MAFS* values are equal to 54.7% for Zhengzhou, 54.0% for Xiamen, and 52.5% for Beijing. Overall, the classification results for these three cities are lower as compared to Wuhan. In Wuhan, using both modalities, categories such as "R", "B", "A", "BA", and "RB" achieve F_1 scores above 50%, with 60% of them representing single-use buildings. Similarly, for Zhengzhou, the categories "R", "B", "I", "BA", and "RA" have F_1 scores above 50%, with 60% of them being single-use buildings. For Xiamen, categories "R", "B", "I", and "RB" have F_1 scores above 50%, with 75% of them representing single-use buildings. In Beijing, categories "R", "B", "A", "RBA", and "RB" achieve F_1 scores above 50%, with 60% of them being single-use buildings. This indicates that classifying single-use categories is generally easier than mixed-use categories. Fig. 8 displays the confusion matrix, while the classification maps can be found in Fig. 14-16 in the Appendix.

When transferring the trained model to a new domain, the accuracy usually decreases since different areas have different characteristics. This can be observed in the *MIFS*, *MAFS* and F_1 score of categories such as "R", "B", "A", and "RB" in the three transferred cities as shown in Table 6. In this research, significant improvements have been achieved for the F_1 scores of the "RBA" categories for Zhengzhou, Xiamen, and

Table 7
Truncation ratio of different categories in four cities.

Study area	R	B	A	I	RBA	BA	RB	RA
Wuhan (Test data set)	0.008	0.231	0.044	0	0.615	0.528	0.377	0.063
Zhengzhou	0	0.198	0.031	0	0.46	0.349	0.181	0.032
Xiamen	0	0.215	0	0	0.597	0.563	0.309	0
Beijing	0.003	0.117	0.04	0	0.308	0.438	0.093	0.073

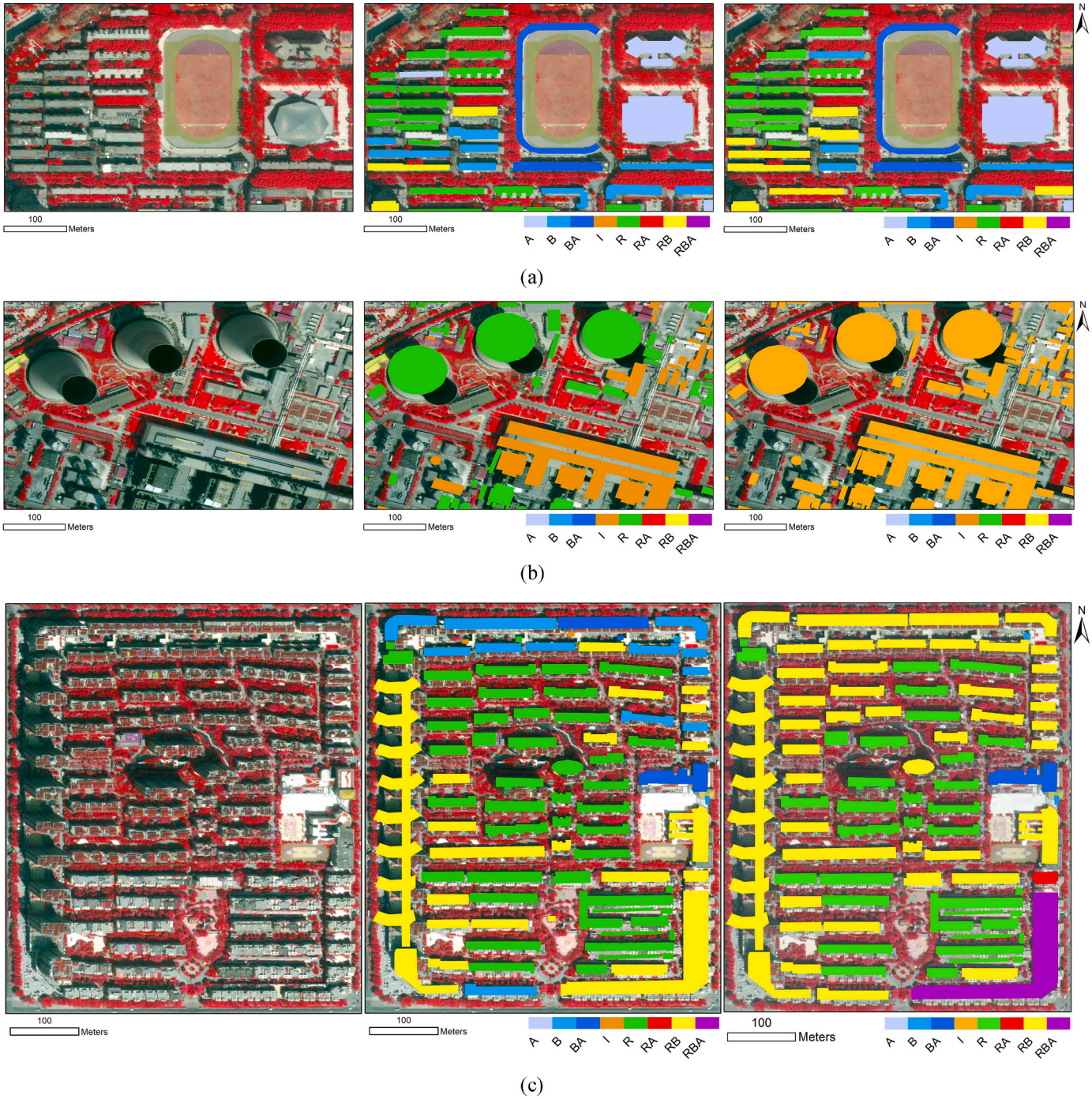
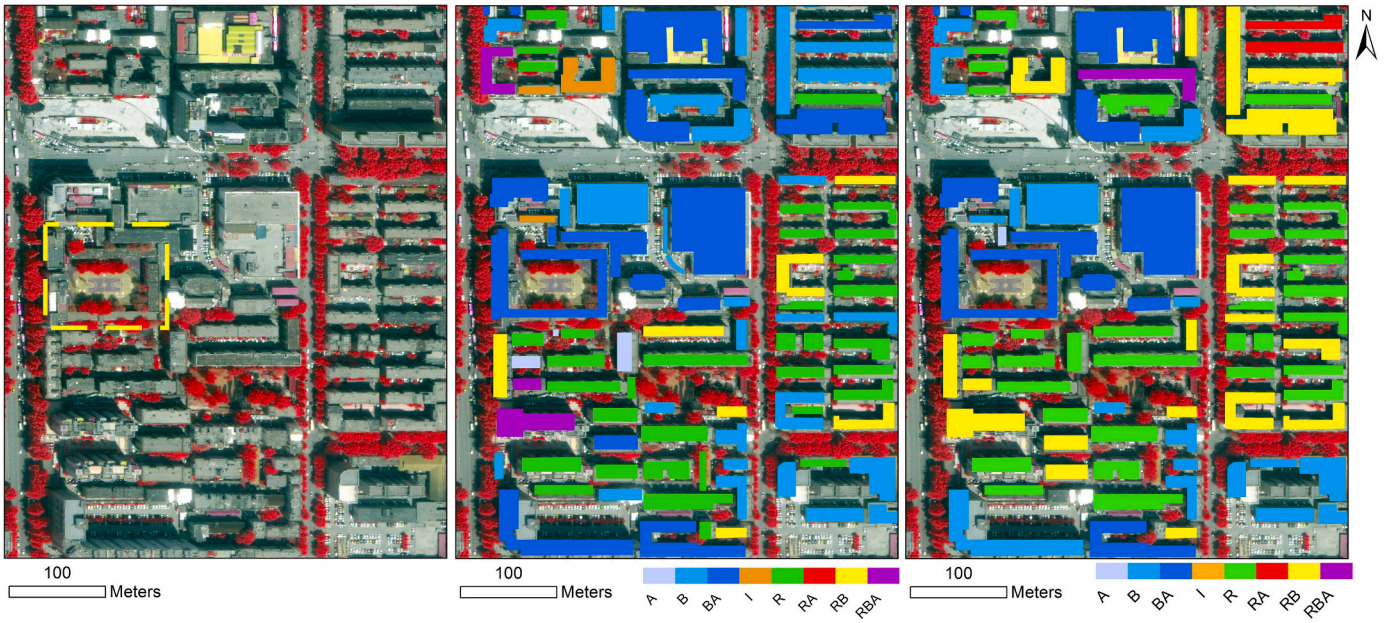


Fig. 9. Detailed building use classification result of Zhengzhou, (a) around a university, (b) around a industrial area, (c) around a living community, (d) around a commercial area. On the left is the RS image, in the middle is the classification result, and on the right is the reference label.



