Remote Sensing of Environment 311 (2024) 114290

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Review

Deep learning for urban land use category classification: A review and experimental assessment

Ziming Li[a], Bin Chen[a,b,c,*], Shengbiao Wu[a], Mo Su[d], Jing M. Chen[e,f], Bing Xu[g,**]

[a] *Future Urbanity & Sustainable Environment (FUSE) Lab, Division of Landscape Architecture, Faculty of Architecture, The University of Hong Kong, Hong Kong SAR, China*
[b] *Urban Systems Institute, The University of Hong Kong, Hong Kong SAR, China*
[c] *HKU Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China*
[d] *Shenzhen Urban Planning and Land Resource Research Center, Shenzhen 518034, China*
[e] *Key Laboratory for Humid Subtropical Eco-Geographical Processes of the Ministry of Education, School of Geographical Sciences, Fujian Normal University, Fuzhou 350117, China*
[f] *Department of Geography and Planning, University of Toronto, Toronto, Ontario M5S 3G3, Canada*
[g] *Department of Earth System Science, Ministry of Education Ecological Field Station for East Asian Migratory Birds, and Institute for Global Change Studies, Tsinghua University, Beijing 100084, China*

ARTICLE INFO

Editor: Marie Weiss

Keywords:
Urban land use type
Deep learning
Remote sensing
Geospatial big data
Sample

ABSTRACT

Mapping the distribution, pattern, and composition of urban land use categories plays a valuable role in understanding urban environmental dynamics and facilitating sustainable development. Decades of effort in land use mapping have accumulated a series of mapping approaches and land use products. New trends characterized by open big data and advanced artificial intelligence, especially deep learning, offer unprecedented opportunities for mapping land use patterns from regional to global scales. Combined with large amounts of geospatial big data, deep learning has the potential to promote land use mapping to higher levels of scale, accuracy, efficiency, and automation. Here, we comprehensively review the advances in deep learning based urban land use mapping research and practices from the aspects of data sources, classification units, and mapping approaches. More specifically, delving into different settings on deep learning-based land use mapping, we design eight experiments in Shenzhen, China to investigate their impacts on mapping performance in terms of data, sample, and model. For each investigated setting, we provide quantitative evaluations of the discussed approaches to inform more convincing comparisons. Based on the historical retrospection and experimental evaluation, we identify the prevailing limitations and challenges of urban land use classification and suggest prospective directions that could further facilitate the exploitation of deep learning techniques in urban land use mapping using remote sensing and other spatial data across various scales.

1. Introduction

Land use and land cover (LULC) are critical factors that determine climate and ecosystem changes through biophysical and biochemical processes (Foley et al., 2003; He et al., 2017). It has been widely recognized that LULC patterns and changes are important factors that affect the energy balance, hydrological cycle, carbon dynamics, and ecosystem services (Findell et al., 2017; Foley et al., 2005; Schilling et al., 2008; Schneider et al., 2010; Tang et al., 2021). The intensified human modifications to LULC, particularly the rapid urbanization over the past few decades, have resulted in prominent impacts on natural systems and human society, including biodiversity loss, food insecurity, natural hazards, poverty, public health, and human well-being (Gong et al., 2012; Liu et al., 2021a; Simkin et al., 2022; Szabo, 2016; Zhang et al., 2018b). Mapping and monitoring the LULC pattern and associated dynamics accurately and timely are of great significance to advance our understanding of the cause and effect of global environmental changes and to facilitate effective and sustainable land management.

Many efforts have been made to enable spatially and temporally explicit LULC mapping, thereby providing essential data support for

* Corresponding author at: Future Urbanity & Sustainable Environment (FUSE) Lab, Faculty of Architecture, The University of Hong Kong, Hong Kong SAR, China.
** Corresponding author at: Department of Earth System Science, Tsinghua University, Beijing 100084, China
E-mail addresses: binley.chen@hku.hk (B. Chen), bingxu@tsinghua.edu.cn (B. Xu).

https://doi.org/10.1016/j.rse.2024.114290
Received 24 March 2024; Received in revised form 30 May 2024; Accepted 22 June 2024
Available online 14 July 2024

outlining the spatial patterns and temporal dynamics of land surface conditions. Representative datasets includes (1) coarse resolution products with a spatial resolution ranging from 100 m to 1 km, such as the Global Land Cover Characterization (GLCC) datasets (Loveland et al., 2000), University of Maryland Global Land Cover (UMD-GLC) maps (Hansen et al., 2000), Global Land Cover 2000 (GLC2000) dataset (Bartholomé and Belward, 2005), Moderate Resolution Imaging Spectrometer (MODIS) land cover datasets (Friedl et al., 2002), global land cover (GlobCover) datasets (Arino et al., 2007); (2) medium resolution products with spatial resolutions ranging from 10 m to 100 m, GlobalLand30 (Chen et al., 2015), FROM-GLC30 (Gong et al., 2013), FROM-GLC10 (Gong et al., 2019), Global Urban Land (Liu et al., 2018b), WorldCover-10 m products (Zanaga et al., 2022); and (3) high resolution products with spatial resolutions finer than 10 m, such as the 3-m China land cover map derived from Planet images (Dong et al., 2022) and the 1-m resolution national-scale land-cover map of China (SinoLC-1) (Li et al., 2023b). The majority of these available datasets focus on land cover types since the physical attributes can be well captured and differentiated by satellite sensors. However, the products of large-scale land use classification with fine-grained classification schemes are still very limited, especially for urban areas. Compared with land cover, land use contains more socio-economic properties and is more challenging to interpret (Fang et al., 2022; Huang et al., 2020b). As the highest-level human modification of the land, urban land use reflects the interaction between human activities and heterogeneous urban landscapes, as a result of the objective and outcome of exploiting land resources (Gong et al., 2020). To date, more than half of the world's population lives in urban areas, and this number is expected to increase to 68% by 2050 (United Nations Department of Economic Social Affairs, 2019). Therefore, mapping the detailed urban land use categories, such as the distribution and composition of residential, commercial, industrial, and public service lands, is critically important. Such information can provide direct data support and reference to city governors, urban planners, and landscape architects for designing, building, and managing a livable and sustainable urban environment. Moreover, the availability of urban land use category information can further benefit a wide range of urban environmental applications and implications, such as exploring urban environmental problems, monitoring urban morphology changes, simulating urbanization process, and informing human behaviors and public health studies (Chen et al., 2021b).

Initially, the urban land use mapping was mainly conducted through on-site land surveys, which were highly inefficient, labor-intensive, and severely limited its geographical extent (DeVries, 1928). With the rapid development of satellite techniques and the growing amounts of Earth observation platforms since the 1970s, remote sensing data with various spatial, spectral, and temporal resolutions have become more available for understanding multi-scale and multi-dimensional urban environments. Remote sensing data record the electromagnetic signals of observed objects, including the reflection, emission, and scattering, which are mainly determined by the nature of objects' materials. Such physical signals are inherent indicators of differentiating land cover types and thus making remote sensing data widely used in generating LULC maps (Joshi et al., 2016). Before 2000, satellite data such as multispectral data (e.g., Landsat and SPOT) and microwave data (e.g., ERS) were commonly used in land use classification research, because of the large coverage, data availability, and relatively high spatial resolution at that time (Barnsley and Barr, 1996; DeVries, 1928; Gong and Howarth, 1992; Pathan et al., 1989). However, due to the large number of mixed pixels, serious spectral confusion, and heterogeneous urban environment impeding the land use classification, the performance did not achieve substantial improvement for decades (Wilkinson, 2005; Wu and Murray, 2003). It was concluded that making efforts to optimize modeling methods is of little value (Manandhar et al., 2009; Rozenstein and Karnieli, 2011). Instead, researchers seek to develop and utilize more advanced and powerful remote sensing data or other alternative data. High-resolution (HR) and very-high-resolution (VHR) remotely sensed data (e.g., GeoEye, Worldview, QuickBird, Gaofen series, etc.), can differentiate subpixel information of ground objects, making it possible to transform land cover classification to land use classification since more potentially useful features such as texture, geometry, size, adjacency, and contextual information can be exploited from HR and VHR data. The other types of remote sensing data including Light Detection And Ranging (LiDAR), nighttime light data, and hyperspectral data were also introduced into the mapping framework since they can collect distinct information on the observed objects. Advances in the Internet of Things and mobile internet technologies also provided feasible alternatives to conduct land use category mapping by leveraging social sensing big data via various kinds of ubiquitous sensors including webcam, mobile phone, fixed monitoring appliances, etc. (Liu et al., 2015; Wang et al., 2019). These emerging sensors facilitated a substantial reduction in the cost of data acquisition and provided new data sources for reflecting socioeconomic attributes and human dynamics (Guo et al., 2024). For example, many previous research studies have utilized point of interest (POI) (Liu and Long, 2016), social media data (e.g. Twitter, Facebook, Weibo, etc.) (Frias-Martinez and Frias-Martinez, 2014; Steiger et al., 2015; Tu et al., 2017), mobile phone data (Pei et al., 2014), smart card data (Long and Shen, 2015; Wang et al., 2021) and GPS trajectories (Pan et al., 2013), to map urban land use categories in the built-up areas. Compared with the well-structured remote sensing images, the structure and organization of social sensing data are of higher complexity and diversity. Subjective bias and uncertainty of individual behavior and human interactions also impede the applicability of social sensing data (Galesic et al., 2021). It is therefore challenging to extract useful socio-economic attributes and human activity patterns from these data (Guan et al., 2021; Yao et al., 2022; Zhang et al., 2015). Despite the notable advantage of social sensing data in capturing socioeconomic characteristics and reflecting human activities, researchers did not exclusively depend on these data for land use classification. The ongoing advancements in remote sensing and social sensing technologies are creating more opportunities to represent the urban space comprehensively. A growing number of recent studies are integrating remote sensing and social sensing data so that physical and socio-economic characteristics can be fully exploited and incorporated to differentiate complicated land use categories in the urban area (Du et al., 2020; Liu et al., 2017; Wang et al., 2023; Xie et al., 2024).

Besides on-site surveys, early urban land use mapping could also rely on visual interpretation of air-borne or space-born images. However, this manual process had the drawbacks of being highly subjective and demanding considerable labor and time. To meet the growing demands for up-to-date urban land use information, more automatic and efficient methods for urban land use classification have subsequently been developed (Barnsley and Barr, 1996; Solberg et al., 1994). The advancements in artificial intelligence and computing technologies provide considerable potential for automating urban land use mapping. Machine learning based classification algorithms such as Maximum Likelihood Estimation (MLE), Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP), have been increasingly adopted in remote sensing community, given their capability to create optimal models in a data-driven approach (Dixon and Candade, 2007; Garg et al., 2021; Li et al., 2019a; Paola and Schowengerdt, 1995; Wang et al., 2022). These shallow classification models require meticulous feature engineering to render accurate decisions for classification tasks. Traditional methods typically extract the spectral, geometry, texture, and other shallow visual characteristics using feature construction approaches such as scale-invariant feature transform (SIFT), histogram of oriented gradient (HOG) and Gray-level cooccurrence matrix (GLCM) (Dalal and Triggs, 2005; Guo et al., 2023b; Haralick et al., 1973; Lowe, 2004). These features often fall short in representing distinctive and universal characteristics of complex terrestrial features, thereby undermining the generalizability and discrimination accuracy of the downstream tasks (Huang et al., 2018; Li et al., 2019a). Alongside the growth of big data, a series of data mining techniques, especially topic modeling such as Bag of Visual Words (BOVW), Latent Dirichlet allocation (LDA), and probabilistic Latent Semantic Analysis (pLSA), were used to exploit and fuse higher-level semantic information in remote sensing and social sensing data (Blei et al., 2003; Hofmann, 1999; Yang and Newsam, 2010). Due to the lack of spatial correlations in such topic modeling, inspired by the word embedding in natural language processing, Word2Vec, Place2Vec, Block2Vec, and Traj2Vec were developed to model potential spatial relationships and embed them into representative vectors (Sun et al., 2021; Yan et al., 2017; Yao et al., 2016; Zhang et al., 2020b). Recently, deep learning models that utilize neural networks with deep architectures have rapidly evolved, shifting the focus of the research community and making them become the mainstream algorithms in computer vision and natural language processing due to their robust capacity to implement representative learning in an end-to-end way (Arel et al., 2010; LeCun et al., 2015). Given its remarkable performance, deep learning has been extensively utilized in the fields of remote sensing and earth system science since 2014, including data fusion, scene understanding, and LULC mapping and change detection (Ma et al., 2019; Reichstein et al., 2019; Zhu et al., 2017). Advanced deep network architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and Transformers, have been employed in Earth observation and environmental monitoring applications, significantly outperforming traditional models (Huang et al., 2018; Ansith and Bini, 2022; Scheibenreif et al., 2022; Xiao et al., 2022). Fig. 1 illustrates the difference between traditional machine learning and deep learning methods when executing land use classification. In traditional machine learning, the process of feature extraction for classification relies heavily on human-designed feature engineering, due to the classification model's limited capability of learning features independently. On the other hand, deep learning methods minimize the need for human intervention in feature engineering, as their deep network layers are able to extract multi-level features through data-driven learning from massive samples. Within the context of mapping urban land use categories, deep learning models can effectively extract latent semantic information and investigate the relationships between different objects. This capability significantly aids in differentiating complex urban land use categories characterized by high intraclass variance and low interclass variance. With their significant advantages, deep learning-based methods have increasingly contributed to urban land use classification studies in recent years (Zang et al., 2021). In the initial stage of applying deep learning in urban land use classification, the pre-trained deep models were directly utilized as a feature extractor to derive hierarchical features for classification (Hu et al., 2015; Marmanis et al., 2016; Zhao et al., 2017a). This approach, however, faced the challenges of the significant mismatch of the model architectures, data source, and pre-trained tasks, limiting the effectiveness of deep learning. Subsequently, a series of studies proposed more effective model architectures tailored for specific data and tasks, constructed task-specific datasets, and retrained their models (Huang et al., 2018; Nogueira et al., 2017; Zhang et al., 2020a; Zhang et al., 2018a). As a result, plenty of the variants of foundational deep learning models have been proposed and compared for urban land use mapping. These studies focused on designing model architectures that could enhance models' capability of extracting discriminative features from specific types of data sources. Currently, there is a noticeable trend toward more effective architecture and strategy for multimodal applications that integrate multisource remote sensing and social sensing data for capturing complexities and diversity of land use patterns within urban areas (Lu et al., 2022; Su et al., 2024; Yan et al., 2024; Yu et al., 2023). Relevant research studies work on addressing the following three challenges: feature representation for each modality, feature alignment between modalities, and feature fusion of multiple modalities (Guo et al., 2024; Li et al., 2023a; Ouyang et al., 2023).

Despite substantial improvements achieved by leveraging deep learning approaches, the field remains in an emerging and evolving state, necessitating regular reviews to summarize advancements, pinpoint challenges, and suggest future directions. While many reviews have explored deep learning's applications, the majority have focused on broader missions or subjects, overlooking the specifics of urban land use (Chen et al., 2023a; Ma et al., 2019; Yuan et al., 2020; Zhu et al., 2017). Urban land use, shaped by both human and natural influences, presents distinct challenges for deep learning applications, including a variety of data sources, mapping units, and classification models. The optimal strategies for employing deep learning in mapping urban land use categories, such as choosing the appropriate analysis units, models, data sources, and parameter configurations for classification practices, remain unclear. Therefore, it is crucial to review the characteristics of

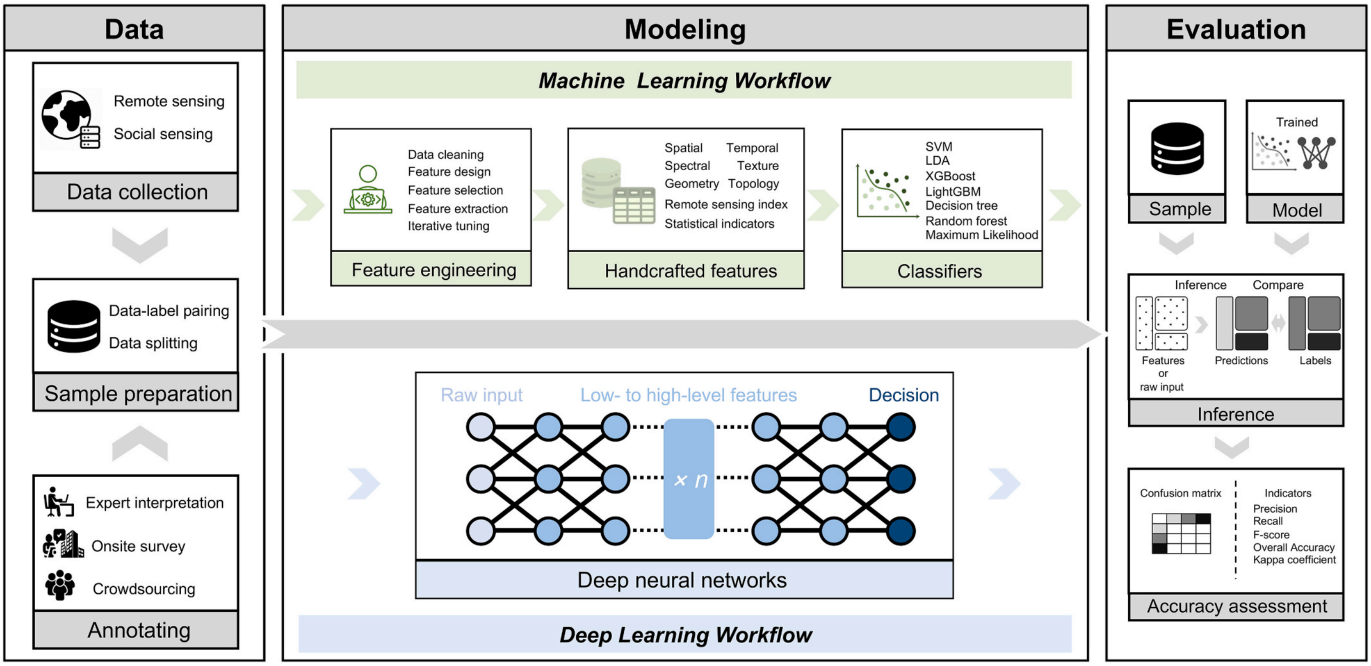

Fig. 1. Workflow comparison between traditional machine learning and deep learning methods for urban land use category classification with geospatial big data.

deep learning in urban land use category classification from existing studies, to guide future research in accurately identifying land use categories and understanding the functional patterns of urban land uses.