(d)

Fig. 9. (*continued*).

Beijing, the "RA" category for Zhengzhou, and the "I" category for Xiamen.

Due to the fixed text length required by the model, the input text is truncated if it exceeds the specified length. The truncation ratios for different categories in different cities have been calculated and are presented in Table 7. Categories like "RBA" and "BA" have relatively high truncation ratios. Analyzing the confusion matrices depicted in Fig. 7 and Fig. 8 for the four cities, it can be observed that, the function "A" has not been correctly identified, resulting in misclassifications into categories "B" or "RB". This misidentification rate equals 54% in Wuhan, 45% in Zhengzhou, 48% in Xiamen, and 6% in Beijing. Generally, higher truncation ratios correspond to higher rates of misidentification, while lower truncation ratios correspond to lower rates of misidentification.

Another reason influencing the F_1 score of "RBA" category in Wuhan is that 23% of the samples with the function "R" have not been correctly identified. For the "BA" category, 28% of the samples in Wuhan with

(a)

(b)

Fig. 10. Detailed building use classification result of Xiamen and Beijing. (a) Classification result of Xiamen (b) Classification result of Beijing. On the left is the RS image, in the middle is the classification result, and on the right is the reference label.

"BA" category have been misclassified as "B" without the correct identification of the function "A". In Zhengzhou, Xiamen, and Beijing, the proportions are equal to 12%, 22%, and 10%, respectively. These reduction ratios are 52.8%, 34.9%, 56.3%, and 43.8%, indicating that the reduction ratio can significantly impact the identification of building use.

The truncation ratio of "RA" is low, indicating less influence on the identification of "RA". Among the confusion matrices of the four cities, 19% of the samples have been classified as "A", without identifying the "R" function. In Zhengzhou, this proportion equals 0%. This suggests that identifying the "R" function in the mixed-use "RA" buildings is more challenging in Wuhan as compared to Zhengzhou. Xiamen, being a tourist city situated on an island, has fewer industrial buildings compared to the other research areas. In our test dataset, we only have 7 samples of the "I" category from Xiamen. These samples exhibit more pronounced industrial characteristics as compared to other cities, but the limited sample size might introduce sampling bias. This could explain why Xiamen has a higher F_1 score for the "I" category when compared to other cities.

Fig. 9 shows the detailed classification of Zhengzhou, confirming that most buildings are correctly classified. For instance, Fig. 9(a) shows objects around a university. We note that buildings used for education have been correctly classified into categories "A", and "BA". Four buildings in the living communities are of mixed-use "RB" but have been classified as "B", which is only partly correct. In contrast, Fig. 9(b) shows the objects in an industrial area. Here, approximately 50% of the buildings classified as "I" have been misclassified as "R". Fig. 9(c) illustrates a typical living community in China, featuring four gates. The buildings located outside the community serve residential functions along with other uses. According to the classification results, the pre-trained model failed to predict the residential function for 13.6% of the buildings, and 24.4% of the buildings were incorrectly labelled as "B". Additionally, buildings categorized as "RBA" were misclassified as "RB". In Fig. 9(d), the classification results for a multiple-use area are presented. The category "RB" is prone to being misclassified as "B". Furthermore, several buildings were mistakenly labelled as "A". The yellow rectangle in Fig. 9(d) represents a primary school, and in the vicinity of these buildings, there are education-related businesses with names containing education-related words. Consequently, these buildings were mistakenly labelled as "A".

Fig. 10 depicts a portion of the detailed building use classification results for the Xiamen and Beijing research areas. In the case of Xiamen, the majority of the buildings has been accurately classified. Still, two samples categorized as "RBA" have been misclassified as "RB", and five buildings labelled as "BA" have been misclassified as "B" without identifying their function as "A". Additionally, 15 buildings categorized as "B" have been classified as "RB", implying the presence of a non-existent function "R". For Beijing, the majority of buildings has also been correctly classified, while eight buildings categorized as "A" have been misclassified as "R", two buildings labelled as "B" have been misclassified as "RB", two as "BA" instead of "RBA", and one as "A" instead of "RA", thereby assigning an incorrect function "R".

Table 8
Number of buildings that contains both POI data and satellite image.

Category	R	B	A	I	RBA	BA	RB	RA	Total
Training	766	581	92	6	107	120	785	31	2488
Validation	240	206	37	2	31	30	259	13	818
Testing	281	201	37	2	26	36	262	16	861

Table 9
The statistics of F_1 score before and after decision fusion.

Dataset	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
LSTM(POI)	0.902	0.679	0.517	0	0.264	0.314	0.756	0.4	0.738	0.479
VGG16(RS)	0.571	0.521	0.192	0	0	0.164	0.340	0	0.454	0.224
DF of LSTM and VGG16 ($\lambda = \mathbf{0.85}$)	0.904	0.682	0.517	0	0.275	0.310	0.761	0.4	**0.741**	**0.481**
FF of our method	0.914	0.751	0.767	0	0.321	0.514	0.793	0.483	**0.787**	**0.568**

DF: decision fusion; FF: feature fusion.

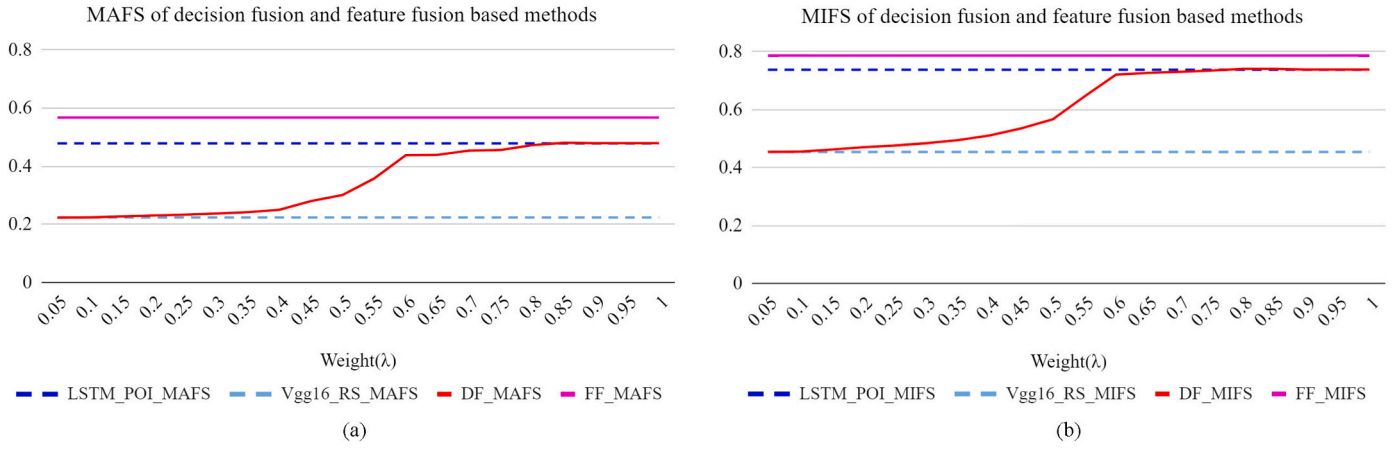


Fig. 11. The results of state-of-the-art decision fusion method and proposed feature fusion method. (a) The *MAFS* of decision fusion and feature fusion based methods. (b) The *MIFS* of decision fusion and feature fusion based methods.

Table 10
The statistics of F_1 score before and after decision fusion.

Dataset	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
Transformer (POI)	0,915	0,711	0,704	0	0,37	0,418	0,761	0,529	0.764	0.551
Transformer (RS)	0,604	0.622	0.074	0	0.111	0	0.266	0	0.488	0.210
DF of Transformer ($\lambda = 0.6$)	0.906	0.725	0.725	0	0.385	0.431	0.762	0.529	**0.764**	**0.551**
FF of Transformer	0.914	0.751	0.767	0	0.321	0.514	0.793	0.483	**0.787**	**0.568**

DF: decision fusion; FF: feature fusion.

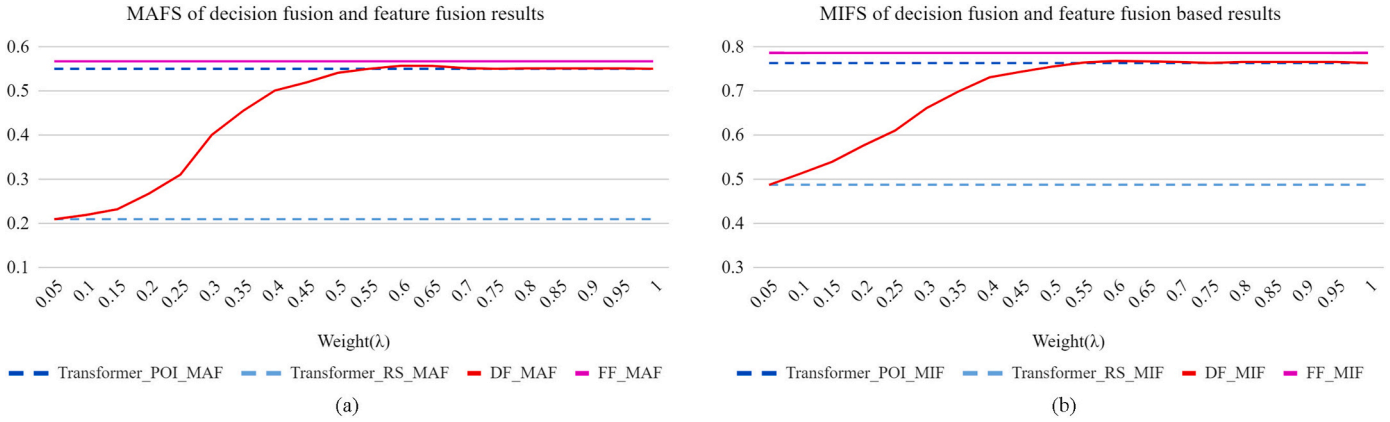


Fig. 12. The results of decision fusion strategy and feature fusion strategy. (a) The *MAFS* of decision fusion and feature fusion based methods. (b) The *MIFS* of decision fusion and feature fusion based methods.

4.4. Comparative experiments

We have compared the proposed method with a state-of-the-art building use classification method: a decision-based multimodal deep learning method (Häberle et al., 2022). This comparative method uses a bi-directional long short-term memory network (LSTM) (Graves et al., 2005) to classify text information, and generated the probability of each category for each data set in the vector form. VGG16 (Simonyan and Zisserman, 2014) was used to classify each image, and generate its corresponding category probability vector. The decision method using the two data sets is shown in Eq. (10), where P_t indicates the category probability vector for the text data, P_i is the category probability vector for the image data, and λ is the weight given to text information.

$$f_b = argmax[\lambda^* P_t + (1-\lambda)^* P_i] \tag{10}$$

Table 8 shows the samples that contain two modalities. We used the training data in Table 8 to train the LSTM model, and the training data in Table 2 which contain more image samples to train the VGG16 network. For these two networks, we used the same test data as shown in Table 8 to evaluate their performance. The decision fusion results are shown in Table 9 and Fig. 11.

From these experiments we see that: 1) Decision fusion represented by the comparative method has not effectively improved the building detailed use classification results. Compared to LSTM and VGG16 based decision fusion, our proposed feature fusion improved the *MIFS* by 6.2% and 3.0% respectively. 2) For classification based on single modality, the *MIFS* of the POI classification is substantially higher than that of RS. For the decision fusion experiments, the highest *MAFS* and *MIFS* occur when assigning POI classification results a weight of 0.85 and results based on RS images 0.15. Therefore, POI data contribute more to buildings' accurate use classification. 3) Using decision fusion, the contribution of RS images for building classification is limited and has not effectively improved the classification of POI.

4.5. Comparison between feature fusion and decision fusion strategy

We used the decision fusion strategy shown as Eq. (10) to fuse the classification results of POI data and aerial images classified by the Transformer network. The same network was used in decision fusion and feature fusion experiments. Results are shown in Table 10 and Fig. 12.

The MIFS and MAFS of the decision fusion are almost the same as that only using POI, which means in this decision fusion progress, adding RS classification results have not improved the overall results. The results of the proposed feature fusion method show that using feature fusion led to better classification accuracy, which shows that the relationship between different modality features can help improve classification results.

As compared to the decision fusion result in Table 9, Transformer performs better than LSTM in each category when dealing with POI data. When classifying RS images, the Transformer model demonstrates better performance in categories such as "R", "B", and "RBA", whereas the VGG16 model excels in categories like "A", "BA", and "RB". As shown in Table 2, samples with POI data in categories "R", "B" and "RB" are more than that in the other categories, and the F_1 score of "R" and "B" are higher than the other categories, irrespective of the use of VGG16 or Transformer. Hence, the *MIFS* of the Transformer model is 3.4% higher than that of VGG16, while the *MAFS* of VGG16 is 1.4% higher than that of the Transformer model. Considering the performance on different modalities, we conclude that the Transformer network outperforms alternative deep learning architectures and is highly effective for buildings use classification.

5. Discussion

5.1. The comparison between revised MMBT and the original MMBT

We conducted a performance comparison between our revised network and the original MMBT network. Both models were trained using the dataset presented in Table 2, utilizing two modalities. For

Table 11
Building use classification result of original and revised MMBT model.

MMBT	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
Revised	0.93	0.74	0.72	0.49	0.32	0.52	0.79	0.48	0.809	0.623
Original	0.92	0.78	0.65	0.00	0.30	0.53	0.78	0.51	0,806	0,558
Compare	**−0.01**	+0.04	**−0.07**	**−0.49**	**−0.02**	+0.01	**−0.01**	+0.03	**−0.03**	**−0.065**

Table 12
Building use classification result using different search radius to match POI data with buildings.

Search radius	R	B	A	I	RBA	BA	RB	RA	*MIFS*	*MAFS*
0 m	0.88	0.72	0.6	0	0.29	0.55	0.77	**0.55**	0.783	0.544
2.5 m	0.92	**0.74**	0.57	0	**0.38**	**0.58**	0.77	**0.55**	0.797	0.563
5 m	**0.93**	**0.74**	**0.72**	**0.49**	0.32	0.52	**0.79**	0.48	**0.81**	**0.62**
7.5 m	0.91	**0.74**	0.67	0.36	0.26	0.49	0.77	0.48	0.786	0.587
10 m	0.9	0.73	0.66	0.42	0.31	0.35	0.74	0.35	0.769	0.559

feature extraction, DenseNet and ResNet152 were employed as the backbones for MMBT models. The training progress of the original MMBT model can be observed in Fig. 17, which is included in the Appendix.

Table 11 presents the comparison results. The original network exhibits better performance in the categories of "B", "BA", and "RA", while performing worse in the remaining five categories, particularly in "I" and "A". However, when considering the overall results represented by *MIFS* and *MAFS*, it is evident that our revised MMBT model generally outperforms the original one.

5.2. The matching problem between POI and building footprints

The performance of data-driven models is highly dependent on the quality and composition of the input data. The POI data plays a significant role in building use classification. In order to optimize the matching process between POI and building footprints, we experimented with different search radii and selected the radius with the highest F_1 score as the optimal search radius. The training progress for each radius is illustrated in Figs. 18 to 21 in the Appendix.

Table 12 presents the classification results obtained using different search radii. It was observed that the category "RA" achieved the optimal F_1 score when using 0 m and 2.5 m as the search radius, although these radii failed to recognize the "I" category. The "B" category attained the highest F_1 score when using search radii of 2.5 m, 5 m, and 7.5 m. For the categories "RBA" and "BA", the optimal F_1 score was achieved with a search radius of 2.5 m. The categories "R", "A", "I", and "RB" obtained their highest F_1 scores with a search radius of 5 m. In general, 62.5% of the categories achieved their optimal F_1 score with a 5 m search radius, and this radius also yielded the highest *MIFS* and *MAFS*. Consequently, we recommend using a search radius of 5 m when matching POI with building footprints in China.

6. Conclusion

In this work, we proposed a new multimodal transformer-based deep learning method based upon feature fusion for building use classification. Based upon our research, we found that by integrating POI and RS data, we can extract detailed mixed-use information. Our proposed method outperforms the state-of-the-art methods in terms of performance.