In this context, this study carries out a comprehensive review and comparative analysis of urban land use classification, focusing on recent advancements that utilize deep learning algorithms. Specifically, a full set of experimental assessments is conducted on the ubiquitous Shenzhen datasets to investigate the sensitivity of crucial hyperparameters in major deep learning models. Based on literature review and experimental results, we identify ongoing challenges and future directions for deep-learning-based urban land use classification. Its primary goal is to serve as a comprehensive guide for mapping large-scale, detailed urban land use categories with deep learning models, addressing the following three main questions: (1) What kinds of data sources, mapping units, and deep learning models have been adopted in the existing methodological framework, and how have they been utilized for the urban land use mapping tasks? (2) In the classification practices, how do variations in parameter settings in terms of data, models and samples influence model performance, especially for deep learning driven urban land use classification? (3) What are the remaining key challenges and prospective opportunities for future investigation and applications of deep learning in urban land use classification?

The rest of this review is structured as follows: Section 2 presents an overview of research on deep-learning-based land use mapping, across the spectrum of data sources, classification units, and fundamental deep learning frameworks. Section 3 outlines extensive experiments of evaluating the effectiveness and differences of deep learning models in urban land use mapping. Section 4 discusses current challenges and future opportunities. Section 5 marks conclusions on key findings and their implications.

2. Advances of deep learning in land use mapping

2.1. Data source

Diverse geospatial big data including remote sensing and social sensing data (Fig. 2), are produced and collected every day. Appropriate data with high-quality information directly determine the classification performance. The purpose of this section is thus to summarize the commonly used remote sensing and social sensing big data in recent studies applying deep learning in urban land use classification.

2.1.1. Remote sensing

2.1.1.1. Multispectral remote sensing. Multispectral remote sensing sensors generate multi-band images, usually with 3–10 bands, and each band records the radiation of electromagnetic waves at specific wavelengths reflected or emitted from the observed surfaces. In general, multispectral imaging systems capture broadband spectral responses from visible to infrared wavelengths, typically covering a range of 10–100 nm. By analyzing the values of different bands in multispectral images, several key land cover types with significant inter-class differences (e.g., water, vegetation, soil, and built-up area) can be distinguished due to their distinct spectral characteristics, providing an essential delineation of the physical environment (Friedl et al., 2002; Verpoorter et al., 2012; Xie et al., 2008; Xu, 2008). Moreover, the large coverage and rich historical records of multispectral remote sensing data facilitate long-term monitoring and analysis of land use change across large regions over decades, making it possible to track and understand the trend of ecological and urban development.

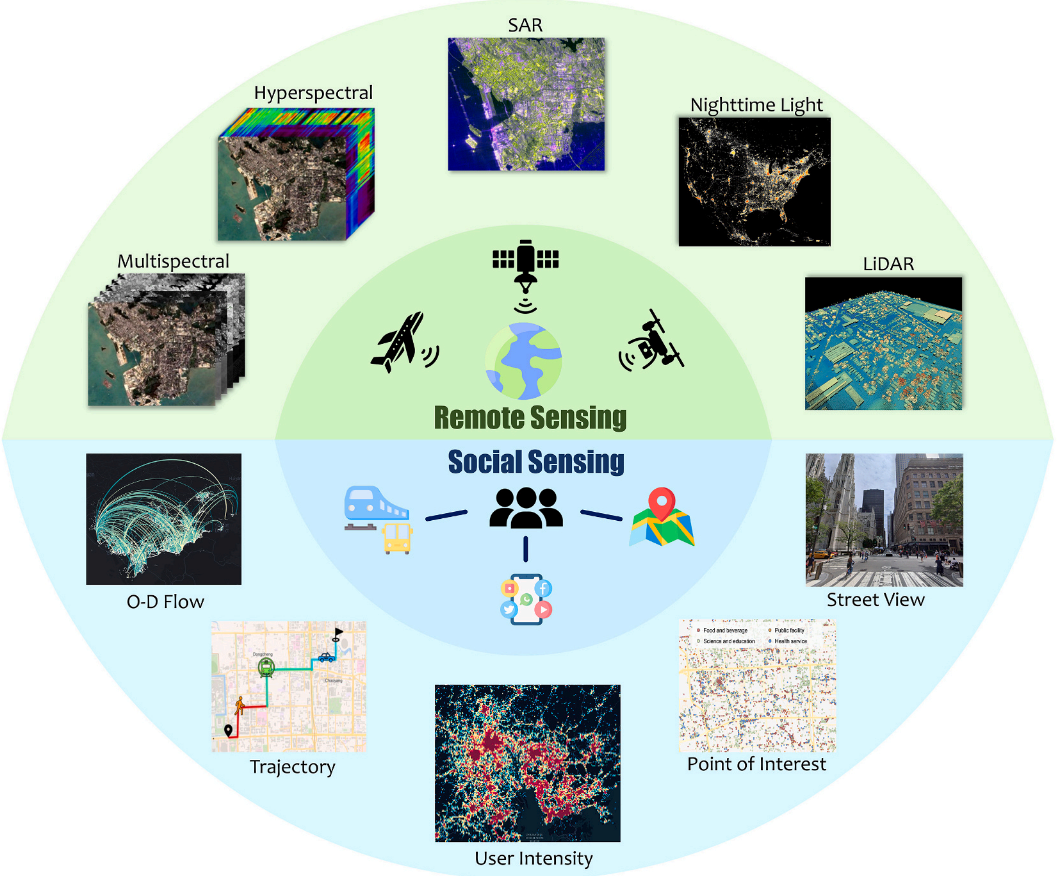

Fig. 2. Examples of common geospatial big data in terms of remote sensing and social sensing.

2.1.1.2. Hyperspectral remote sensing. Built upon imaging spectroscopy, hyperspectral sensors detect the contiguous electromagnetic energy with higher spectral resolutions. The capability of capturing high levels of spectral detail enables hyperspectral data to identify materials and derive biophysical parameters with superior accuracy compared to using multispectral satellite images (Yokoya et al., 2017). In general, hyperspectral sensors produce data with hundreds of bands with narrow bandwidths (<10 nm) in visible-to-NIR region (Jafarbiglu and Pourreza, 2022). Because of the strength in discrimination of subtle differences between objects, many applications have used hyperspectral data to classify LULC categories (Chen et al., 2017a; Kalluri et al., 2010; Xu and Gong, 2007; Yuan et al., 2022). In urban areas characterized by high heterogeneity and complexity, hyperspectral data is able to identify the infrastructure materials, vegetation species, and even their conditions (e.g., aging and health) in finer granularity (Abbas et al., 2021; Heiden et al., 2012). Such sensitivity to manmade and natural objects makes it a more powerful indicator than multispectral data for detailed urban land use classification. The commonly used data sources include EO-1 Hyperion, PROBA-1 CHRIS, HJ-1 A, GF-5, PRISMA, etc., most of which were of moderate spatial resolutions. Due to the costly acquisition and insufficient resolution, the hyperspectral images are limitedly used to investigate the spatial pattern and dynamic of LULC across large scales.

2.1.1.3. Nighttime light remote sensing. Nighttime light observed from space is closely related to the artificial lights from human activities. To monitor nighttime lights over large spatial scales, a number of space-borne sensors were developed over the past decades, including the Defense Meteorological Satellite Program's Operational Line-scan System (DMSP/OLS), the Visible Infrared Imaging Radiometer Suite (VIIRS) instrument onboard the Suomi National Polar Partnership satellite (NPP) fitted with Day and Night Band (DNB), and the Cubesat Multispectral Observing System (CUMULOS), LUOJIA-1 and SDGSAT-1(Guo et al., 2023a). It has been widely recognized as an effective proxy for investigating human activities and deriving socioeconomic variables, which are hardly achievable by the aforementioned daytime remote sensing imagery. Meanwhile, unlike other social sensing data, nighttime light data exhibits superior strength in wide coverage and spatial continuity. The synthesis of daytime remote sensing data and nighttime light data has thereby become a common combination in characterizing urban land use categories from a more holistic perspective (Chen et al., 2021a; Chen et al., 2023b; Huang et al., 2021; Tu et al., 2020). However, the coarse and medium resolution (from 130 m to 1 km) data generated from the abovementioned sensors may undermine the mapping performance, especially for fine-grained pixel-based classification. To overcome this limitation, the Israeli EROS-B (Levin et al., 2014) and the Chinese JL1-3B (Jilin-1) satellite (Zheng et al., 2018) offer decimeter-level nighttime light data, enabling to investigate human activities and urban land use compositions in finer details from the space.

2.1.1.4. Microwave remote sensing. Microwave remote sensing records passive or active electromagnetic waves with wavelengths ranging from 1 mm to 1 m. Radar is the most common form of imaging active microwave sensors, with strength in all-weather and day or night imaging (Moreira et al., 2013). Synthetic Aperture Radar (SAR) provides finer spatial resolutions than conventional radar by artificially manufactured antenna systems to utilize the Doppler Effect (Javali et al., 2021). In recent years, an increasing number of high-resolution or very-high-resolution SAR sensors were developed including Gaofen-3, HJ-1C, Cosmo-Skymed, ALOS PALSAR, ALOS-2, RADARSAT-2, TerraSAR-X and Sentinel-1. Due to the potential to characterize complex urban structures and reveal multi-dimensional geometrical features, microwave sensors are also important data sources in differentiating detailed land use types (Chen et al., 2013; Schulz et al., 2021). The recorded backscattering coefficient and phase information in SAR data can reflect the urban structural information, orientation, shape, roughness, and height(Chen et al., 2013; Frantz et al., 2021). Compared with optical remote sensing, another strength of SAR lies in the capability of all-day, all-weather observation and a certain degree of penetration, contributing to the broad applications in cloud-prone regions(Vaglio Laurin et al., 2013). The joint use of optical and microwave data provides complementary information on the physical surface.

2.1.1.5. LiDAR. Light Detection and Ranging (LiDAR) is a popular active remote sensing method used for measuring varying distances of objects from the Earth's surface by pulsed lasers and obtaining accurate three-dimensional information about the object. Different from the optical remote sensing imagery, the LiDAR point cloud possesses the intrinsic strength in capturing precise structural information, robustness to light conditions, no relief displacement, and penetration of tree canopies (Yan et al., 2015). Therefore, it has been frequently applied to land cover and land use mapping (Mo et al., 2024; Pan et al., 2020b; Zhong et al., 2017). Compared with SAR data, which also shows potential in describing three-dimensional information, LiDAR data mostly has higher resolution and provides more detailed and precise measurements of the target. The urban structure parameters extracted from LiDAR data play important roles in delineating urban layouts from 2D to 3D and even extracting building-level functions, especially in megacities with substantial three-dimensional components (Ma et al., 2015a). The LiDAR-derived parameters such as height, intensity, skewness, and sky view factors have been found to be useful in differentiating different land use categories (Man et al., 2015; Sanlang et al., 2021). However, limitations related to spatial coverage, data availability, data cost, and updating temporal frequency impede the widespread application of LiDAR data for large-scale land use classification.

2.1.2. Social sensing

With the emergence of big data and Information and Communication Technologies (ICT), social sensing has become a new and promising approach to quantify and understand socioeconomic environments (Liu et al., 2015). Social sensing big data have the potential to trigger novel datasets, approaches, tools, and insights to characterize spatiotemporal patterns of human activities, serving as an important complement to remote sensing data (Chen et al., 2021b). Given the multi-source and multi-modal nature of social sensing data, the challenge of bridging the modality gap and mitigating the heterogeneity issue is significant. Deep learning, known for its capability of automatically learning representative features, presents a promising approach to integrating diverse data sources and types for a better understanding of the human-environment relationship.

2.1.2.1. Social media. Social media provides a unique lens for portraying users' spatiotemporal preferences and mobility patterns (Chen et al., 2019). Social media applications record tons of geotagged information every day (Ilieva and McPhearson, 2018; Shen and Karimi, 2016). For example, Twitter data have been viewed as a feasible data source for characterizing urban land use category (Soliman et al., 2017). Weibo check-in data and the Tencent user density data are two primary data used to reveal the population dynamics and identify urban functions in China due to their large user base (Chen et al., 2017b). Besides the spatial and temporal information provided by the location-based service network, geotagged images and texts on social media also contain rich semantics for understanding the interplay between users and their nearby environments (Häberle et al., 2019; Hoffmann et al., 2023). By leveraging advanced deep learning techniques for effective image analysis and natural language processing, this rich geotagged semantic content can be transformed into representative features for further analysis. Additionally, Point-of-interest (POI) data emerges as another popular data source owing to its widespread availability and detailed socioeconomic activities occurring, representing micro-level land use patterns within areas (Huang et al., 2024; Liu et al., 2020;

Yang et al., 2022). These user-generated data unveil the intensity and modalities of users' interaction with urban spaces, offering valuable insights into the socioeconomic attributes of regions. Consequently, such implicit knowledge enables social media data to serve as an alternative or complementary approach to exploring how cities function (Zhou and Zhang, 2016).

2.1.2.2. Mobile device data. Mobile device data have been validated as a more direct means of capturing human activity information, contributing to a better understanding of mobility patterns and social dynamics in urban areas(Gao et al., 2024; Sun et al., 2022). Currently, the mobile phone is the most common mobile device due to its widespread use and portability. It serves as an ideal proxy to uncover individual human mobility patterns, offering insights into the spatial organization of human networks (Ríos and Muñoz, 2017). Cellular networks, composed of geographic zones around phone towers, can accurately locate each mobile phone call using network-based positioning methods (Deville et al., 2014). Mobile phone records detailing location and time have been recognized as good proxies for revealing human activity patterns and inferring urban land use (Toole et al., 2012; Widhalm et al., 2015). Additionally, vehicle-based and GPS data are another important source for urban land use mapping because they offer high spatiotemporal accuracy in capturing individual human movements (Hu et al., 2021; Liu et al., 2012). Given the intrinsic interdependence between land use and traffic patterns, mobile device data can benefit the exploration of urban land use structures by excavating the implicit spatial and temporal rhythm of human mobility (Liu et al., 2012; Zhang et al., 2018c).

2.1.2.3. Proximate sensing. Different from the overhead images from spaceborne or airborne sensors, proximate sensing provides ground-level, georeferenced visuals of nearby objects and scenes (Leung and Newsam, 2010; Qiao and Yuan, 2021). These images captured from perspectives different from remote sensing data in very high resolution can describe more detailed, representative, and heterogeneous visual characteristics related to urban land use attributes (Wu et al., 2023). Street view imagery, the most popular type of proximate sensing data, has become an important source for urban analysis. Services like Google Street View, Bing StreetSide, Mapillary, Tencent Street View, and Baidu Total View lead this domain with their extensive spatial coverage (Biljecki and Ito, 2021). Compared with overhead images, these street-level images offer not only an intimate view of the built environment but also great potential to understand the socioeconomic attributes(Fan et al., 2023; Gebru et al., 2017). The latent socioeconomic characteristics are mostly inferred from the composition and configuration of fine-grained urban objects that can only be recognized and discriminated by the ground-level high-resolution image (Zhao et al., 2022a). Therefore, these images have been widely used in identifying land use characteristics of different scales (Li et al., 2017; Zhang et al., 2023). However, their applications are also with practical challenges including issues related to image quality, obstructions, uneven coverage, as well as variability in availability and update frequency (Bin et al., 2020; He et al., 2020; Hou and Biljecki, 2022).

2.1.2.4. Volunteered geographic information. Volunteered geographic information (VGI) refers to the geospatial data provided voluntarily through crowdsourcing (Goodchild, 2007). OpenStreetMap (OSM) stands out as the most prevailing VGI database because of its global coverage, high flexibility in data importation, and open access to the latest volunteered contributions (Flanagin and Metzger, 2008; Koukoletsos et al., 2012). The rich geographic features in OSM have intrigued many researchers to leverage it for mapping urban land use patterns, demonstrating its utility in urban studies (Gong et al., 2020; Johnson et al., 2022). However, since the information was collected from the volunteers regardless of their varying expertise and background, VGI data suffers from the problem of inconsistent quality and completeness, potentially undermining its reliability (Bordogna et al., 2014; Senaratne et al., 2017; Vargas-Munoz et al., 2021).

2.1.2.5. Other data sources. Beyond remote sensing and social sensing data, various auxiliary datasets can also contribute to urban land use category mapping. Census data, for instance, provide official statistics on a range of demographic factors including migration, education, health, and employment. These data can directly represent the socioeconomic attributes of different census units, providing essential insights into urban land zonings and land use classifications (Theobald, 2014). Additionally, municipal services data recording the consumption of resources in daily life, are emerging as new sources for delineating human activities over time and space. With the advantages of high spatiotemporal resolutions, comprehensive coverage of population, and long timespan, datasets such as time-series of water and electricity usage have the potential to analyze urban land use patterns (Guan et al., 2021; Pan et al., 2020c; Yao et al., 2022).

2.2. Mapping units

The spatial patterns and phenomena observed within a region can considerably vary depending on the type of mapping units used, making the selection of these units critical to the purpose. In the field of land use mapping, this variation is particularly evident, where three hierarchical levels of mapping units including the pixel, object, and scene, are commonly used as basic spatial entities to represent homogeneous landscapes.

The size of pixel-level land use units is determined by the spatial resolution of raster data, predominantly from remote sensing imagery. With finer spatial resolutions, the homogeneity within pixel-level units also gets more refined. The pixel stands as the most fundamental one among the three mapping units, offering the advantage of scalability. This means that pixels can be aggregated to represent larger land surfaces, a process less straightforward for higher-level mapping units due to the absence of specific, finer-scale details. In this sense, pixel-level land use mapping holds great potential. However, it also presents key challenges, especially in urban areas, where the high variability within classes (i.e., intra-class differences) and the minimal difference between classes (i.e., inter-class differences) can render traditional classification methods ineffective. To address these problems, recent years have witnessed a surge of research into deep learning-based segmentation techniques specifically for pixel-level land use mapping.

The object-level unit composed of pixels, represents image objects that are meaningful entities with homogeneous attributes, which are differentiable at a certain scale in the image (Blaschke et al., 2014). These object-level units are generated and identified through segmenting objects from remote sensing images using object-based image analysis (OBIA) methods. Over the past decade, OBIA has become a dominant approach in land use mapping, offering several advantages over pixel-based methods, including improved classification accuracy and a better representation of the real world's heterogeneous features (Whiteside et al., 2011; Yu et al., 2006). While many studies have successfully applied OBIA methods to produce land use maps for local and regional studies, extending these applications to larger datasets for mapping at continental and global scales remains a challenge. This is due to the spatial scale effect and the requirement of effective human intervention and experiment (Myint et al., 2011; Zhang et al., 2022). Moreover, from a practical perspective, the use of such segmented units cannot be easily applied in real-world applications like land use planning and resource management (Zhong et al., 2020). Besides, buildings, as representative objects carrying most of the urban functions, are also widely used as analysis units of urban land use mapping (Du et al., 2024; Zheng et al., 2024).

The scene-level unit represents specific extents defined by human interpretations, encompassing socio-economic contexts such as traffic analysis zones, cadastral fields, and street blocks or simple grids of a certain size. Grids, in particular, have gained popularity in the development of land use classification models based on deep learning due to their regular shape and straightforward preprocessing. Several scene-based remote sensing land use datasets, such as UC Merced Land Use Dataset (Yang and Newsam, 2010), WHU-RS19 (Dai and Yang, 2011), and Aerial Image Dataset (AID) (Xia et al., 2017), are publicly available and have been widely used to evaluate these models. However, scene-based mapping has not been widely adopted in existing land use maps due to its tendency to cause serious mosaic phenomenon, characterized by zigzag boundaries and ambiguous geographical meanings (Zhang et al., 2022). Recently, street blocks, defined by roads and rivers, have become the preferred scene unit because they typically represent a relatively homogeneous urban function and align better with the basic unit of urban planning and land management. However, generating such units can be challenging in areas lacking ancillary geographic data like road networks, and spatial contexts might be difficult to delineate and characterize with specific rules (Oliva-Santos et al., 2014; Zhang et al., 2018a). Another concern about these scene-based classifications is the prevalence of mixed land use in urban areas, which can compromise mapping accuracy due to the ambiguity of highly mixed scenes.

2.3. Deep learning-based approaches

2.3.1. Base models

As deep learning techniques gain enormous popularity in computer vision and natural language processing applications, their advancements have significantly influenced the field of land use mapping. The emergence of monumental model architectures, exemplified in Fig. 3, has provided an important reference of base models for researchers to utilize and develop to facilitate the study of the urban environment.

2.3.1.1. Multi-layer perceptron. Multi-Layer Perceptron (MLP), also called artificial neural network, is the prototype of most advanced deep learning models. The simplest architecture of an MLP consists of one input layer, one hidden layer, and one output layer. The neurons in one layer are fully connected with the neurons in the following layer. Activation functions, appended after the fully connected operation, introduce non-linearity to enable the model to learn complex representations. During model development, two main computation processes, forward computation, and backward propagation, are executed alternatively. In forward computation, data are passed through the network from the input layer to the output layer. A loss function is used to measure the discrepancy between the model output and the true target. To minimize the loss, a gradient descent algorithm is then utilized to optimize the parameters in each layer during backward propagation. This three-layer structure can be further deepened by increasing the number of hidden layers to help the model fit more complex functions. MLPs have been applied to LULC classification since the late 1980s (Atkinson and Tatnall, 1997; Bischof et al., 1992; McClellan et al., 1989). The fully connected layers in MLP enable the model to thoroughly explore the relationships between every feature, thereby embedding them into higher-level features (Zhang et al., 2019). This type of dense connectivity structure is also frequently used to integrate heterogeneous features from multi-source big data, thereby better approximating the relationships between multiple features and complex land use categories. However, MLPs were less popular than other machine learning algorithms such as RF and SVM in most research fields for the following reasons: firstly, the "black-box" characteristic of neural networks leads to low interpretability; secondly, although theoretically deep neural networks are able to approximate any complex function (Hornik et al., 1989), increasing depth not only strains computational costs but also results in issues such as vanishing/exploding gradients, overfitting, and network degradation, all of which can seriously undermine model performance. These challenges have received wide attention with the surge in deep learning interest, particularly with the rapid development of convolutional neural networks over the past decade.

2.3.1.2. Convolutional neural network. Developed from the artificial neural network, a novel network architecture characterized by the convolution operation gradually occupies the dominant position of the computer vision field. Regarding the first CNN—LeNet proposed as the classic template of network architecture, a series of CNNs, such as AlexNet, InceptionNet, ResNet, SENet, and HRNet, appeared subsequently and acquired improvement in both accuracy and efficiency. These CNNs are constructed of two sub-networks: (i) the feature extractor network, comprising a hierarchical structure of convolutional layers (which apply a series of learnable filters to the input data to extract features by sliding these filters across the input spatially, performing element-wise multiplications with the part of the input they are currently on, and summing up the results into an output feature map) and subsampling layers for learning high-level feature representations

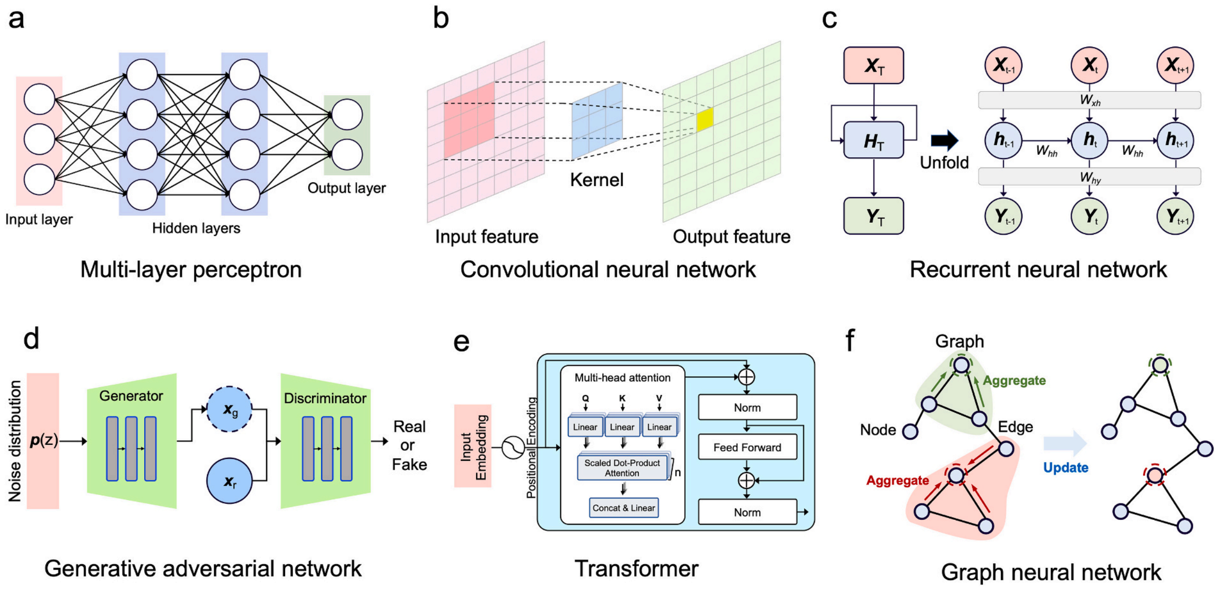

Fig. 3. Illustrative diagrams of base model architectures for (a) multi-layer perceptron, (b) convolutional neural network, (c) recurrent neural network, (d) generative adversarial network, (e) transformer, and (f) graph neural network.

and (ii) the classifier network, constructed of successive fully-connected layers that make the final classification or estimation (Chan et al., 2017; Paoletti et al., 2019). These two sub-networks can be trained together in an end-to-end manner following the same procedure as MLP (i.e., forward computation and backward propagation). Utilizing their exceptional ability to capture local spatial relationships, CNNs are often employed in conjunction with remote sensing imagery (e.g., multispectral, hyperspectral, microwave, and nighttime light data) and ground-level images. They can extract high-level semantic features that represent physical environments and socioeconomic attributes from the imagery, thereby enhancing the accuracy of complex urban land use classification. These CNNs have been used in scene-based or object-based land use information extraction from remote sensing data. Penatti et al. (2015) investigated the generalization ability of CNNs trained on natural images in the remote sensing land use classification tasks with 21 land-use classes. The results demonstrated the superiority of CNNs in accuracy compared to other low-level and mid-level descriptors. By introducing a multi-scale input strategy, Luus et al. (2015) proposed a multi-view CNN to improve the performance in land use classification, which achieved competitive accuracy compared to other models. However, the architecture decides that the model can only assign categories to specific regions and produce coarse-resolution mapping results. Although some researchers have attempted pixel-by-pixel classification by combining the CNN model classification for centered pixels of a local region with sliding window calculation, these methods often suffer from high computational costs and show unsatisfactory performance especially on the border between ground surface objects. As an alternative, a fully convolutional network (FCN) designed by Long et al. (2015), employs a stack of convolutional layers to replace the fully connected layers, turning CNN from an image classification model to a semantic segmentation model. FCN and its extensions were introduced to perform pixel-wise classification of land cover and land use categories on remote sensing images (Paisitkriangkrai et al., 2016; Volpi and Tuia, 2017; Wang et al., 2017). Volpi and Tuia (2017) employed the downsample-then-upsample architecture of FCN to perform full-resolution land cover mapping on sub-decimeter overhead images, obtaining state-of-the-art performance without using any post-processing and additional handcrafted features. Owing to the outstanding performance, CNNs are still the mainstream model being widely adopted and further advanced for land cover and land use classification tasks (Du et al., 2021; Zhang et al., 2020a; Zhang et al., 2022).

2.3.1.3. Recurrent neural network. As an important branch of deep learning algorithms, recurrent neural networks (RNN) have shown promising capability to handle sequential data, because “recurrent” in the context of neural network refers to connections that feed the output of a neuron back to its input, enabling the network to maintain a form of memory by considering its previous state in its current processing (Zhu et al., 2017). Derived from feedforward neural networks, RNNs are augmented models with the inclusion of connections that span adjacent phases, introducing the notion of time or order to the model (Lipton et al., 2015). These cyclic connections feed the network activations from a previous time phase as inputs to influence the network of current states, enabling RNNs to capture the dependency over the elements of different time phases (Sak et al., 2014; Sharma et al., 2018). There are three main categories of RNN architectures currently: vanilla RNN (Williams and Zipser, 1989), Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) network (Cho et al., 2014), and these models have been extensively used in earth science studies, especially for the applications involving temporal modeling and hyperspectral data processing. Satellites can observe a certain area dynamically and the time interval is decided by the revisit period. The temporal information of the land surface is beneficial for differentiating some land cover and land use categories (e.g., deciduous forests and evergreen forests). In light of the benefit, many research works used RNNs to extract the temporal features of the input time series of remotely sensed data and improve the classification performance (Campos-Taberner et al., 2020; Lyu et al., 2016; Rußwurm and Körner, 2017). Given that hundreds of spectral bands are provided in hyperspectral remote sensing data, it is also a popular method for hyperspectral land use classification to view these spectral bands as a sequence and then use RNNs to characterize spectral correlation and band-to-band variability (Hang et al., 2019; Mou et al., 2017; Pan et al., 2020a). Moreover, social sensing data that reflects human activities, such as user check-in and trajectory data, also exhibit strong temporal characteristics. These sequential models can effectively capture periodicity, regularity, and fluctuations of these data, introducing discriminative features that help distinguish between different land use categories.

2.3.1.4. Generative adversarial network. Generative adversarial networks (Goodfellow et al., 2014), featured by the two sub-networks (i.e., generator and discriminator) competing against each other, are one of the most innovative ideas in deep learning and have shown very impressive performance on generative modeling. In a GAN, a model acting as a generator performs the mapping from a given noise input to a particular data distribution of interest, while another model viewed as a discriminator is built to differentiate the fake/generated data produced by the generator from the real data (Creswell et al., 2018). The goal of this adversarial process is that the generator can synthesize plausible data that cannot be easily discriminated by the discriminator. Both the generator and the discriminator used in vanilla GANs are MLP, but it is common to extend it using CNN or RNN in order to process image or sequential data. Plenty of research works have already directly employed GANs to classify or segment remote sensing images (He et al., 2019; Jozdani et al., 2022; Xu et al., 2018). Besides, the studies that make use of the generative capability of GANs have also increasingly emerged for advancing land cover and land use mapping from different aspects. Shang et al. (2022) proposed a GAN-SRM model that used low-resolution pixels to predict high-resolution land cover maps, and the mapping results demonstrated that the introduction of GAN helped restore high-frequency details. Han et al. (2020) exploit GANs to generate samples including high-resolution remote sensing images and corresponding labels to support the land use classification tasks with insufficient annotated samples. To cope with the domain shift problem, Ji et al. (2021) developed a GAN-based domain adaptation method that fully aligned the source and target images from image space and feature space for land use classification using multiple-source remote sensing images.

2.3.1.5. Transformer. Transformer (Vaswani et al., 2017) is the state-of-the-art deep learning architecture, which was first developed for solving machine translation tasks and then became the leading model in NLP fields due to its capacity to capture long-range dependencies. The Transformer has an encoder-decoder architecture which only uses stacked self-attention and pointwise fully connected layers. The success of Transformer comes from the multi-head attention mechanism that enhances the representation capacity by jointly learning the dependencies of tokens at different positions from different representation subspaces. Beginning with Vision Transformer (ViT) (Dosovitskiy et al., 2020), the great success of Transformer has led to the extensive investigation of Transformer-based architectures for CV tasks and also intrigued researchers to explore the potential and feasibility of Transformers in classifying land cover and land use (Aleissaee et al., 2023; Liu et al., 2021b; Xie et al., 2021). For example, as the first work that purely applied Transformer to hyperspectral image classification, Hong et al. (2022) developed a transformer-based network – SpectralFormer and devised groupwise spectral embedding and cross-layer adaptive fusion modules to overcome the intrinsic weakness of Transformer. Chen et al. (2022b) employed a bitemporal image transformer to efficiently and effectively capture contextual information over time and space for change detection on remote sensing data. Transformer has shown great tremendous potential to replace the leading structures such as CNN and RNN in processing and fusing diverse modalities of data including remote sensing and social sensing for urban land use mapping (Liu et al., 2022; Zhou et al., 2023). Moreover, inspired by the recent technical breakthrough of ChatGPT (https://openai.com/blog/chatgpt) powered by transformer-based large language models, more advanced works are expected to exploit its potential in developing large pre-trained models for earth science (Cong et al., 2022; Sun et al., 2023; Wang et al., 2024).