We draw five conclusions. 1) POI and RS images effectively reflect the building's detailed use information, including mixed-use. 2) Compared to RS images, POI data reflect more functional information, while combining two modalities provides more information than using a single modality. 3) The proposed feature fusion strategy performs better than the state-of-the-art decision fusion method, and it increases the classification accuracy. 4) Single use categories usually have a higher accuracy than mixed-use categories. 5) Based upon the four case studies we hypothesise that the proposed method has a good generation ability for other major Chinese cities.

The performance of our deep learning method heavily relies on the number of training samples. The accuracy is varying among different categories, for example, classification results of "RBA", "RA", "I", and "BA" with less samples are inferior to those with sufficient samples. Different application scenarios may have varying requirements for data accuracy. It is crucial to consider the uncertainty associated with the generated classification results before utilizing them. In our future research, we will enlarge our data set and add new types of data sources such as street view images to improve the building use classification result. We will then also explore use of our method in different parts of the world and will explore urban structures and unequal access to services.

CRediT authorship contribution statement

Wen Zhou: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Claudio Persello:** Conceptualization, Writing – review & editing, Supervision. **Mengmeng Li:** Resources, Data curation. **Alfred Stein:** Conceptualization, Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Wen Zhou reports financial support was provided by China Scholarship Council.

Data availability

Data will be made available on request.

Appendix A. Fig.13-Fig.21

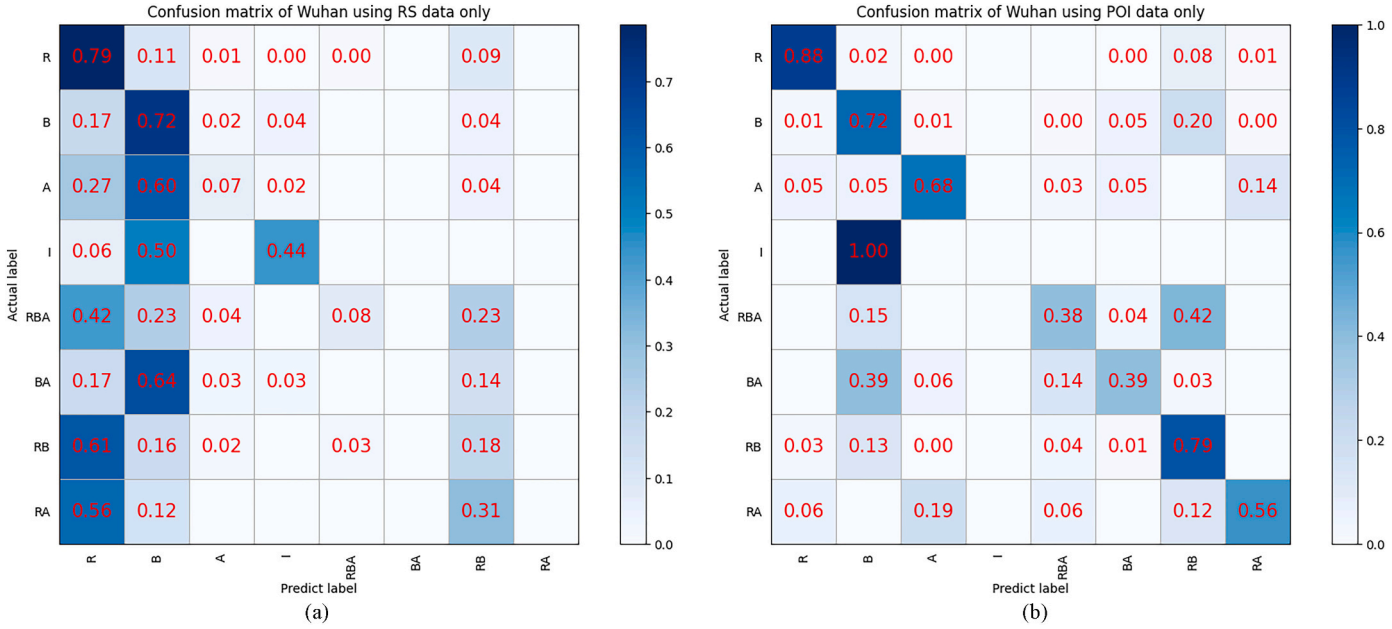


Fig. 13. Confusion matrix when using RS and POI data for the classification based on MMBT model. (a) Result only using RS. (b) Result only using POI.

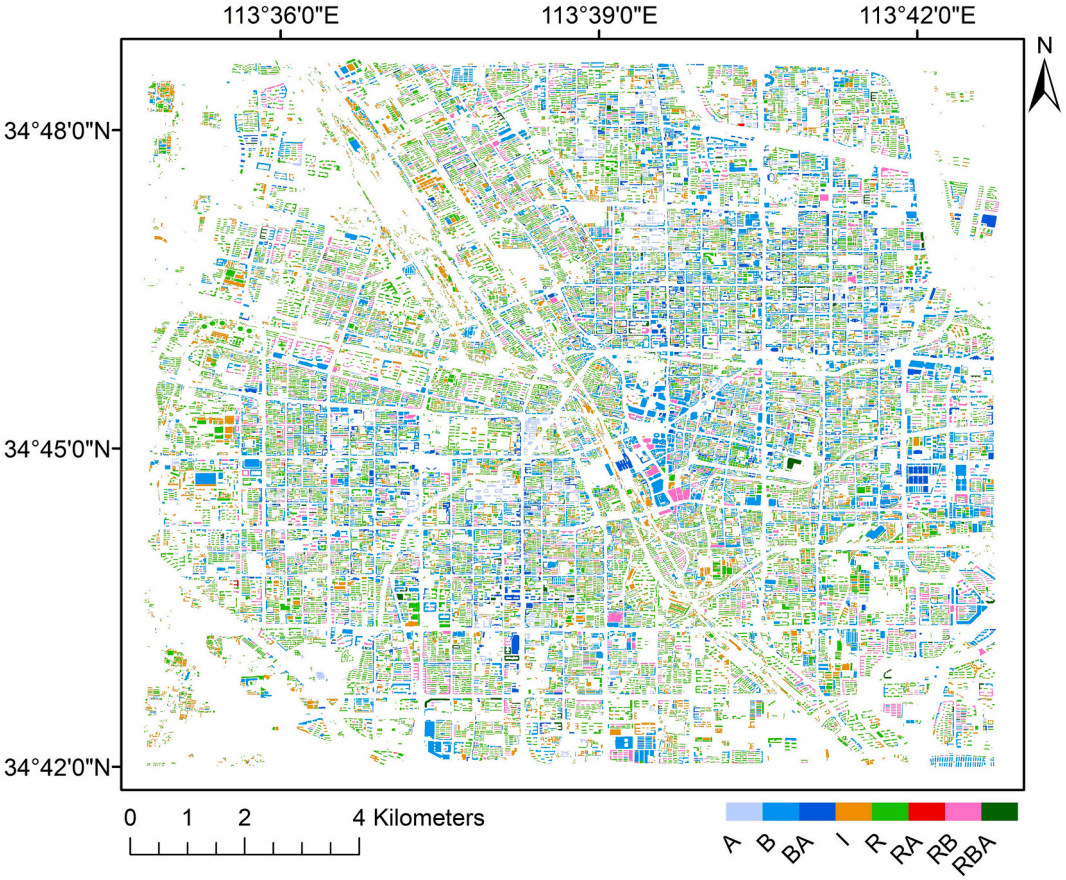


Fig. 14. Building use classification of Zhengzhou.