2.3.1.6. Graph neural network. Different from images, which are composed of regular pixels, real-world entities are more complicated and irregular. Graphs consisting of nodes representing entities and edges representing the relationship between nodes, offer a more adaptable structure for describing a border array of real-world systems, such as social networks and transportation networks. As shown in Fig. 3f, the graph neural network aims at exploiting the information from adjacent nodes and itself to update the target node so that it can realize graph-based feature embedding. The embedded features are further used to carry out node-level, edge-level, and graph-level tasks. GNNs are able to effectively model spatial topological relationships and exploit contextual information from geospatial data, which have been demonstrated to be effective in understanding place characteristics (Kong et al., 2024). Furthermore, GNNs possess the unique ability to derive features from the connected local neighborhood even in semi-supervised scenarios with limited labels, making them particularly suitable for urban land use classification, a task that often suffers from label scarcity (Zhang et al., 2023). In terms of disclosing urban functions using geospatial big data, GNNs are also applied to model socio-economic data with irregular structures such as POI and road networks. Xu et al. (2022b) structured all the POI data in each land parcel as a graph and employed graph convolutional networks to extract the spatial context of POI data for the graph-level urban land use classification. By Extracting semantic information from street view images as node features, Zhang et al. (2023) constructed a graph network with streets as nodes to identify urban functions at the street scale.

2.3.2. *Multi-model ensemble*

The abovementioned deep learning models have been exclusively used in land use classification tasks by employing different kinds of data and overperformed traditional classifiers in terms of classification results. Nevertheless, each model possesses its own strengths and limitations in specific aspects, which have aroused researchers' interest in seeking more effective model architectures that can further refine mapping accuracy. A feasible and prevailing approach is to combine these deep learning models or even incorporate other reliable non-deep learning algorithms, within a single mapping framework. This integration leverages the intrinsic strengths of each internal element to offset existing weaknesses, thereby enhancing the efficacy and efficiency in tackling complex classification tasks (Penatti et al., 2015).

2.3.2.1. DL with Non-DL. The first type of integration is to combine a deep learning model with a non-deep learning model. Before deep learning models gained considerable popularity in LULC studies, traditional machine learning classifiers almost dominated relevant mapping studies. Even though deep learning architecture has shown promising results in effective land use classification, researchers tended to utilize deep networks as feature extractors, producing representative features for shallow machine learning classifiers, such as decision tree, random forest, and SVM (Leng et al., 2016; Li et al., 2019b; Nijhawan et al., 2019; Xia et al., 2022). For example, Dong et al. (2020) investigate the performance of a method based on the fusion of a random forest classifier and CNN on land use mapping of forest areas using VHR images. Basically, there are two reasons for taking such measures: firstly, these shallow classifiers are more interpretable than advanced deep learning models which were viewed as “black box” due to their weak interpretability; secondly, the combination of shallow learning models have stronger robustness and strength in small datasets. Apart from machine learning classifiers, traditional feature extractors (e.g., bag-of-visual-words model) and post-processing algorithms (e.g., conditional random fields (CRF) and Markov random fields (MRF) model) are also common measures to be integrated for improving classification performance.

2.3.2.2. DL with DL. With the advent of increasingly flexible and ingenious network architectures, integrating different deep learning models has emerged as a key approach for enhancement. According to the data modality of classification, these integrations of DL with DL models can be categorized into two groups: mono-modal and multi-modal methods. As for mono-modal methods, the entire LULC mapping framework, which integrates two or more networks, tackles only one type of input data, mostly images or other raster data, aiming to excavate the maximum potential of mono-modal data for mapping land use attributes. For example, Xiao et al. (2022) employed a CNN-GRU model to extract spatiotemporal neighborhood features and capture long-term dependency from the input time series data and verified its effectiveness in modeling land use dynamics of the Hexi Corridor, China. Song et al. (2023) explored the potential of the joint use of CNN and Transformer for urban scene understanding from remote sensing images. The local detail features and long-range relationships can be simultaneously exploited in the proposed model, thus improving the accuracy of semantic segmentation.

Given that different deep learning models have their own strengths in processing specific types of data (e.g., CNN for images/raster data, RNN for sequential/time series data, and GNN for graph data), the multi-modal models have been proposed in recent years to handle each kind of data effectively, thereby providing multifaced information to recognize land use types accurately. Yao et al. (2022) proposed an end-to-end land use identification model comprised of CNN, FCN, and LSTM models to mine physical attributes from remote sensing images and socioeconomic attributes from time series of electricity. In Fang et al. (2022)'s study, two modalities of data are included: the street-view images and the graph constructed by their spatial location. To cope with the multi-modal input, they developed an integrated model for urban land use classification where the CNN extracts high-level semantic image features, and the graph convolution network model simulates the spatial context and interaction.

3. Experimental assessments

3.1. *Materials and methods*

3.1.1. *Study area*

Shenzhen, located in southern China (22°27′ N- 22°52′ N, 113°46′ *E*-114°37′ E), is selected as the primary study area for the experimental assessment in this study. After becoming China's first special economic zone in 1979, Shenzhen has experienced unprecedentedly rapid urbanization, transforming the city from a cluster of fishing villages to one of the world's largest metropolitans. It is now home to over 13 million residents and spreads across a total area of 1996 km^2. As shown in Fig. 4, Shenzhen consists of 10 administrative districts: Baoan, Guangming, Longhua, Nanshan, Futian, Luohu, Yantian, Longgang, Pingshan, and Dapeng. The city's complex economic structure and diverse population distribution have resulted in intricate and varied urban land use categories across the city. This complexity and diversity in land use patterns make Shenzhen an excellent testbed for evaluating the performance of different deep learning models in generating urban land use category classification.

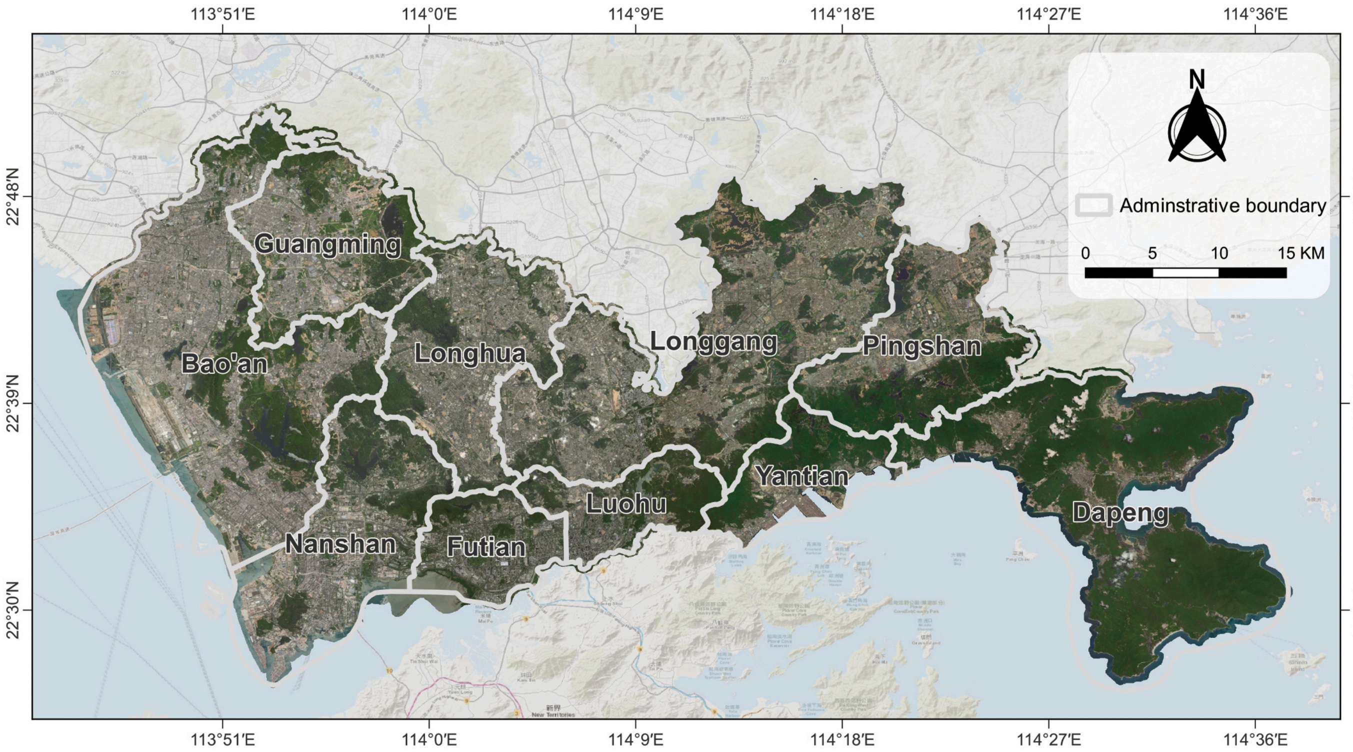

Fig. 4. Geographical location of Shenzhen, China and its administrative boundaries of 10 districts.

3.1.2. Urban land use dataset

Built upon the Shenzhen Urban Land Use Classification system (SULUC), the urban land use classification categories in this research are structured as a two-level classification scheme comprised of 5 Level-I categories and 13 Level-II categories, as shown in Table 1. Compared with SULUC, the scheme included fewer categories because we aggregated several classes with limited samples. As the basic analytical unit of this study, three-level land parcels were obtained by parcel segmentation with a road network. Only the parcels within the built area were retained while the other areas such as water surface, farmland, forestland, and bareland were excluded.

For each parcel, we assign urban land use categories by referring to the land survey data that covers the entire city of Shenzhen. A series of sample verification and quality checks including indoor processes and in situ investigation have been conducted to ensure the accuracy and reliability of urban land use category labels (Su et al., 2020), which are some of the most important factors for building optimal machine-learning models. By doing these, we obtained a multi-level parcel-based urban land use dataset in Shenzhen with complete coverage and accurate categories, facilitating our following experiments.

3.1.3. Methodology

Fig. 5 presents the workflow for the deep-learning-based urban land use category classification framework employed in this study. The data included satellite images, auxiliary data, land-use survey data and the segmented street blocks of Shenzhen. The first step involves data alignment to preprocess satellite images and auxiliary data. This process includes converting data to raster, standardizing the projection, resampling images, and aligning images. The aligned data are then fused by stacking and clipped into block-level patches according to the street block extents. In the model development stage, these data patches are matched with the land-use survey data to construct a sample set. Each sample in this set contains corresponding input data and a specific land use class label. This sample set is further proportionally split into a training set and a test set. The training set is used to train the deep learning model. During the training process, the data are input into the model, and forward computation results in prediction outputs. The loss function calculates the discrepancy between these predictions and the reference. The backward propagation of the loss gradient helps the model obtain the gradients of trainable layers, which are used to update and optimize the model weights. This forward computation and

Table 1
Two-level classification scheme of urban land use categories in Shenzhen.

Level-I	Level-II	Description
01 Residential	101 Urban village	Original rural resident housing currently mostly surrounded by city blocks
	102 Urban residential	Land used for residential housing and related facilities
02 Commercial	201 Business and finance	Commercial and service land used for business operations
	202 Golf course	Golf course and service housing and facilities
	203 Storage	Land used for stockpiling and temporary storage for distribution
	204 Other commercial	Retail, wholesale, production and sales, services, and entertainment land
03 Industrial	301 Industrial	Land used for production, product processing, manufacturing, machine repair, and other related facilities
04 Transportation	401 Road	Transportation land
	402 Transportation Station	Land used for service facilities, such as stations, transfer stations, parking facilities
	403 Harbor	Land used for harbors or related facilities
05 Public Service	501 Instructional and research	Land used for instruction, research, design, surveying, testing, environmental assessment, extension, etc.
	502 Parks and green space	Parks, zoos, gardens, squares, and other green space for recreation
	503 Public infrastructure	Land used for public infrastructure

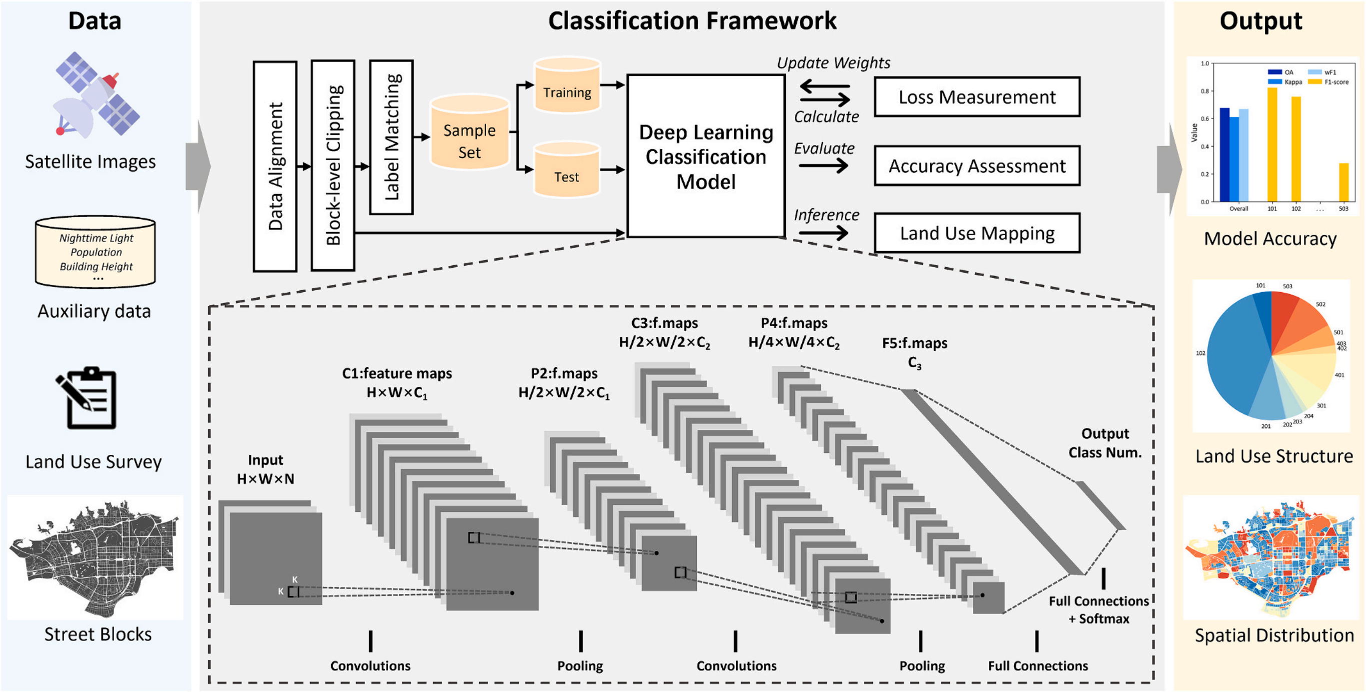

Fig. 5. Workflow of the entire urban land use category classification framework.

backward propagation process is iterated until a set number of iterations is reached. The best model saved during iterations is evaluated by conducting an accuracy assessment comparing the prediction with the reference values of the test set. In the inference stage, block-level patches, without any prior labels, are directly input into the trained model, resulting in an urban land use map. This map can be used to analyze urban land use structure and spatial pattern. The majority of deep learning models used in this study are CNN models, so we present a simplified architecture of a CNN, containing only key elements in this workflow. Specifically, this model takes an input image patch with the dimensions of height × width × N, where N denotes the number of input channels corresponding to the number of spectral bands and auxiliary data. A convolutional layer with a kernel size of K processes the input data and produces the feature maps C1. These feature maps are then spatially aggregated by a pooling layer to reduce computations and enhance translation invariance. After passing the feature maps into another group of convolution and pooling layers, the generated P4 feature maps are transformed into a one-dimensional vector through flattening or global pooling. This vector passes through several fully connected layers and finally, a softmax function processes the vector to produce the prediction of class probabilities, which can be assigned to specific category. Compared with the basic CNN, deeper models append more convolutional, pooling and fully connected layers. Other state-of-the-art models enhance this basic architecture by adding novel and effective modules to improve performance in certain aspects (e.g., dropout, batch normalization, residual connection, etc.).

3.2. Experimental design

3.2.1. Basic setting and implementation

To conduct extensive experiments on land use classification, the basic configuration of the baseline method is detailed as follows. MobileNet is selected as the main classification network because of its lightweight architecture and computational efficiency. Worldview images with red, green, and blue bands are the main data source. The images are cropped according to the boundaries of level-3 land use parcels, and then resampled to image patches with a uniform size of 224 × 224 pixels. The complete dataset is split into a 70% portion for network training and the remaining 30% for evaluation. The spatial distribution of two sets is shown in Fig. 6. Due to the constraints in computing capacity, the size of each mini-batch is set to 16, and the deep network is trained for 100 epochs. The optimal model weights saved during the training course are further assessed by the test dataset. To balance the sample sizes across different land use types, a weighted random sampler is used to select samples from training set to form batches for training. The loss function employed is the cross-entropy loss. Network parameters are optimized by the Adam algorithm with an initial learning rate of 0.001. Data augmentation methods such as horizontal flip, vertical flip, and clockwise rotation by 0, 90, 180, and 270° are applied to the training samples to alleviate overfitting problems. Each of the experiments mentioned below adjusted only one specific component, while the rest of the model remains consistent with the baseline method unless explicitly stated otherwise, ensuring a fair comparison.

To evaluate the performance of models generated in the following experiments, we employed four widely used evaluation metrics: (1) overall accuracy (OA), which represents the proportion of correctly classified samples out of all test samples; (2) F1-score, which incorporates the precision and recall for one class by computing the harmonic mean; (3) weighted F1-score (wF1), which synthesizes all of the class-wise F1-score by a weighted average whose weights are determined by the proportion of each class, and (4) Kappa Coefficient, which measures the agreement between classification results and the ground truth reference.