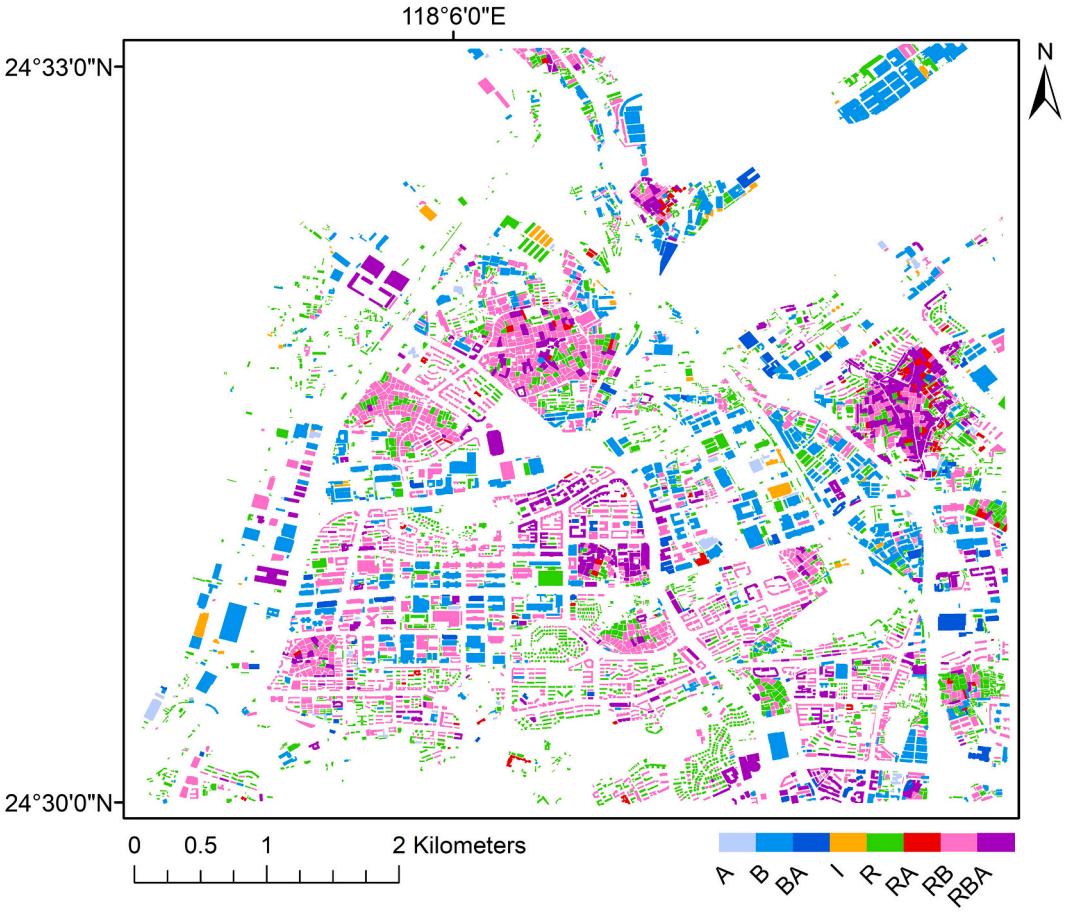


Fig. 15. Building use classification of Xiamen.

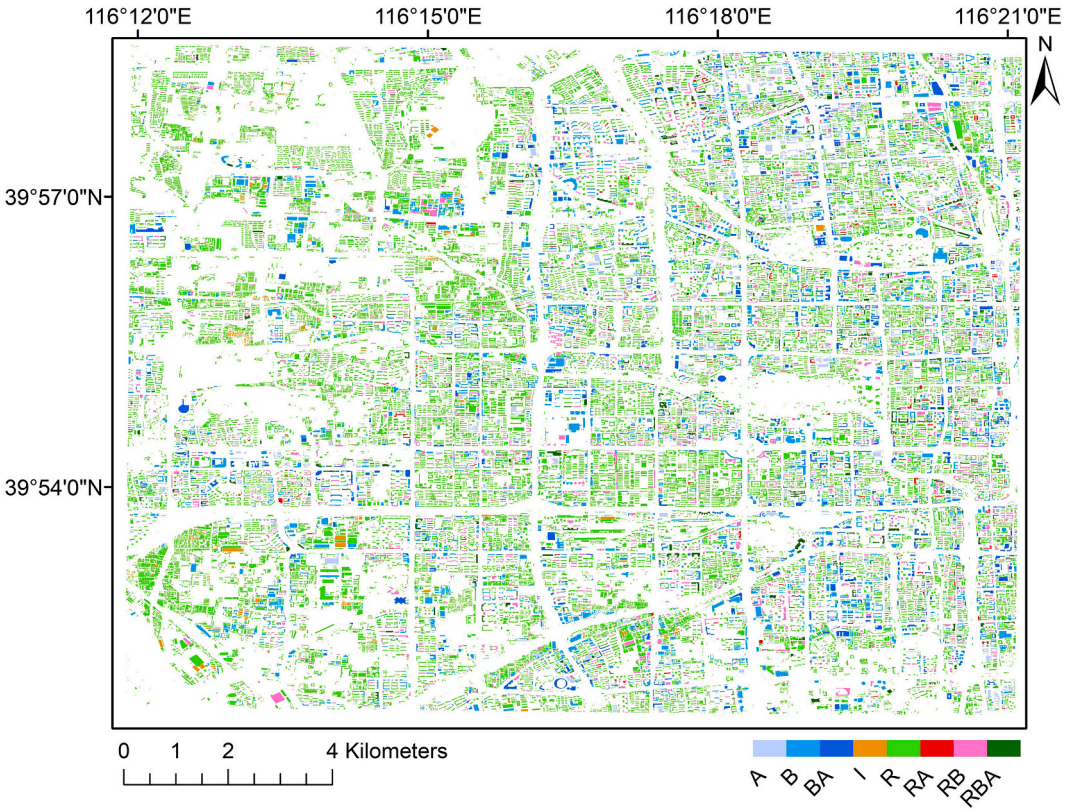


Fig. 16. Building use classification of Beijing.

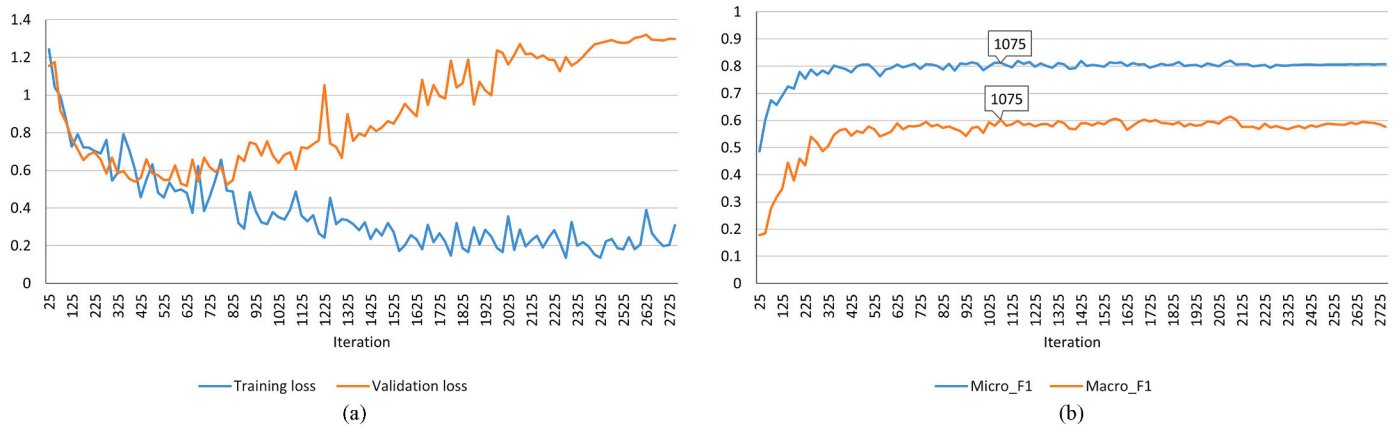


Fig. 17. The training progress of using original MMBT model for building use classification based on two modalities. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

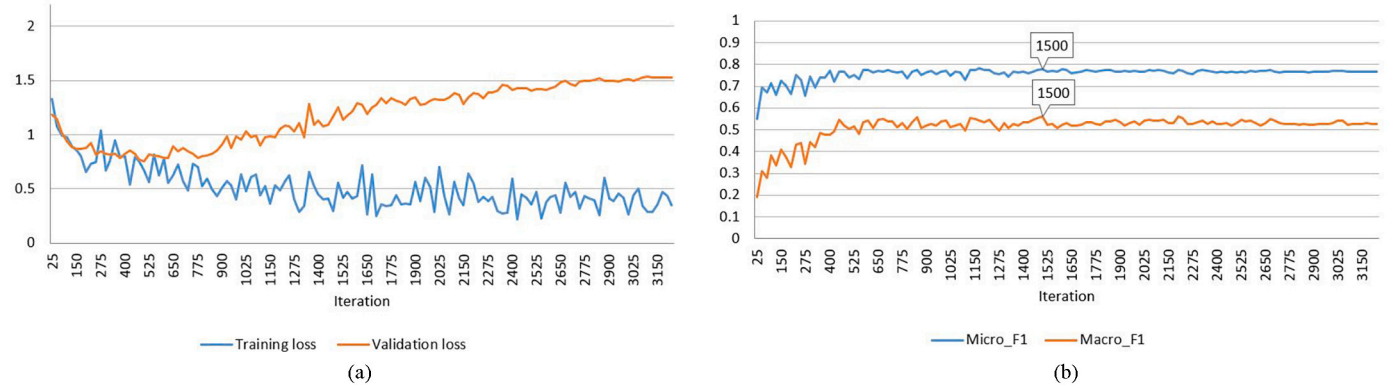


Fig. 18. Building use classification result using 0 m search radius to match POI data with buildings. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