3.2.2. Experimental details

We designed eight sets of comparative experiments (Fig. 7) with the aim to investigate the impact of different configurations on classification performance in model development for urban land use mapping. The intention is to provide comprehensive references for researchers employing deep learning models in urban land use classification studies.

The first experiment investigates how different types of input data affect the performance of the models. The input data are divided into two categories: remote sensing data and auxiliary data. The remote sensing data used in this experiment include 0.5-m WorldView imagery, 3-m PlanetScope imagery, 10-m Sentinel-2 imagery, 30-m Landsat-8

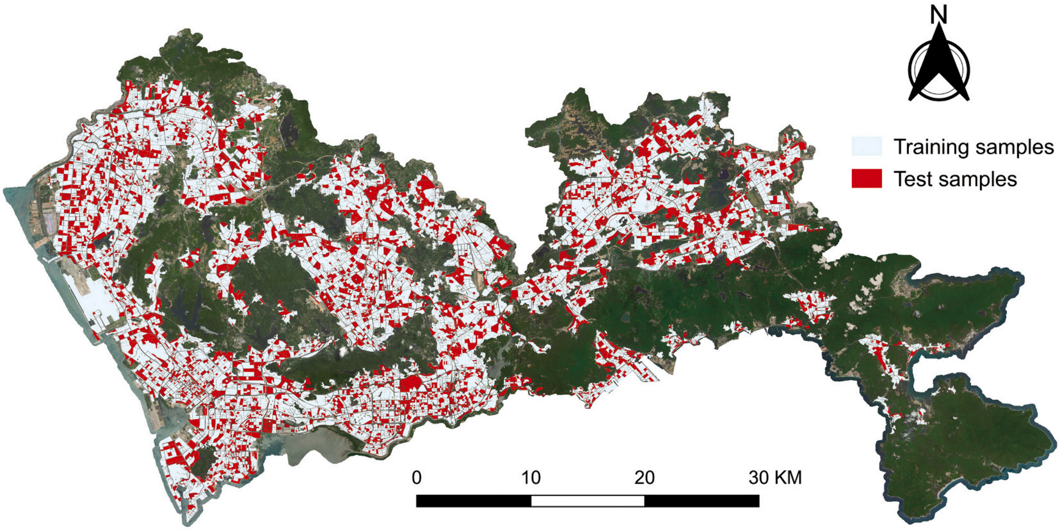

Fig. 6. Spatial distribution of the training and test samples for urban land use classification.

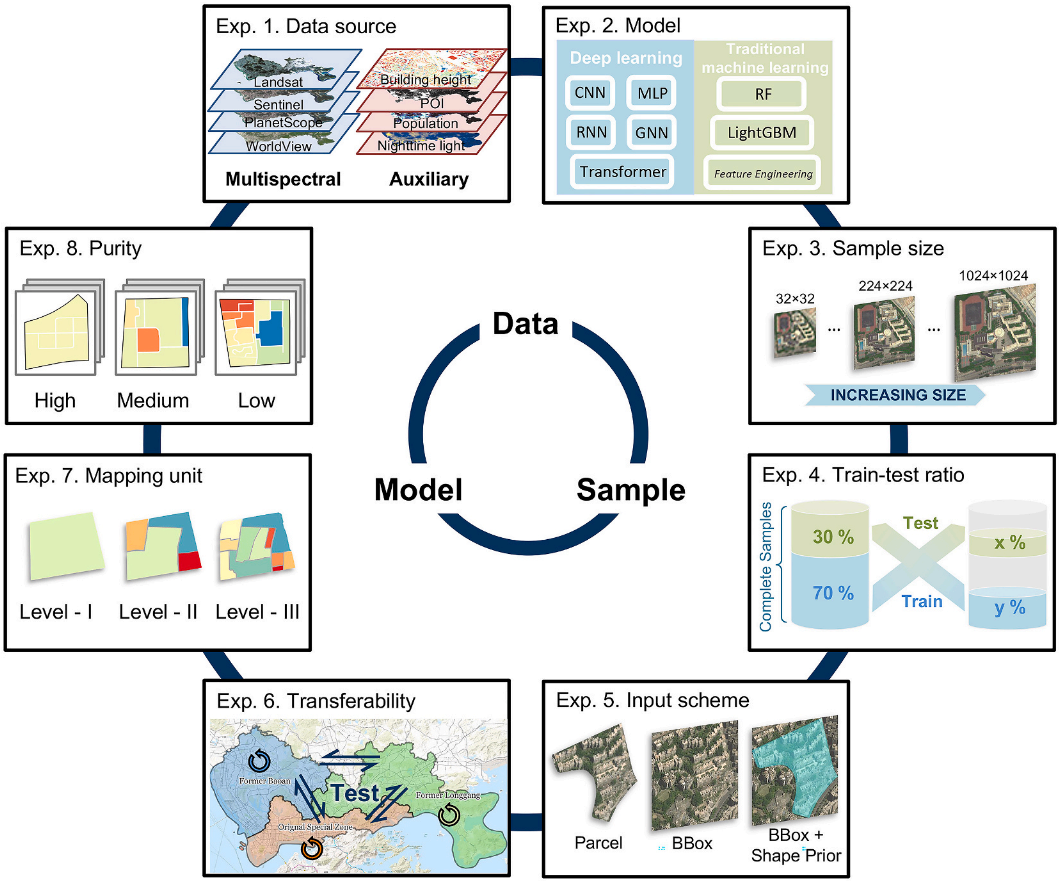

Fig. 7. Illustrative diagram of eight comparative experiments conducted in this study.

data, 30-m Sentinel-1 SAR imagery, Suomi NPP VIIRS Day-Night Band imagery. All these remote sensing data were collected in 2020. The auxiliary data includes Gaode POI data, WorldPop population, and building height that are assumed to supplement socio-economic attributes for urban land use classification. The detailed information on these datasets is listed in Table 2. Before model training, these collected multi-source data were first preprocessed to ensure compatibility and facilitate subsequent integration. Given that raster is the primary modality in the planned experiments, data from other modalities (i.e., POI and building footprint) were transformed into raster format. Specifically, Gaode POI data, originally in point vector format, were converted to raster, with the count of each POI category within each raster cell assigned as raster values to represent the spatial density of urban functions. Building footprint data was similarly converted from the polygon vector to the raster with building heights encoded as the raster values to provide three-dimensional information on urban structures. After unifying the

Table 2
An overview of the datasets used in the experiments.

Category	Data	Resolution	Year	Description
Multi-spectral	Worldview-2	0.5 m	2020	Includes three primary bands (blue, red and green) in visible spectrum
	Planetscope	3 m	2020	Includes four spectral bands of red, green, blue and near-infrared
	Sentinel-2	10 and 20 m	2020	Offers a broad range of spectral bands including blue, green, red, four red edge bands at 705, 740, 783, and 865 nm, near-infrared, and two short-wave infrared bands at 1610 and 2190 nm.
	Landsat-8	30 m	2020	Provides comprehensive spectral coverage with bands of coastal/aerosol, blue, green, red, near-infrared, two shortwave infrared, and cirrus bands.
SAR	Sentinel-1	10 m	2020	Contains backscatter coefficient of VV and VH polarization from C-Band Ground Range Detected (GRD) scenes
Nighttime light	VIIRS nighttime light	500 m	2020	Provides stray light corrected VIIRS Day/ Night bands radiance
POI	Gaode POI	–	2018	Consists of names, location coordinates, and specific urban functional types for each point-of-interest data.
Building	Building footprint and height data	–	2020	Provides building outline and height in vector polygon format
Population	WorldPop population	100 m	2020	Offers annually estimated number of people residing in each grid cell

data format, we use the Align Rasters tool in QGIS to align the extents, coordinate reference systems, and spatial resolution of different datasets. As parts of our experiment aimed to compare the influence of auxiliary data on Sentinel-2 and Worldview-1 data, these two data served as the reference layer for alignment. Bilinear interpolation was chosen as the primary resampling method to adjust low-resolution data to match the resolution of reference data. Upon completing these steps, we segmented the aligned data into patches confined to the boundaries of the land use parcels. Depending on the specific design of each case, we stack these aligned multisource data along the channel dimension to form a three-dimensional patch. Paired with the corresponding land use type labels, these patches were ready for the subsequent model training and evaluation. Based on this dataset, we conducted a range of experiments to examine the effects of spatial resolution, the number of spectral bands, and the incorporation of auxiliary data on the classification performance.

The second experiment examines the performance of various artificial intelligence models, spanning from shallow learning to deep learning to gauge their strengths and weaknesses when applied to urban land use category classification. In this experiment, we have considered a number of popular classification algorithms: deep learning networks such as Visual Geometry Group Network (VGGNet) (Simonyan and Zisserman, 2015), Residual Neural Network (ResNet) (He et al., 2016), Wide Residual Network (WideResNet) (Zagoruyko and Komodakis, 2016), Densely Connected Convolutional Network (DenseNet) (Huang et al., 2017), MobileNet (Howard et al., 2017), MLP, LSTM, Vision Graph Neural Network (ViG) (Han et al., 2024) and Visual Transformer (ViT) (Dosovitskiy et al., 2020), along with shallow machine learning algorithms including Random Forest (RF) (Breiman, 2001) and Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017). For deep learning models, we also compared the performance of models with and without pretrained weights. As for traditional machine learning algorithms, we studied the effect of sample balancing and feature engineering. To balance the sample, we increased the sample size of each land use category to match the size of the category with the most samples, which is "Industrial land use" in our study. Feature engineering for the traditional machine learning models involved 25 commonly used features from previous research, calculating the mean, standard deviation, maximum and minimum values of three spectral bands (i.e., red, green, and blue) and eighteen texture features (i.e., contrast, dissimilarity, homogeneity, energy, correlation, angular second moment for each spectral band) generated by gray-level co-occurrence matrix, and also the parcel area.

The third experiment compares deep learning models that are trained with different patch sizes. For efficient mini-batch training, input data of varying sizes need to be rescaled to the same size to train the deep neural networks. Because the parcels in this study are divided based on road networks, each parcel is of a different size. Downsampling an image from a large patch to a smaller one leads to the loss of detail, whereas upsampling a smaller parcel increases computational burden. As such, an optimal scale is required for resampling the image to achieve accurate and efficient classification. In this experiment, the clipped images were resampled into square patches with uniform lengths of 32, 64, 96, 128, 160, 192, 224, 256, 384, 512, 768, and 1024 pixels. These rescaled patches were then used to train and evaluate deep learning models.

The fourth experiment contrasts the performance of models trained and evaluated with different amounts of training and test samples. The quantity and diversity of training samples affect the classification accuracy and generalizability of the deep learning model, while the volume and representativeness of test samples reflect the confidence in evaluation outcomes. In this experiment, the entire dataset was first randomly divided into the training set and test set with a ratio of 7:3 for each class. We reduced the amount of training sample from 100% to 1%, which was then used for model training. The goal here was to explore the influence of training sample size on classification performance. Another set of experiments where the amount of test samples was decreased from 100% to 1% was implemented to investigate its impact on evaluation results.

The fifth experiment examines the influence of different input data strategies. Achieving accurate remote sensing image classification largely depends on providing rich, representative, and task-related data. For parcel-level land use classification, the internal information within the parcel is crucial as it directly signifies the parcel's attributes. Nevertheless, contextual information from outside the parcel could also provide supplementary information to enhance the precision of the deep learning model's classification. To this end, we have considered three data input strategies based on this assumption: (1) the sample only contains pixels within the parcel, (2) the images that preserve all pixels within the parcel' envelop, and (3) the envelop image combined with the parcel's binary mask.

The sixth experiment explores the spatial transferability of deep learning models across different areas of the same city. As a result of the Special Economic Zone policy in Shenzhen, the city can be divided into three spatially continuous regions (Fig. 8a): the original special zone (OSZ), former Bao'an and former Longgang districts until 2010 when the extent of special economic zone was expanded to the entire Shenzhen. The 30-year policy and legal difference resulted in imbalanced development of infrastructure and economy in these three regions. The spatial strategy rezoning in 2010 expected to balance development accelerate the reformation of land use and the upgrade of infrastructure construction in underdeveloped area, but as shown in the Fig. 8b, the district-level guidelines lead to the difference in the urban function structures(Xiao et al., 2023). The varying development and land use

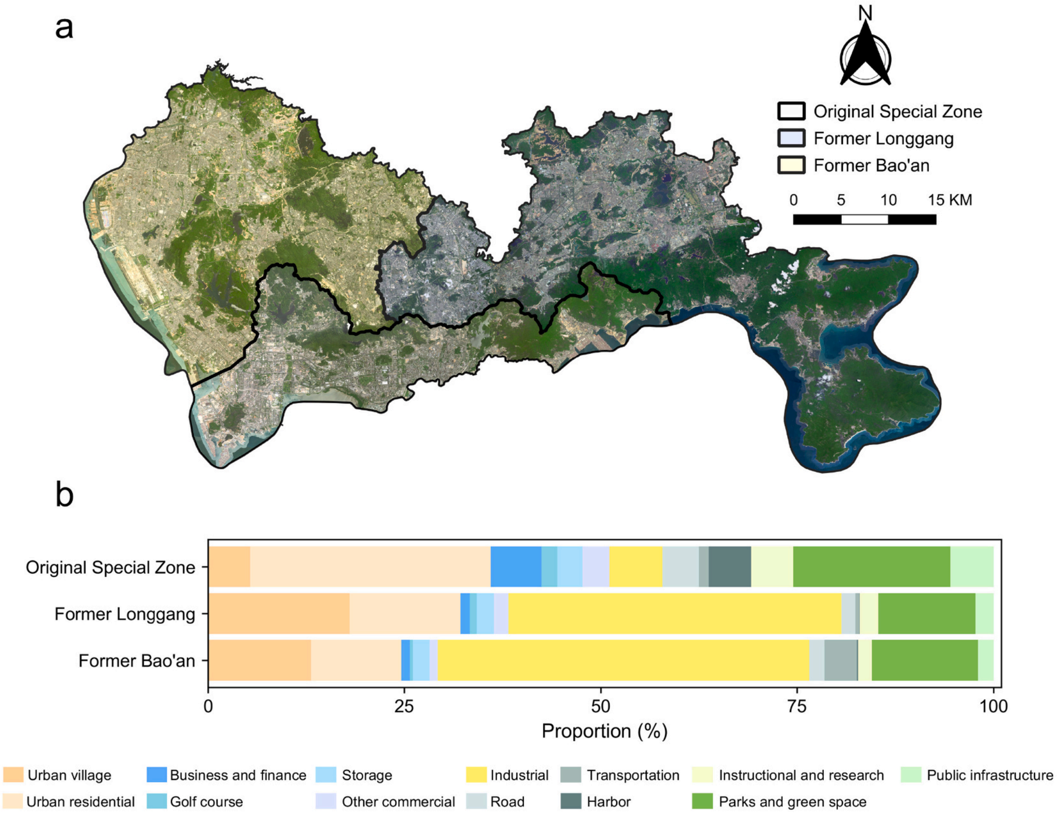

Fig. 8. Geographical distribution (a) and urban land use composition (b) of three study subregions in Shenzhen.

compositions make this three-zone partition an ideal testbed for the transferability of land use classification models. Besides, rather than the district-level partition, the three-zone partition scheme can ensure enough samples for each class in training and evaluation across different regions. For each of these regions, the sample set was split to a 70% training set and a 30% test set. This led to the creation of three groups of experiments. In each group, one of the regions served as the source domain, from which a region-specific classification model was trained. This model was not only evaluated on the test set of the source region but also on the other two regions, acting as target domains, to assess the spatial transferability of the deep learning methods. Moreover, we investigated the efficacy of the fine-tuning technique by evaluating the performance gains through the incorporation of additional samples from the target domain. In these fine-tuning experiments, the models were initially trained on the source region data, and subsequently fine-tuned using the training samples from the target region.

The seventh experiment compares the performance of urban land use classification at different parcel levels. Parcels were generated by using different levels of road networks to divide the urban area, leading to parcels of different levels. The lower-level parcels are subsets of the upper-level parcels. In this experiment, we cropped the data for three levels of parcel, creating three datasets to develop deep learning models. We then assessed the performance of these models to examine the impact of different segmentation scales on land use classification.

The eighth experiment examines the impact of sample purity on the performance of urban land use classification. Mixed land use is a common occurrence in urban areas and often complicates the assignment of a definitive label to a given parcel. Typically, the parcel category is determined by the predominant land use category that accounts for the largest portion of the parcel. This areal proportion of the major land use category is defined as the parcel's purity. In this experiment, the complete set of land use samples was divided into three subsets based on purity: the high-purity subset (where each parcel has a label purity of 0.9 or higher), the medium-purity subset (where each parcel has a label purity ranging from 0.6 to 0.9), and the low-purity subset (where each parcel has a label purity of <0.6). Each subset was further split into a 70% training set and a 30% test set. Each training set was used to train a model, which was subsequently validated using three different test sets, each with a different level of purity.

3.3. Results

3.3.1. Data input

The quantitative results of experiments using the same MobileNet model, but different inputs of data sources are presented in Table 3. Four multi-spectral data sources (Landsat-8, Sentinel-2, PlanetScope and WorldView) represent remote sensing images with different spatial resolutions and spectral bands. When the models only used red, green and blue bands as input, an upward trend can be observed in three overall evaluation metrics as the spatial resolution of the image increases, which indicates that the higher resolution data are able to provide more useful features spatially so that the model can improve the classification performance. Apart from three-primary colors, the electromagnetic signals of the other spectral bands are also recorded by the sensors of Landsat, Sentinel, PlanetScope, providing a more comprehensive description of terrestrial objects. As for Landsat-8 data, using all spectral bands as model input brings an increment of 5.39% in the overall accuracy compared to only using the bands of three-primary colors. Inputting all spectral bands of Sentinel-2 and PlanetScope images can only gain 2.32% and 1.88% increase in the overall accuracy. It demonstrates that low-resolution data, including more spectral bands, can gain larger benefits than higher-resolution data. We also conduct the experiments using only the Sentinel-1 SAR data with VV and VH intensity or POI data. The results indicate that either SAR data or POI data alone fails to effectively capture the urban functional attributes, with the overall accuracy not exceeding 50%. Besides, the effect of adding

Table 3
Classification performance of models trained with different input data sources and features. The column names encoded with numbers from 101 to 503 represent different Level-II urban land use categories defined in Table 1.

Data	OA	wF1	Kappa	Class-wise F1-score												
				101	102	201	202	203	204	301	401	402	403	501	502	503
L8(RGB)	47.34	45.50	36.00	52.41	57.31	20.18	28.57	9.76	9.38	54.87	47.48	2.94	52.63	14.43	41.51	10.05
L8(All)	52.73	50.90	42.60	60.34	63.38	26.96	40.00	8.23	20.9	60.18	50.19	6.45	70.37	24.12	43.35	14.34
S2(RGB)	60.66	58.84	52.36	71.98	68.32	33.33	57.14	17.97	27.50	66.33	61.54	12.90	74.58	44.34	54.23	14.95
S2(RGBN)	61.59	60.26	53.68	73.73	69.76	36.59	33.33	21.56	22.50	68.31	63.25	14.08	78.57	42.01	51.81	18.11
S2(All)	62.98	61.23	55.19	73.85	71.99	35.62	44.44	18.53	30.59	70.43	66.89	2.86	80.70	43.78	50.81	15.25
S2(All) + S1	63.55	62.10	55.94	77.05	72.38	34.31	75.00	22.68	26.67	70.96	65.00	5.8	73.33	39.8	50.51	22.67
S2(All) + POI	65.58	64.44	58.61	75.97	74.45	47.94	57.14	30.08	20.59	73.36	63.31	5.97	68.97	58.01	49.94	33.85
S2(All) + NTL + BH + POP	64.6	63.11	57.24	75.92	72.72	36.77	50.00	16.6	42.11	73.10	64.97	7.89	81.36	43.40	54.19	24.27
S2(All) + S1 + POI + NTL + BH + POP	66.25	65.36	59.33	73.78	73.49	44.44	57.14	34.93	29.73	72.92	67.34	28.57	79.25	58.72	55.33	37.50
PL(RGB)	61.61	59.79	53.50	75.29	69.70	34.75	50.00	18.45	31.58	68.46	64.10	6.45	58.82	41.28	48.01	15.93
PL(RGBN)	63.49	61.87	55.86	78.18	70.73	34.02	75.00	24.54	30.95	70.79	63.80	15.62	76.67	42.15	49.61	18.02
WV(RGB)	68.28	67.22	61.82	83.51	76.51	45.57	57.14	27.27	35.00	75.17	65.99	21.92	76.00	55.40	55.32	25.20
WV(RGB) + S1	66.09	64.50	58.94	81.67	72.08	39.27	28.57	30.00	38.46	73.17	66.44	13.33	78.43	49.54	52.16	18.58
WV(RGB) + POI	66.41	65.49	59.66	75.37	74.42	47.2	75.0	33.22	28.21	74.2	64.7	5.71	75.0	59.82	52.99	38.78
WV(RGB) + NTL + BH + POP	66.61	65.14	59.71	80.26	74.33	36.28	44.44	25.59	32.43	74.20	65.58	20.59	78.57	46.58	55.07	25.00
POI density	47.94	46.75	38.16	55.89	51.79	34.36	0	24.63	12.61	59.02	44.75	5.26	5.56	53.11	25.28	28.83
S1 (VV + VH)	44.18	42.67	32.31	48.38	45.91	30.96	57.14	9.56	2.99	53.86	49.65	3.39	62.96	10.65	43.5	9.61

auxiliary features from remote sensing and social sensing was also tested in Sentinel-2 and Worldview data. For sentinel-2 data, adding features from Sentinel-1 leads to 0.57% increase of OA, while integrating POI density map gains largest improvement of OA (2.60%). Combining all of the auxiliary features with Sentinel-2 multi-spectral features can help the model achieve the OA of 66.25%. It shows that such assistant information can reflect the socio-economic properties of land surface can supplement the shortage of spectral information and thus the model classification accuracy can be further improved. But the accuracy of the model using all spectral bands and auxiliary data input was still far from that of the model only using RGB bands of Worldview data. In contrast, for worldview data, introducing auxiliary data results in varying degrees of accuracy loss. This decrease may be due to the fact that the very-high-resolution images of the Worldview satellite contain fine and rich socio-economic information implicitly. However, adding these auxiliary data, most of which are coarse-resolution, introduces noisy information and ambiguity to the model, undermining the original classification performance of the deep network. As a whole, increasing the resolution of input spectral data can gain more benefits than adding extra features when developing a deep learning-based urban land use classification model. As shown in Fig. 9, deep learning classification methods using VHR images can accurately map the spatial composition and pattern of urban land use function (Fig. 9b), which is visually in high consistency with the reference land use map (Fig. 9a). As the spatial resolution increases, less classification error can be observed in the three example regions.

L8: Landsat-8; S2: Sentinel-2; S1: Sentinel-1; NTL: Nighttime light; BH: Building height; POP: Population.

3.3.2. Classification models

The performance of different machine learning models, including traditional machine learning algorithms and deep learning models, are reported in Table 4. The class-specific sample size and mean performance are also illustrated in Fig. 10. Among ten deep learning networks, WideResNet outperformed other models, achieving the best overall accuracy, wF1 and Kappa. Comparing to the vanilla resnet, the strength of WideResNet is that it enlarges its width instead of the depth. A deeper network with identity mapping in residual blocks suffers from diminishing feature reuse, which means only some layers can learn meaningful representation while the features learned by other layers are washed out after repeated operation. In WideResNet, larger width (channel size) and dropout are two key mechanisms that help it avoid this issue and obtain superior performance over other deep learning models. MobileNet, the network with lightweight architecture, achieved the second-best performance. With a basic motivation similar to WideResNet, MobileNet also tries to preserve a wide structure as much as possible by introducing linear bottlenecks and inverted residual modules to enlarge the channel size so that it can alleviate the serious representation loss issues. For this urban land use classification task, the complex model does not always perform better than the simple one. The two most complicated models with more parameters, DenseNet and ViT achieve worse performance than other deep learning models. We also compare the performance of MLP, LSTM, and ViG, achieving the OA of 35.34%, 58.22%, and 61.13%, respectively. Such poor performance is attributable to the weakness of these model structures in leveraging local spatial correlations inherent to image representations. The results of ResNet series models also show that model performance gets worse as the depth of a model increases. It is because, in challenging tasks like urban land use classification with insufficient samples, complex models are more prone to encountering the problem of overfitting and thus affect the model performance on the test set. Fig. 10 shows the distribution of classification performance of ten deep learning models for each land use category. It demonstrates that the categories with large sample sizes can achieve better accuracy and stability. Notably, the harbor class also shows high accuracy, which is mainly because unique characteristics such as sea surface and containers in the scene make this class more differentiable even though limited samples are used. Moreover, training the models without pretrained weights leads to a 5.33% drop for MobileNet and a 7.42% drop for WideResNet, emphasizing the importance of pretrained models or transfer learning when dealing with such difficult tasks without sufficient samples. Compared with deep learning networks, the traditional machine learning models achieve relatively low accuracy. The LightGBM model trained with data processed by feature engineering, the best model among them, obtains an overall accuracy of 48.42%, which is still over 10% lower than the overall accuracy of deep learning models. Moreover, feature engineering is crucial to traditional machine learning algorithms. Even though the applied feature engineering is simple, it helps random forest and LightGBM improve their overall accuracy by 8.68% and 5.95%. Nonetheless, due to the limitations of human-designed features, its representation ability is still not as effective as the features generated by deep learning networks leveraging data-driven representation learning. A class imbalance problem also seriously affects the performance of machine learning models. For the machine learning algorithms with

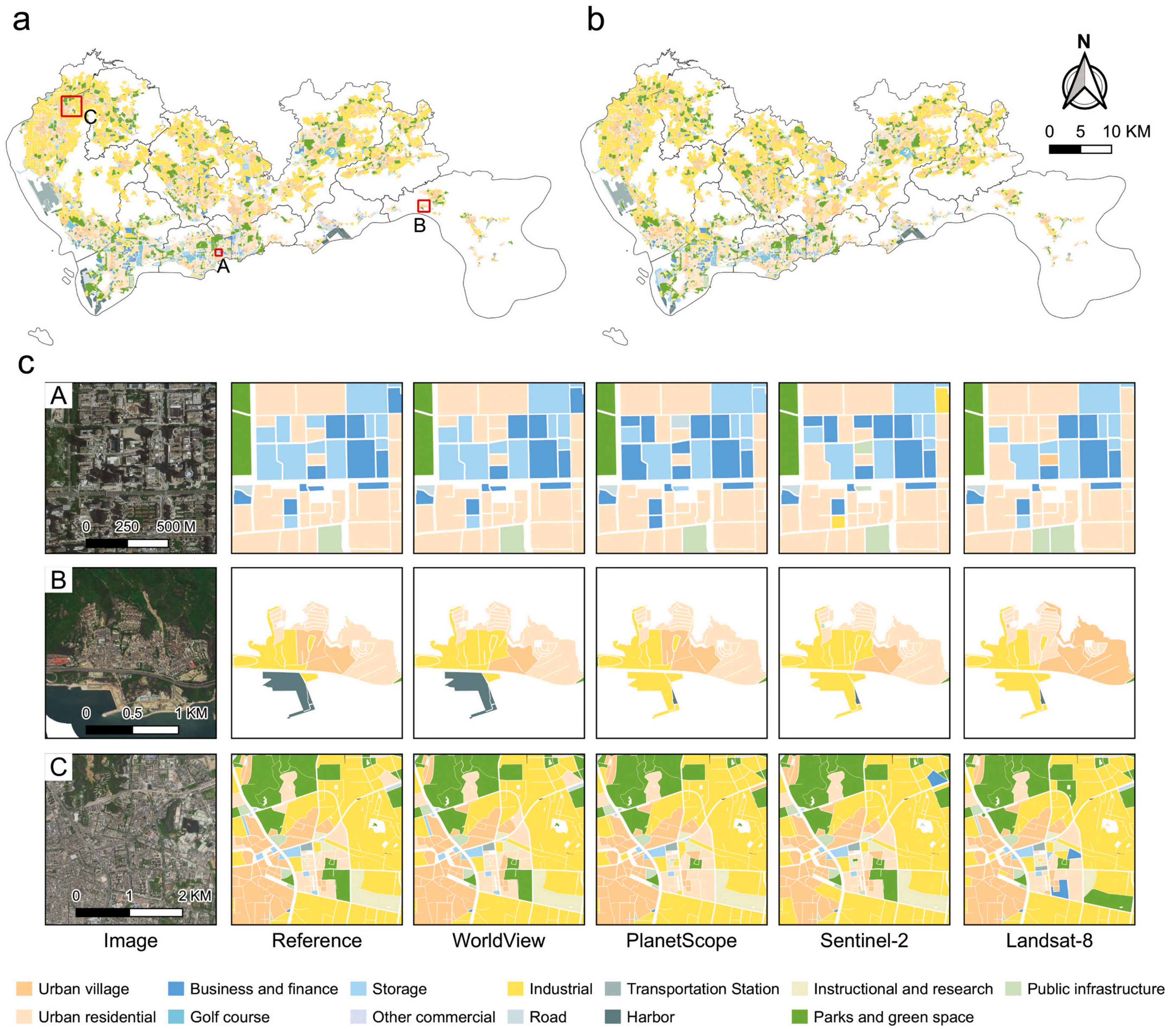

Fig. 9. Land use categories maps of Shenzhen: (a) reference map and (b) classification map by deep learning method, and (c) three zoom-in subsets of high-resolution images, reference map, and classification maps with different satellite images as input.

unbalanced samples, the performance of some minority classes is extremely low or even achieves zero F1-score. Despite using over-sampling methods to balance the data improved the performance of minority classes as the table shows, such balanced data led to a significant decrease in overall performance. In terms of the computational cost, deep learning models significantly outperform machine learning which can be attributed to the end-to-end architecture of deep learning models and its strength in tackling raw data input. Among the deep learning methods, VGGNet is the most efficient, but its accuracy falls behind the two lightweight models and WideResNet. WideResNet, achieving the highest accuracy, demands more time than lightweight models. Considering the trade-off between accuracy and efficiency, MobiletNet emerges as a better option, serving as the basic model architecture in the other experiments. Within the machine learning methods, the computational cost of feature engineering nearly doubles computational time, without gaining comparable accuracy improvements to deep learning approaches. It highlights both the efficiency and effectiveness of deep learning-based mapping. Fig. 11 presents the spatial distribution of land use parcels classified by Wide-ResNet (Fig. 11b) and the reference data (Fig. 11a). From the three subsets (Fig. 11c), it is evident that deep-learning-based methods (Wide-ResNet and MobileNet) achieve higher consistency with reference distribution than random forest. Thereinto, without manmade feature engineering, the model taking flattened pixel values as input failed to classify easy scenes such as the greenspace in example C.

3.3.3. *Image scale*

Table 5 presents the overall performance of MobileNet trained with different resampling scales and also the performance for each original scale group. The highest overall accuracy was obtained by the model trained with the resized 384*384 pixels patches. The overall performance shows an increasing trend from the size of 32 pixels to 384 pixels, and then the performance fluctuates when the resampled scale keeps rising. Additionally, the inference time increases with the size of the resampling scale, which highlights a trade-off between achieving higher accuracy and maintaining computational efficiency. By inspecting the

Table 4
Classification performance of deep learning models and traditional machine learning models.

Methods	OA	wF1	Kappa	Class-wise F1-score													Inference Time (s)
				101	102	201	202	203	204	301	401	402	403	501	502	503	
VGGNet	67.07	66.51	60.65	83.51	75.47	45.88	50.00	26.37	24.39	75.78	66.23	22.99	74.07	50.42	55.76	16.22	78.76
MobileNet	67.59	66.83	61.04	82.35	75.77	45.96	80.00	33.57	40.86	74.59	64.32	11.27	84.21	51.72	54.59	27.64	98.25
ShuffleNet	67.20	66.06	60.54	82.85	75.19	43.04	54.55	25.96	34.67	75.82	62.29	21.62	76.36	51.28	52.36	24.41	107.04
ResNet18	66.43	65.16	59.48	81.79	75.35	39.47	50.00	28.57	29.27	73.74	61.20	8.82	73.47	51.38	53.69	23.24	80.25
ResNet34	66.09	64.67	59.07	81.42	73.52	41.15	66.67	21.68	37.04	72.83	64.62	10.96	76.00	46.73	56.25	20.72	96.90
ResNet50	65.17	63.20	57.98	79.72	71.96	42.06	0	20.69	35.62	72.33	64.51	6.56	76.00	48.00	53.52	13.76	113.39
ResNet101	65.19	63.56	57.81	78.39	72.45	43.29	50.00	24.70	22.50	73.28	62.67	8.82	80.00	48.40	53.03	19.82	131.61
WideResNet	68.80	67.34	62.27	83.8	76.41	44.84	75.00	29.66	31.58	75.88	65.37	11.76	82.35	56.46	55.41	17.57	111.19
DenseNet	64.39	63.70	57.43	79.29	74.02	41.25	44.44	22.44	37.97	71.13	62.46	14.63	82.14	43.97	53.53	23.79	151.44
ViT	62.28	62.64	55.51	80.31	70.56	45.14	44.44	21.66	11.11	72.71	62.57	6.06	78.43	51.23	49.57	16.19	102.59
ViG	61.13	61.03	53.96	75.22	71.72	41.38	28.57	17.86	27.74	67.46	63.99	15.52	73.68	51.64	51.25	16.36	165.06
LSTM	58.22	57.48	50.33	72.54	67.78	40.12	25.0	14.23	19.82	64.41	60.37	11.11	66.67	39.26	51.8	4.85	68.44
MLP	35.34	33.62	22.31	41.11	43.40	21.43	0	5.53	7.34	35.47	44.64	6.9	21.95	0	34.93	0	64.23
Without pretraining																	
MobileNet	62.26	61.90	55.05	76.74	70.3	41.56	50.00	18.77	25.23	69.67	63.47	15.22	70.00	54.90	54.46	15.71	98.15
WideResNet	61.38	59.67	53.13	76.22	70.26	30.92	66.67	19.38	2.82	69.27	58.36	5.19	76.00	43.56	49.75	12.79	110.06
Imbalanced samples																	
RF-F	37.97	32.54	21.61	28.12	47.93	1.47	0	0	7.69	46.38	48.31	0	0	0	19.46	0	279.63
RF-FE	46.65	42.21	33.62	52.12	51.54	8.86	0	4.85	7.14	55.46	54.24	0	12.50	1.50	32.45	3.73	580.23
LightGBM-F	42.47	37.73	27.48	37.81	54.27	5.59	0	1.15	4.00	48.19	45.61	0	6.45	0	35.52	0	279.42
LightGBM-FE	48.42	44.55	36.04	54.90	53.56	18.82	0	8.65	3.64	56.52	53.45	0	18.18	4.38	37.87	4.71	580.45
Balanced samples																	
RF-F	35.26	32.56	20.56	31.32	44.4	14.29	8.33	4.17	6.74	43.27	48.05	2.33	25.00	0	20.36	2.38	279.45
RF-FE	36.57	37.85	26.58	51.81	43.98	19.14	7.02	6.36	10.48	44.35	54.11	5.13	16.57	9.30	23.79	7.94	580.25
LightGBM-F	38.79	37.36	25.74	37.66	48.75	19.67	28.57	6.95	10.64	46.43	48.71	1.64	22.86	6.49	32.32	5.08	279.42
LightGBM-FE	40.26	40.98	30.37	53.61	45.29	21.23	15.38	14.72	11.36	48.95	54.96	8.18	25.81	12.69	30.63	6.58	580.51