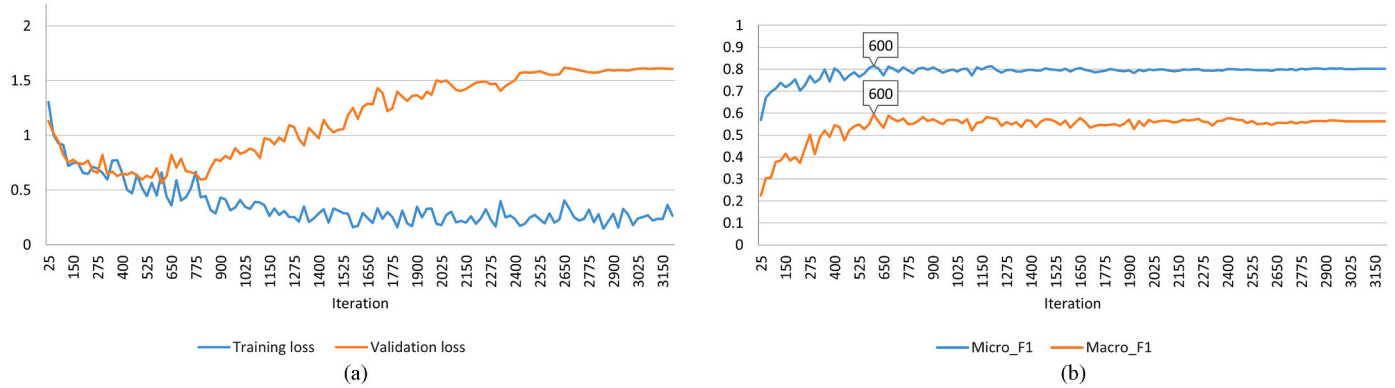


Fig. 19. Building use classification result using 2.5 m search radius to match POI data with buildings. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

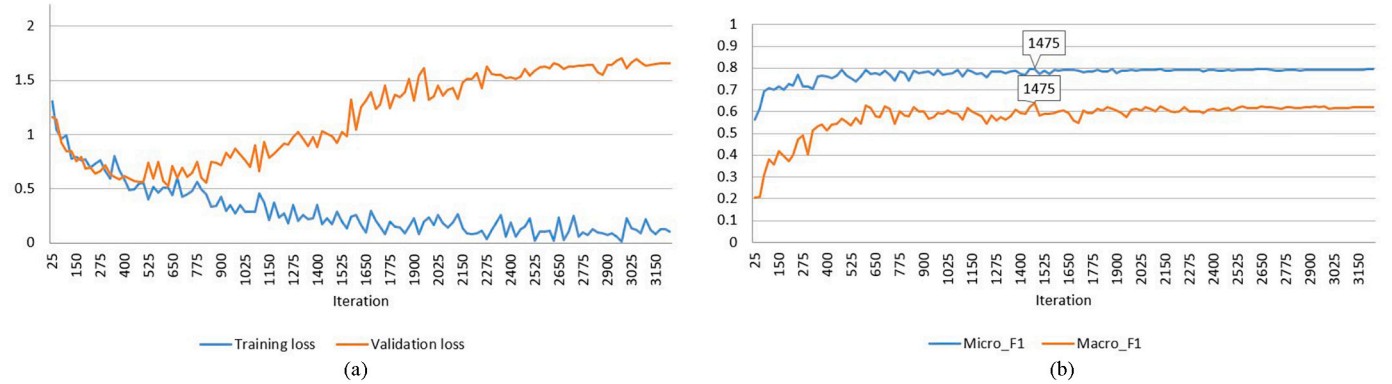


Fig. 20. Building use classification result using 7.5 m search radius to match POI data with buildings. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

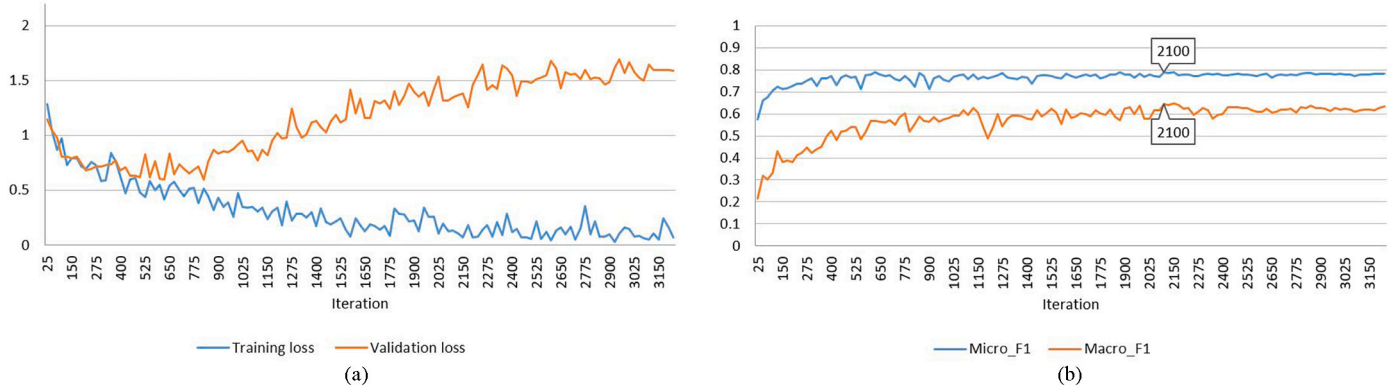


Fig. 21. Building use classification result using 10 m search radius to match POI data with buildings. (a) Loss value at different iteration. (b) F_1 score of validation data at different iteration.

Appendix B. The Land use classification system used in this research

Table 13
The land use classification system used in this research.

Category code			Categories	Description
Coarse classes	Medium classes	Detailed classes		
R			Residential Land	Residence and its corresponding service facility.
	R1		Residential land one	Land with complete facilities and good environment, mainly covered by low-rise houses.
	R2		Residential land two	Land with complete facilities, good environment, mainly covered by multi-, medium-, and high-rise residential land.
	R3		Residential land three	Land lacking facilities and with a poor environment, and mainly covered by simple houses that need to be renovated, including dilapidated houses, shanty towns, and temporary housing.
B			Business-related and commercial service facilities land	Land for commercial, business, entertainment, and sports facilities, excluding land for service facilities in residential land.
	B1		Commercial land	Commercial, catering, hotel, and other service industry land.
	B2		Business land	Comprehensive office land for finance, insurance, art, media, R&D and design, technical services, etc.
	B3		Land for entertainment and sports facilities	Land for entertainment, sports, and other facilities.
	B4		Land for public facilities and business outlets	Land for retail fuel, gas, telecommunications, postal, and other public facilities business outlets.
I			Industrial land	Land used for production workshops, warehouses, and ancillary facilities of industrial and mining enterprises, including land for special railways, wharves and ancillary roads, parking lots, etc., excluding land for open-pit mines.
	I1		Industrial land 1	Industrial land that basically has no interference, pollution, and safety hazards to the residential and public environment, including industrial land that focuses on industrial research and development, pilot trials, and small-scale production.
	I2		Industrial land 2	Industrial land that has certain interference, pollution, and safety hazards to the residential and public environment

(*continued on next page*)

Table 13 (*continued*)