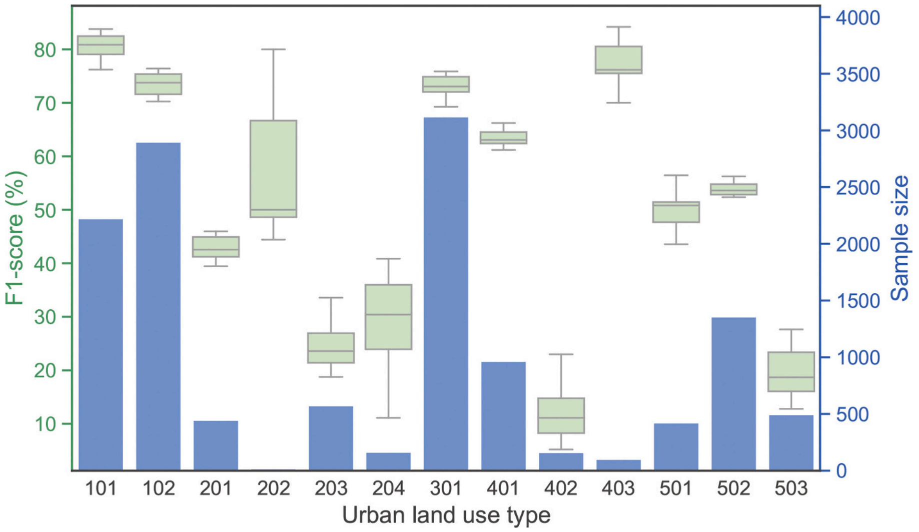

Fig. 10. Class-wise distribution (green box) of F1-score across different deep learning models and the amounts of samples (blue column) for each land use category. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accuracy of each scale group, similar trends of accuracy like overall performance can also be seen. Moreover, it can be found that not all the groups obtained their highest performance at 384 pixels size. For example, the images of scale group [200,300) got their best accuracy at 256 pixels size, while the highest accuracy of scale group [1500, $+\infty$) was achieved by a resampling size of 768 pixels. For small-scale images, the peak was reached at small resampling sizes. As the image scale increases, the optimal resampling size also increases. This is because, for small-scale images, a small resampling size is sufficient to contain effective information for land use classification. Although larger resampling sizes can help large-scale images make better decisions, they may bring no improvement in small scales but increase the unnecessary computational burden. Therefore, it suggests that multi-scale attributes of land use parcels should be carefully considered, adopting a more effective way such as developing a scale-robust model or multi-scale training rather than simply resampling to the same size.

3.3.4. *Size of training and test samples*

The quantitative results of deep learning models trained with different amounts of training samples are listed in Table 6. As the size of the training set gets larger, the overall performance experienced a rising trend with higher amplitudes in smaller sizes and lower amplitudes in larger sizes. It indicates that large training samples are helpful to improve the model's performance and the effect conforms to the law of diminishing marginal utility. It also demonstrates that even using a fairly small sample set such as only 10% of the original dataset can allow a deep learning model to achieve 60.94% in overall accuracy, which is much higher than the traditional machine learning algorithms shown before. Besides training size, we also investigated the impacts of different sizes of the test set on the classification performance. The test set is mainly used to validate the performance of models on a new dataset and thus the larger sample size tends to represent a higher confidence level of evaluation. From Fig. 12, we can see that a large variance of evaluation exists when the sample size is small. As the sample size increases, this evaluation uncertainty decreases. When the ratio of the test set exceeds 40%, the standard deviation of the evaluation metrics is $<$1%. This indicates that using less than half of the samples can estimate the model's performance with relatively low uncertainty. If users find this level of uncertainty acceptable, we can transfer some of the samples from the test set to the training set, allowing the model to benefit from more training data.

3.3.5. *Data input scheme*

As shown in Table 7, the performance of using a bounding box input scheme is better than only inputting the pixels inside the parcels. This suggests that including more context information in the input can help the model differentiate land use types. Then, we can also see that the third scheme, a bounding box with a parcel mask got the highest performance among the three schemes. It may be attributed to the input not only containing the nearby information but also providing a mask of the parcel indicating the area to which the model should pay more attention.

3.3.6. *Spatial transferability*

Transferability is one of the most important indicators that represent how well the model will perform in unseen regions, especially for mapping urban land use categories for country-scale or global-scale. We tested the spatial transferability of MobileNet across Shenzhen, and the statistical results are shown in Table 8. As a whole, the results demonstrate that even in a city, spatial transferability is still a serious problem which causes a major decrease of performance when a model trained in one region is applied to another region in the same city. Each source dataset achieves its best performance on the target datasets from the same region. Among night results, the model both trained and tested using Baoan datasets achieved the highest overall accuracy reaching 72.58%. For models trained using Baoan and Longgang datasets, they have better transferability in each other regions and they both achieved the worst performance on the OSZ dataset. It indicates that Baoan and Longgang share more common urban land use patterns. This similarity may be attributed to the fact that they are not the earliest special economic zone since 1979 like the OSZ region, which makes both two regions develop the urban area at a slower pace than OSZ. Fine-tuning technique effectively enhanced the trained models' performance in the target region, consistently yielding substantial improvement in OA, wF1, and Kappa coefficient across six cases. Notably, the fine-tuned models exhibited comparable or inferior performance relative to the performance of models trained from scratch in the test region.

3.3.7. *Levels of parcel hierarchy*

Table 9 shows that obvious differences exist between the performance of different parcel levels. Among them, Level-2 parcels obtained the highest accuracy, which is over 69%, followed by Level-3 and Level-1. Looking into the performance in specific classes, Level-1 indeed performs worst in some classes, which is consistent with the overall evaluation. Especially for 403 Harbor, Level-1 parcels degraded the detection of this class dramatically. However, Level-2 parcels are not

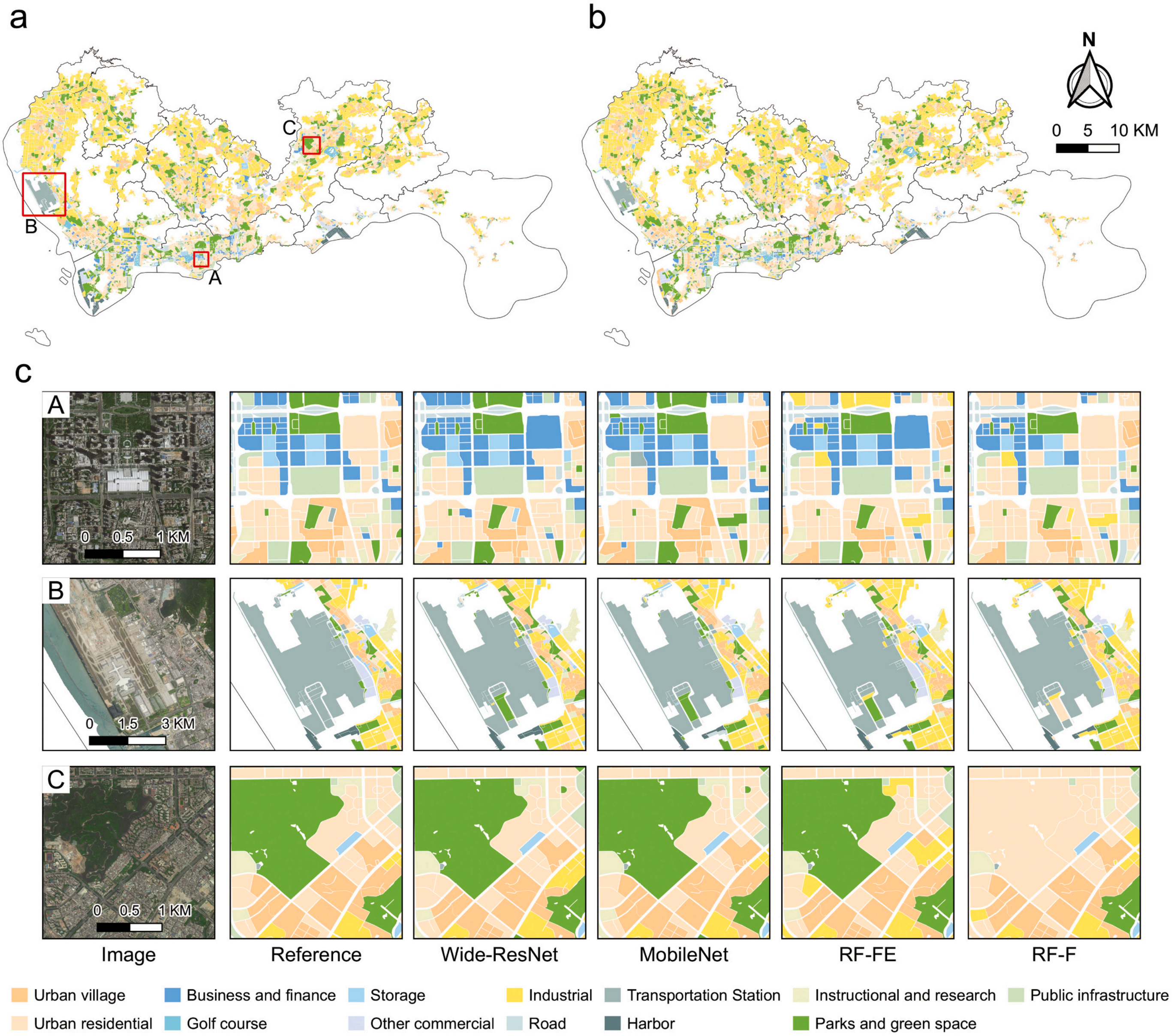

Fig. 11. Land use categories maps of Shenzhen: (a) reference map and (b) classification map by Wide-ResNet, and (c) three zoom-in subsets of high-resolution images, reference map, and classification maps using different classification methods.

always the most suitable level for some classes, such as 201 Business and finance and 202 Golf course. By contrast, better performance was obtained by using higher-level parcels as input.

3.3.8. Purity of land parcels

The performances of models trained and evaluated by data with different levels of purity are shown in Table 10. Among all results, the model trained with high-purity samples and tested in the same purity group achieved the best performance, with not only the highest overall accuracy (78.89%), but also the best F1-score for class 101, 102, 401, 402, 502, and 503. The medium-purity groups also achieved satisfying performance with an overall accuracy of 75.02%, ranking second in these results. For low-purity data, the model achieved better performance with an overall accuracy of 67.11% when it was tested in medium-purity data. For each training group, all of them performed worst on the low-medium test sets, which indicates that mixed land use is still an important issue that should be carefully considered and addressed in the land use mapping framework, otherwise it will significantly undermine the mapping accuracy. Moreover, we can see that models trained with higher-purity datasets obtain better OA, wF1, and Kappa in the entire test set. It suggests that when collecting reference data for developing deep learning models, the researcher should prioritize the collection of high-purity data in order to train a better model with limited labels.

4. Discussion

4.1. Insights from the experimental assessment

Our experimental investigation into urban land use classification in Shenzhen reveals that generally, deep learning methods outperform traditional machine learning methods, because of its automatic process of feature extraction, transitioning from low to high levels through multiple layers. The extraction of task-specific features is more advanced and descriptive than the manually defined features often used in traditional machine learning (Penatti et al., 2015). Upon a deeper dive into

Table 5
Classification accuracy (OA) and computational time cost of multi-scale samples under the supervision of different resampling scales.

	32	64	96	128	160	192	224	256	384	512	768	1024
Scale range												
[0, 200)	33.01	46.12	51.46	51.46	50.49	51.94	50.49	50.97	48.54	47.57	49.51	49.51
[200, 300)	42.39	59.23	59.23	62.68	61.46	61.05	63.08	64.71	62.07	60.45	62.07	62.47
[300, 400)	43.50	63.21	65.04	64.43	68.29	66.46	64.23	68.09	67.68	66.67	65.65	63.21
[400, 500)	46.72	62.93	63.13	64.48	66.99	66.60	66.02	64.67	70.08	68.92	66.22	67.18
[500, 600)	49.65	62.30	65.11	70.73	70.02	68.85	68.85	69.09	69.32	69.56	67.68	68.62
[600, 700)	47.60	71.86	73.95	71.26	74.55	72.46	78.14	75.15	78.14	75.75	75.45	73.95
[700, 800)	44.40	65.25	66.80	70.66	69.50	69.11	67.57	67.95	69.50	71.43	69.88	67.57
[800, 1000)	41.03	61.67	62.16	66.09	66.34	70.02	65.36	68.55	72.24	65.11	68.30	67.32
[1000, 1500)	36.49	53.83	60.81	60.81	62.84	63.51	64.41	60.36	66.22	62.84	65.32	63.06
[1500, +∞)	30.65	55.17	59.39	65.13	61.69	64.37	67.43	65.13	62.84	66.28	68.97	65.13
Overall	42.34	60.69	62.85	64.91	65.71	65.71	65.84	65.97	67.41	65.86	66.15	65.20
Inference Time (s)	71.38	71.64	72.48	89.01	96.38	96.84	100.1	102.22	109.35	121.35	153.88	203.45

Table 6
Classification performance of deep learning model trained on different ratio of training data.

Training Ratio	OA	wF1	Kappa	101	102	201	202	203	204	301	401	402	403	501	502	503
5	57.05	54.82	48.05	68.16	64.53	36.68	0	15.33	14.29	67.14	54.52	10.17	66.67	36.79	41.65	8.57
10	60.94	58.64	52.61	74.17	70.84	20.96	40	14.84	20.69	68.62	61.63	3.03	73.68	41.79	47.91	9.95
15	61.69	60.23	53.71	76.37	69.14	37.29	26.67	18.45	18.18	70.37	60.57	16.22	73.08	50	48.71	10.79
20	63.96	62.26	56.46	79.07	72.72	36.89	33.33	22.57	15.58	72.57	59.88	9.52	73.33	47.8	50.12	13.51
25	63.11	61.45	55.4	78.65	71.71	39.81	80	19.01	24.18	71.81	54.94	12.99	68.97	42.65	52.34	8.26
30	65.12	63.14	57.67	79.76	74.37	36.87	28.57	23.24	25	71.92	61.08	8.33	79.25	52.58	51.29	13.13
35	64.6	62.94	57.17	79.09	72.39	37.07	57.14	20.07	20.51	72.25	63.16	8.7	83.02	51.56	52.94	15.53
40	65.45	63.82	58.26	81.28	73.7	40.82	50	25.83	29.27	73.03	60.27	9.52	70.37	54.46	48.61	20.69
45	66.15	64.7	59.11	81.29	75.17	41.32	57.14	24.52	30.56	73.7	61.12	11.27	80.77	49.11	52.17	21.4
50	66.2	65.11	59.29	80.56	75.71	44.44	28.57	27.1	32.08	73.15	63.46	15.38	75.47	49.76	53.89	20.25
55	66.15	64.71	59.08	80.24	74.24	43.72	28.57	22.49	15.58	72.89	63.59	9.84	77.78	54.46	55.49	25.9
60	66.66	65.34	59.84	82.23	74.65	41	60	22.39	28.24	74.66	65.02	6.45	80.77	51.52	54.67	18.9
65	65.99	64.79	58.99	81.04	74.44	42.86	66.67	25.76	21.51	73.7	61.37	9.52	84.62	49.53	54.83	20.69
70	66.64	65.37	59.83	82.45	74.25	47.43	36.36	27.96	33.71	75.61	62.2	15.15	74.51	54.81	49.68	17.32
75	66.82	64.96	59.7	81.41	74.61	39.79	75	35.86	29.33	72.85	62.18	10.71	80	53.4	52.52	17.78
80	67.72	66.72	61.16	82.34	75.81	44.92	66.67	38.1	31.33	74.36	64.01	15.87	79.25	54.3	52.17	30.4
85	66.74	66	59.98	80.83	73.99	48.33	66.67	33.02	28.57	74.47	63.62	15.87	76	54.47	54.99	25.66
90	68.31	67.58	61.87	81.83	76.3	47.46	60	28.26	42.35	76.69	62.46	11.76	80.7	48.93	56	29.41
95	67.15	66.2	60.47	82.25	74.4	44.18	66.67	27.72	33.33	74.94	65.43	17.5	84.62	52.63	55.57	21.69
100	67.59	66.83	61.04	82.35	75.77	45.96	80	33.57	40.86	74.59	64.32	11.27	84.21	51.72	54.59	27.64

deep learning-based methods, we explored the impacts of various factors and configurations on urban land use classification in three primary areas: data, models, and samples.