Category code			Categories	Description
Coarse classes	Medium classes	Detailed classes		
	I3		Industrial land 3	Industrial land that with serious interference, pollution, and safety hazards to the residential and public environment.
W			Logistics and warehousing land	Land for material storage, transit, distribution, etc., including land for affiliated roads, parking lots, and trucking company fleet stations
	W1		Logistics and warehousing land 1	Logistics and storage land that is basically free of interference, pollution, and safety hazards to the residential and public environment.
	W2		Logistics and warehousing land 2	Logistics and storage land that has certain interference, pollution, and safety hazards to the residential and public environment
	W3		Logistics and storage of dangerous goods	Special logistics and storage land for dangerous goods such as flammable, explosive, and highly toxic.
A			Land for public management and public service facilities	Land for administrative, cultural, educational, sports, health, and other institutions and facilities, excluding land for service facilities in residential land.
	A1		Land for administrative office	Land for administrative office and related facilities such as political party and government agencies, social organisations, institutions, etc.
	A2		Land for cultural facilities	Land for public cultural event facilities such as libraries and exhibitions.
	A3		Educational land	Land for colleges and universities, secondary professional schools, middle schools, primary schools, and their auxiliary facilities, including land for students living in a separate area built for schools.
	A4		Sports land	Land for stadium and sports training bases, excluding land for sports facilities dedicated to schools and other institutions.
	A5		Land for medical and health	Land for medical treatment, health care, sanitation, epidemic prevention, rehabilitation, and first aid facilities.
	A6		Social welfare land	Facilities and ancillary facilities land for providing welfare and charity services to the society.
	A7		Land for historical sites and cultural relics	Ancient sites, ancient tombs, ancient buildings, cave temples, representative modern buildings, revolutionary memorial buildings, and other lands with conservation value. Excluding historic sites and cultural relics that have been used for other purposes.
	A8		Research land	Land for scientific research institutions and their ancillary facilities.
U			Utility land	Land for supply, environment, safety, and other facilities
	U1		Supply facility land	Land for water supply, power supply, gas supply, and heating facilities
	U2		Land for environmental facilities	Land for rainwater, sewage, solid waste treatment, and other environmental protection facilities and their auxiliary facilities.
	U3		Land for safety facilities	Public facilities and ancillary facilities land for such as fire fighting and flood control to protect city safety.
	U4		Land for funeral facilities	Land for funeral parlor, crematoriums, ashes depository, and cemetery.
G			Green space and square land	Public open spaces such as parks, green spaces, and squares.
	G1		Parks and green space	green space opening to the public, with recreation as the main function, and also with the functions of ecology, beautification, and disaster prevention.
	G2		Protective green space	Green space with sanitation, isolation, and safety protection functions.
	G3		Square land	Urban public event venue with functions of recreation, commemoration, assembly, and hedge as the main function.
S			Land for roads and transportation facilities	Land used for urban roads, transportation facilities, etc., excluding residential land, industrial land, and other internal roads, parking lots, etc.
	S1		Urban road land	Land for expressway, main road, secondary road, and branch road, including land at its intersection.
	S2		Urban rail transit land	The above-ground part of the line and site land of the urban rail transit in an independent section
	S3		Land for the transportation hub	Land for railway passenger and freight station, long-distance passenger terminal, passenger port terminal, bus hub, and its affiliated facilities land.
	S4		Land for traffic station	Land for traffic service facilities, excluding land for the traffic command center and traffic police.

References

Akroyd, J., Harper, Z., Soutar, D., Farazi, F., Bhave, A., Mosbach, S., Kraft, M., 2022. Universal digital twin: land use. Data-Centr. Eng. 3, e3.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 2425–2433.

Bao, H., Ming, D., Guo, Y., Zhang, K., Zhou, K., Du, S., 2020. DFCNN-based semantic recognition of urban functional zones by integrating remote sensing data and POI data. Remote Sens. 12.

Barlacchi, G., Lepri, B., Moschitti, A., 2021. Land use classification with point of interests and structural patterns. IEEE Trans. Knowl. Data Eng. 33, 3258–3269.

Bergado, J.R., Persello, C., Stein, A., 2020. Land use classification using deep multitask networks. In: ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 17–21. XLIII-B3-2020.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., Veit, A., 2021. Understanding Robustness of Transformers for Image Classification eprint arXiv: 2103.14586.

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.

CAUPD, C.A.O.U.P.D., 2018. Code for Classification of Urban and Rural Land Use and Planning Standards of Development Land. Ministry of Housing and Urban-Rural Development of the People's republic of China, Beijing.

Chen, M.X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Parmar, N., Schuster, M., Chen, Z., 2018a. The best of both worlds: combining recent Advances in neural machine translation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Chen, W., Huang, H., Dong, J., Zhang, Y., Tian, Y., Yang, Z., 2018b. Social functional mapping of urban green space using remote sensing and social sensing data. ISPRS J. Photogramm. Remote Sens. 146, 436–452.

Chen, Y., Song, Y., Li, C., 2020. Where do people tweet? The relationship of the built environment to tweeting in Chicago. Sustain. Cities Soc. 52.

Chen, B., Tu, Y., Song, Y., Theobald, D.M., Zhang, T., Ren, Z., Li, X., Yang, J., Wang, J., Wang, X., Gong, P., Bai, Y., Xu, B., 2021a. Mapping essential urban land use categories with open big data: results for five metropolitan areas in the United States of America. ISPRS J. Photogramm. Remote Sens. 178, 203–218.

Chen, B., Xu, B., Gong, P., 2021b. Mapping essential urban land use categories (EULUC) using geospatial big data: progress, challenges, and opportunities. Big Earth Data 1–32.

Deng, Y., Chen, R., Yang, J., Li, Y., Jiang, H., Liao, W., Sun, M., 2022. Identify urban building functions with multisource data: a case study in Guangzhou, China. Int. J. Geogr. Inf. Sci. 36, 2060–2085.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

Feng, Y., Huang, Z., Wang, Y., Wan, L., Liu, Y., Zhang, Y., Shan, X., 2021. An SOE-based learning framework using multisource big data for identifying urban functional zones. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 7336–7348.
Galletti, C., Myint, S., 2014. Land-use mapping in a mixed urban-agricultural arid landscape using object-based image analysis: A case study from Maricopa, Arizona. Remote Sens. 6, 6089–6110.
Gong, P., Chen, B., Li, X., Liu, H., Wang, J., Bai, Y., Chen, J., Chen, X., Fang, L., Feng, S., Feng, Y., Gong, Y., Gu, H., Huang, H., Huang, X., Jiao, H., Kang, Y., Lei, G., Li, A., Li, X., Li, X., Li, Y., Li, Z., Li, Z., Liu, C., Liu, C., Liu, M., Liu, S., Mao, W., Miao, C., Ni, H., Pan, Q., Qi, S., Ren, Z., Shan, Z., Shen, S., Shi, M., Song, Y., Su, M., Ping Suen, H., Sun, B., Sun, F., Sun, J., Sun, L., Sun, W., Tian, T., Tong, X., Tseng, Y., Tu, Y., Wang, H., Wang, L., Wang, X., Wang, Z., Wu, T., Xie, Y., Yang, J., Yang, J., Yuan, M., Yue, W., Zeng, H., Zhang, K., Zhang, N., Zhang, T., Zhang, Y., Zhao, F., Zheng, Y., Zhou, Q., Clinton, N., Zhu, Z., Xu, B., 2020. Mapping essential urban land use categories in China (EULUC-China): preliminary results for 2018. Sci. Bull. 65, 182–187.
Graves, A., Fernández, S., Schmidhuber, J., 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja, E., Zadrożny, S. (Eds.), Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005, pp. 799–804.
Häberle, M., Werner, M., Zhu, X.X., 2019. Building type classification from social media texts via geo-spatial textmining. In: IGARSS 2019–2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 10047–10050.
Häberle, M., Hoffmann, E.J., Zhu, X.X., 2022. Can linguistic features extracted from geo-referenced tweets help building function classification in remote sensing? ISPRS J. Photogramm. Remote Sens. 188, 255–268.
He, T., Sun, Y.-J., Xu, J.-D., Wang, X.-J., Hu, C.-R., 2014. Enhanced land use/cover classification using support vector machines and fuzzy k-means clustering algorithms. J. Appl. Remote. Sens. 8.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.
He, J., Li, X., Liu, P., Wu, X., Zhang, J., Zhang, D., Liu, X., Yao, Y., 2021. Accurate estimation of the proportion of mixed land use at the street-block level by integrating high spatial resolution images and geospatial big data. IEEE Trans. Geosci. Remote Sens. 59, 6357–6370.
Hoffmann, E.J., Werner, M., Zhu, X.X., 2019. Building instance classification using social media images. In: 2019 Joint Urban Remote Sensing Event (JURSE), pp. 1–4.
Hoffmann, E.J., Abdulahhad, K., Zhu, X.X., 2022. Using social media images for building function classification arXiv preprint, arXiv:2202.07315.
Hu, S., Wang, L., 2012. Automated urban land-use classification with remote sensing. Int. J. Remote Sens. 34, 790–803.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.
Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269.
Huang, B., Zhao, B., Song, Y., 2018a. Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. Remote Sens. Environ. 214, 73–86.
Huang, R., Taubenböck, H., Mou, L., Zhu, X.X., 2018b. Classification of settlement types from tweets using LDA and LSTM. In: IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 6408–6411.
Huang, Z., Zeng, Z., Huang, Y., Liu, B., Fu, D., Fu, J., 2021. Seeing out of tHe bOx: end-to-end pre-training for vision-language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) eprint arXiv:2104.03135.
Jacob Devlin, M.-W.C., Lee, Kenton, Toutanova, Kristina, 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of NAACL, Minneapolis, Minnesota, pp. 4171–4186.
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.
Jiao, J., Rollo, J., Fu, B., 2021. The hidden characteristics of land-use mix indices: an overview and validity analysis based on the land use in Melbourne, Australia. Sustainability 13.
Kang, J., Körner, M., Wang, Y., Taubenböck, H., Zhu, X.X., 2018. Building instance classification using street view images. ISPRS J. Photogramm. Remote Sens. 145, 44–59.
Karen, S., Andrew, Z., 2014. Very deep convolutional networks for large-scale image recognition arXiv:1409.1556 [cs.CV].
Khorram, S., Brockhaus, J.A., Cheshire, H.M., 1987. Comparson of Landsat MSS and TM data for urban land-use classification. IEEE Trans. Geosci. Remote Sens. GE-25, 238–243.
Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D., 2019. Supervised multimodal bitransformers for classifying images and text arXiv preprint arXiv: 1909.02950.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Curran Associates Inc., Lake Tahoe, Nevada, pp. 1097–1105.
Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.
Li, M., Stein, A., 2020. Mapping land use from high resolution satellite images by exploiting the spatial arrangement of land cover objects. Remote Sens. 12.
Li, C.Y., Yuan, P.C., Lee, H.Y., 2020. What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6434–6438.
Lin, A., Sun, X., Wu, H., Luo, W., Wang, D., Zhong, D., Wang, Z., Zhao, L., Zhu, J., 2021. Identifying urban building function by integrating remote sensing imagery and POI data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 8864–8875.
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. Geogr. Inf. Sci. 31, 1675–1696.
Liu, X., Niu, N., Liu, X., Jin, H., Ou, J., Jiao, L., Liu, Y., 2018. Characterizing mixed use buildings based on multi source big data. Int. J. Geogr. Inf. Sci. 32, 738–756.
Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S., 2020. 12-in-1: multi-task vision and language representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10434–10443.
Lu, W., Tao, C., Li, H., Qi, J., Li, Y., 2022. A unified deep learning framework for urban functional zone extraction based on multi-source heterogeneous data. Remote Sens. Environ. 270.
McGuire, S., 2014. Centers for disease control and prevention. 2013. Strategies to prevent obesity and other chronic diseases: the CDC guide to strategies to support breastfeeding mothers and babies. Atlanta, GA: US Department of Health and Human Services, 2013. Adv. Nutr. 5, 291–292.
Mroueh, Y., Marcheret, E., Goel, V., 2015. Deep multimodal learning for audio-visual speech recognition. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2130–2134.
Ouyang, W., Chu, X., Wang, X., 2014. Multi-source deep learning for human pose estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2337–2344.
Raman, R., Roy, U.K., 2019. Taxonomy of urban mixed land use planning. Land Use Policy 88, 104102.
Rozenstein, O., Karnieli, A., 2011. Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. Appl. Geogr. 31, 533–544.
Santos, A., Canuto, A., Neto, A.F., 2011. A comparative analysis of classification methods to multi-label tasks in different application domains. Int. J. Comput. Inform. Syst. Indust. Manag. Appl. 3, 218–227.
Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition.
Song, Y., Merlin, L., Rodriguez, D., 2013. Comparing measures of urban land use mix. Comput. Environ. Urban. Syst. 42, 1–13.
Song, J., Lin, T., Li, X., Prishchepov, A.V., 2018. Mapping urban functional zones by integrating very high spatial resolution remote sensing imagery and points of interest: a case study of Xiamen, China. Remote Sens. 10.
Srivastava, S., Vargas-Muñoz, J.E., Swinkels, D., Tuia, D., 2018a. Multilabel building functions classification from ground pictures using convolutional neural networks. In: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, pp. 43–46.
Srivastava, S., Vargas Muñoz, J.E., Lobry, S., Tuia, D., 2018b. Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. Int. J. Geogr. Inf. Sci. 34, 1117–1136.
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.
Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J., 2020. Vl-bert: pre-training of generic visual-linguistic representations. In: ICLR 2020.
Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9.
Theobald, D.M., Kennedy, C., Chen, B., Oakleaf, J., Baruch-Mordo, S., Kiesecker, J., 2020. Earth transformed: detailed mapping of global human modification from 1990 to 2017. Earth Syst. Sci. Data 12, 1953–1972.
Urbanism, C.F.T.N., 2000. Charter of the new urbanism. Bull. Sci. Technol. Soc. 20, 339–341.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017).
Workman, S., Zhai, M., Crandall, D.J., Jacobs, N., 2017. A unified model for near and remote sensing. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2707–2716.
Wu, Z., Jiang, Y.-G., Wang, J., Pu, J., Xue, X., 2014. Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: Proceedings of the 22nd ACM International Conference on Multimedia. Association for Computing Machinery, Orlando, Florida, USA, pp. 167–176.
Xia, H., Liu, Z., Efremochkina, M., Liu, X., Lin, C., 2022. Study on city digital twin technologies for sustainable smart city design: A review and bibliometric analysis of geographic information system and building information modeling integration. Sustain. Cities Soc. 84.
Xie, S., Girshick, R., Dollar, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5987–5995.
Yue, Y., Zhuang, Y., Yeh, A.G.O., Xie, J.-Y., Ma, C.-L., Li, Q.-Q., 2017. Measurements of POI-based mixed use and their relationships with neighbourhood vibrancy. Int. J. Geogr. Inf. Sci. 31, 658–675.
Zhan, X., Ukkusuri, S.V., Zhu, F., 2014. Inferring urban land use using large-scale social media check-in data. Netw. Spat. Econ. 14, 647–667.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018a. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70.
Zhang, X., Du, S., Wang, Q., 2018b. Integrating bottom-up classification and top-down feedback for improving urban land-cover and functional-zone mapping. Remote Sens. Environ. 212, 231–248.
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.
Zhang, C., Benz, P., Argaw, D.M., Lee, S., Kim, J., Rameau, F., Bazin, J.C., Kweon, I.S., 2021a. ResNet or DenseNet? Introducing dense shortcuts to ResNet. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 3549–3558.
Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J., 2021b. Vinvl: making visual representations matter in vision-language models. arXiv preprint arXiv:2101.00529, 1, p. 8.
Zhong, Y., Su, Y., Wu, S., Zheng, Z., Zhao, J., Ma, A., Zhu, Q., Ye, R., Li, X., Pellikka, P., Zhang, L., 2020. Open-source data-driven urban land-use mapping integrating point-line-polygon semantic objects: A case study of Chinese cities. Remote Sens. Environ. 247.
Zhou, W., Ming, D., Lv, X., Zhou, K., Bao, H., Hong, Z., 2020. SO–CNN based urban functional zone fine division with VHR remote sensing image. Remote Sens. Environ. 236, 111458.
Zhu, Y., Deng, X., Newsam, S., 2019. Fine-grained land use classification at the city scale using ground-level images. IEEE Trans. Multim. 21, 1825–1838.
Zhu, X.X., Wang, Y., Kochupillai, M., Werner, M., Häberle, M., Hoffmann, E.J., Taubenböck, H., Tuia, D., Levering, A., Jacobs, N., 2022. Geo-Information Harvesting from Social Media Data arXiv preprint arXiv:2211.00543.