Deep learning models are considerably influenced by the input data, specifically in the context of band composition and spatial resolution (Fan et al., 2021). Higher spatial resolutions offer rich and detailed spatial information, enabling deep learning models to learn more effective features related to land use functions (Huang et al., 2018). As such, high-resolution data can compensate for the accuracy deficit brought about by a reduced number of spectral bands and enhance classification outcomes. However, solely using high-resolution data without an effective classifier does not guarantee high accuracy. Our experiments have shown that the full potential of classification capabilities and satisfactory performance can be more potentially achieved by incorporating high-resolution data and deep learning models. Compared with Su et al. (2020)'s work that classified Shenzhen land use categories using machine learning and multi-source data including multispectral images, human activity features (Tencent Mobile-phone locating-based service data), POI data, nighttime light data, and building survey, our method, which employs only deep learning models and high-resolution RGB images can achieve comparable classification outcomes. This suggests that, through deep learning models, high-resolution images can, to some extent, supplant multi-source data in urban land use mappings. By feeding into very-high-resolution RGB images only, Zhong et al. (2023) also indicated that deep learning is the key to bridging the gap between high-resolution remote sensing data and land use functions in the tasks of scene-based land use classification. Nevertheless, for certain classifications, integrating diverse data sources would be beneficial to further improve accuracy. For example, categories such as *402 Transportation Station* and *501 Instructional and research* have low accuracies across all models in this study. The reason is their similar visual representation in overhead areal imagery, irrespective of its spatial resolution. Using only spectral images to distinguish them can be a challenge, even with a powerful deep learning model. In previous studies using traditional classifiers for land use classification (Sun et al., 2020; Tu et al., 2020), incorporating additional data such as POIs that record the categories of urban objects like schools, train stations, and airports, can help the classifier to assign correct land use categories to these indistinguishable land parcels. When applying deep learning, this finding still aligns with the result of the experiment

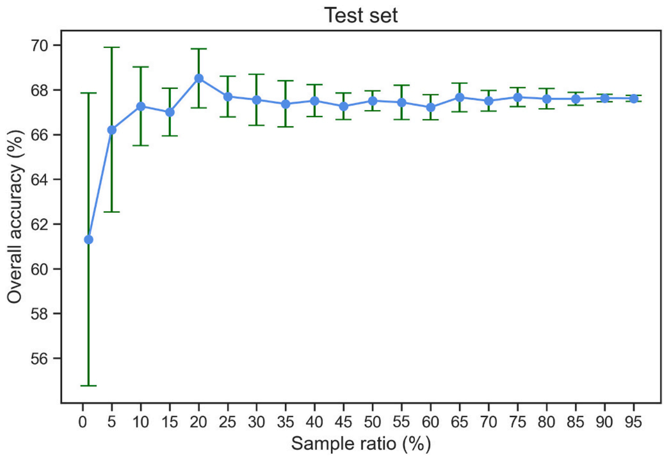

Fig. 12. The overall accuracy of the models using different ratios of the test sample set. The blue line represents the mean value of performances evaluated with random sampling test sets while the range of the green bar represents the variance of them. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which adds POI density as auxiliary data for worldview images.

Although deep learning models significantly outperform traditional shallow machine learning methods owing to the hierarchical learning mechanism of deep neural networks (Hu et al., 2015; Zhao and Du, 2016), the results also indicate that there is no one-size-fits-all deep learning model that guarantees the best performance across all categories. The selection of the model requires a trade-off among overall accuracy, computational efficiency, and performance in categories of interest, depending on the requirements of the tasks and applications. Regarding the patch size of samples, the overall classification accuracy dropped dramatically only in the extreme scenarios; otherwise, variations in accuracy were minimal. A similar empirical finding was also summarized by Hu et al. (2015) who tested the effect of resizing data with three scales into one scale. Besides, samples with different scales also have their optimal resampling scale. Mainstream solutions are designing effective model architectures that excavate multi-resolution and multi-scale information and incorporate them together as the features for classification (Du et al., 2021; Liu et al., 2018a). Due to the difficulty of obtaining labeled data for the whole city, few studies have been able to comprehensively probe into the impact of training and test set sizes on the classification performance based on the complete dataset. Zhao and Du (2016) investigated the effect of increasing data volume in a small and discrete label set and observed a continuous rise in model accuracy when the size of the training set increased. In contrast, with the complete dataset, Su et al. (2020) depicted a curve that rapidly increases at an early stage and then gradually levels off, exhibiting the same trend as our deep learning-based results. The results from Experiment 5 confirmed that including additional information about a target parcel's contexts and shape can boost classification performance. This finding suggests the need to consider spatial relations between the target object and its surroundings for accurate urban function inference. Except for the straightforward methods utilized here for adding contextual information, more advanced approaches such as designing spatial context-aware modules (Zhao et al., 2017b; Zhao et al., 2022b) and employing graph structures' topological relationship (Fang et al., 2022; Xu et al., 2022b) to leverage spatial contextual information in identifying land characteristics have been demonstrated to significantly improve accuracy as well. Model transferability, indicating a model's ability to perform in unfamiliar regions without training data, is also crucial. The results revealed that models show better transferability when the source and target regions are closely matched in natural and socioeconomic characteristics. It aligns with insights from the cross-city scale investigation into land use classification model transferability conducted by Chen et al. (2021a), which also identified regional similarity as a key determinant of transferability. The finding could guide model development for new regions by leveraging the proxy of region similarity and provides a unique lens through which we might understand how AI perceives cities.

Regarding different levels of parcel hierarchy, the variation mainly arises from the effect of increasing region sizes and land use mixture within the small region. Accounting for mixed land uses, there is a need for careful sample selection and labeling. Typically, training samples represent only a fraction of entire cities or regions to lessen the cost of both labor and time in most of the mapping tasks (Gong et al., 2020; Guzder-Williams et al., 2023). Thus, collecting high-quality samples is essential for model building. Experiment 8 indicated that emphasizing samples with consistent land use promotes better classification. For cities with a high mix of urban land use parcels, including a moderate number of mixed samples can aid in identifying low-purity parcels. However, parcels with very low purity should be excluded to prevent model contamination and accuracy reduction. In the meantime, models with soft label supervision and multi-label classification might offer a feasible solution to make full use of these mixed parcels (Hua et al., 2019; Wu et al., 2022).

All eight experiments are conducted in Shenzhen, and the primary reason for selecting this main site is the comprehensive land use inventory and extensive dataset we have collected. This inventory meticulously records urban land use functions at two levels of detail throughout the city, providing a robust foundation for building effective data-driven models. The abundance of precisely annotated labels makes Shenzhen an ideal testbed for examining how variations in input data, sample sizes, and models influence classification outcomes. The inventory is particularly valuable not only for its detailed categorization of land use types but also for its documentation of the degree of urban function mixtures— a distinctive and crucial characteristic that allows for an investigation into the impact of land use diversity on urban function identification. Moreover, since becoming a special economic zone in 1978, Shenzhen has undergone rapid and spatially diverse urbanization, resulting in a complex tapestry of land use patterns with varied urban functions (Ng, 2011; Xiao et al., 2023). This makes it an exemplary location for testing the spatial transferability of our models. However, it is important to note that as Shenzhen is just one of many megacities globally, the findings and conclusions drawn from this experimental comparison might not be universally applicable. Nonetheless, the experiments conducted here should provide a reference framework to assist researchers and practitioners in identifying optimal configurations for mapping urban land use patterns with deep learning techniques in unfamiliar cities. Looking ahead, it is imperative to invest more effort in gathering accurate and extensive urban land use data from cities with different socioeconomic and political backdrops worldwide. Such data will enable the derivation of broader insights and more universally applicable guidance from our studies.

Table 7
Classification performance of deep learning models with three kinds of inputs.

Input Scheme	OA	wF1	Kappa	101	102	201	202	203	204	301	401	402	403	501	502	503
Parcel	67.59	66.83	61.04	82.35	75.77	45.96	80.00	33.57	40.86	74.59	64.32	11.27	84.21	51.72	54.59	27.64
BBox	67.80	66.59	61.13	81.69	75.44	48.10	40.00	27.76	51.16	77.23	64.40	27.40	85.19	53.27	46.97	26.67
BBox+Mask	68.19	67.12	61.64	81.83	76.30	47.46	60.00	28.26	42.35	76.69	62.46	11.76	80.7	48.93	56.00	29.41

Table 8
Classification performance on target datasets using models trained with source datasets from the same or different sub-regions.

Source	Target	OA	wF1	Kappa	101	102	201	202	203	204	301	401	402	403	501	502	503
Baoan	Baoan	72.58	71.42	65.48	83.04	70.21	17.39	0	47.5	53.33	81.82	71.43	41.38	66.67	55.56	58.56	32.73
	Longgang	51.12	50.80	40.52	62.71	48.13	10.71	0	14.56	10.17	62.67	61.11	1.6	0	29.56	41.02	4.79
	Longgang*	70.32	68.58	62.64	85.56	72.29	0	100	22.64	0	79.09	65.00	0	0	51.85	48.68	23.53
	Original	46.72	42.66	35.53	47.45	67.11	15.3	14.29	15.27	9.9	34.95	50.7	16.39	0	41.19	48.04	13.43
	Original*	64.54	62.86	55.06	56.34	78.43	51.92	66.67	21.82	50.0	48.98	72.11	16.67	89.47	50.6	65.17	25.64
Longgang	Longgang	70.32	68.6	62.67	83.54	75	14.29	100	21.05	0	76.27	73.39	0	0	54.05	52	18.75
	Baoan	59.55	58.74	49.17	71.72	50.71	6.29	0	21.76	2.67	74.63	55.34	5.83	0	56.63	44.97	1.83
	Baoan*	70.98	69.68	63.51	84.28	69.19	23.08	0	32.50	47.06	80.77	59.35	16.00	66.67	62.50	62.44	27.45
	Original	44.89	39.88	33.29	38.75	66.93	15.31	21.05	13.29	6.4	30.91	54.72	0	0	43.03	46.54	7.87
	Original*	64.28	62.74	54.82	71.43	80.63	55.93	66.67	29.09	57.14	44.44	64.23	23.53	78.95	42.22	60.11	21.21
Original	Original	65.31	63.88	56.47	66.67	79.61	44.8	66.67	14.81	60.61	56.41	69.5	36.36	89.47	46.91	64.52	36.36
	Baoan	57.44	47.43	48.19	71.15	62.19	11.15	0	20.19	5.13	68.63	54.88	10.69	0	37.33	50.93	13.99
	Baoan*	70.13	68.71	62.19	82.06	67.06	27.27	0	47.50	42.86	77.90	71.83	25.81	0	51.61	55.56	31.37
	Longgang	52.65	44.58	43.18	69.16	58.77	15.13	25	8.66	2.5	56.51	59.21	5.8	0	36.36	42.77	18.45
	Longgang*	68.23	67.34	60.36	82.23	70.69	0	66.67	39.39	0	76.25	70.37	0	0	54.05	46.91	15.79

The target region with * represents that additional training samples are used to fine-tune the model pretrained in the source region.

4.2. *Challenges in deep learning for urban land use classification*

4.2.1. *Imbalance usage of data sources*

The majority of existing deep learning-based methods in urban land use classification emphasize imagery such as remote sensing images and street-view photos. Limited studies incorporate alternative data, including POI data and human activities data. Two major factors contribute to this trend: first, CNNs have garnered substantial attention due to their capabilities in computer vision and earth science (Gu et al., 2018; Guo et al., 2016; Kattenborn et al., 2021). These networks are naturally aligned with spatially continuous data, making them ideal for imagery. As a result, much effort has been directed toward applying and improving CNN-based models for these tasks, often sidelining other data types that might be better processed using RNNs or GNNs. Second, for deep learning models, it is vital to have sufficient input data paired with accurate labels. With the increasing availability of remote sensing data from different sensors and platforms, the wealth of data, combined with its easy access and rich contexts, for example, the ISPRS Vaihingen & Potsdam dataset, the Geofen Image dataset (Tong et al., 2020), and the DeepGlobe 2018 dataset (Demir et al., 2018), all provide image-label combinations for researchers to quickly test and validate their novel approaches to refining CNN models. Consequently, such open datasets significantly boost the progress in image-based land use classification models.

4.2.2. *Computational costs*

In contrast to traditional machine learning, deep learning has higher computational demands, particularly when dealing with large-scale, high-resolution remote sensing images. While it is possible to manage city- or country-scale research using basic neural network architecture on local computational devices, challenges arise with increasing data sizes, more complex models, or expanded study areas. These challenges often push local devices beyond their computational limits, resulting in slow processing times. High-performance computing (HPC) with specialized computer clusters can cater to the intense computational needs of deep learning models. However, setting up and maintaining these HPC facilities is often beyond the reach of many researchers (Ma et al., 2015b). This is where cloud-computing platforms like Google Earth Engine (GEE), Amazon Web Services (AWS), and Planetary Computer step in (Gorelick et al., 2017; Xu et al., 2022a). These platforms leverage cutting-edge HPC techniques such as parallel and distributed computing, and offer a plethora of tools, algorithms, and datasets that empower researchers to efficiently undertake large-scale earth science projects (Gupta et al., 2013; Yang et al., 2018). Regarding deep learning algorithms, platforms such as GEE and Planetary Computer have set protocols for their cloud-based implementation. However, due to the limited flexibility and the high costs associated with computational resources, only a handful of researchers fully developed and deployed their deep learning models on these platforms.

4.2.3. *Sample collection*

The law of large numbers indicates that as the sample size grows, the average of the results should approach the expected value. This principle is fundamental in machine learning. Traditional machine learning models tend to stabilize in performance when data reaches a certain volume, while deep learning models continue to benefit from even larger datasets (Alom et al., 2019). This underscores the importance of sample collection in creating effective and generalizable deep learning models. With the emerging concept of data-centric AI, the efficacy of deep learning models is largely determined by the quality, quantity, and reliability of data (Jarrahi et al., 2023). However, collecting large amounts of high-quality and reliable urban land use samples poses more challenges than natural image recognition or land cover classification tasks. First, determining the scale of sample collection is intricate. The scale should align with mapping requirements, with the pixel being the finest mapping unit (Zhang et al., 2018a; Zhou et al., 2020). Many deep

Table 9
Classification performance of deep learning model on different parcel levels.

Parcel level	OA	wF1	Kappa	101	102	201	202	203	204	301	401	402	403	501	502	503
Level-1	65.68	64.5	58.31	69.41	77.92	50	100	22.86	37.5	77.14	56.21	25	0	41.38	55.38	11.76
Level-2	69.28	67.9	62.5	82.59	75.83	38.83	50	26.09	29.09	79.57	53.2	24.24	80.85	52.73	58.17	19.8
Level-3	67.59	66.83	61.04	82.35	75.77	45.96	80.00	33.57	40.86	74.59	64.32	11.27	84.21	51.72	54.59	27.64

Table 10
Classification performance of deep learning model on different purity levels of parcels.

Train	Test	OA	wF1	Kappa	101	102	201	203	301	401	402	501	502	503
>90	>90	78.89	78.17	74.11	94.15	86.55	41.18	35.29	81.45	78.46	53.33	43.33	65.08	43.37
	60–90	66.44	64.38	58.6	80.79	72.37	51.52	27.69	80.67	32.38	0	50.91	44.72	28.07
	≤60	40.71	41.98	28.65	47.25	46.69	17.39	8.16	56.15	41.46	0	25	19.2	7.41
	Overall	65.18	63.52	57.51	79.82	72.82	38.36	24.24	74.29	60.85	17.39	41.29	43.2	28.87
60–90	>90	65.18	64.99	57.76	86	75.09	39.29	36.36	75.6	28.38	40	43.48	42.97	21.43
	60–90	75.02	73.94	69.15	86.36	83.78	45.45	41.98	83.17	37.29	16.67	72.41	71.09	13.33
	≤60	50.65	50.99	40.6	61.54	54.84	19.51	18.87	63.22	11.54	0	45.83	50.79	18.18
	Overall	65.10	63.24	57.49	80.57	73.89	36.81	34.0	75.48	27.03	18.6	55.26	54.27	17.93
≤60	>90	54.43	50.36	45.53	80.42	63.51	27.59	12.82	62.87	38.39	19.05	15.57	40.37	2.94
	60–90	67.11	64.11	60.19	83.5	70.05	45.07	28.26	79.43	43.01	28.57	62.96	62.16	21.82
	≤60	50	48.86	40.69	57.42	45.85	9.3	25.81	63.64	50	16.67	50	52	28.07
	Overall	57.81	58.09	49.57	76.02	62.41	30.23	22.41	69.63	41.73	20.69	45.59	51.56	16.67

learning models aiming for pixel-wise dense segmentation require large numbers of training image patches with complete annotations for every pixel, which demands a significant manpower commitment. For tasks focused on objects or scenes, larger units are more likely to grapple with the mixed land use problem, potentially compromising sample quality (Du et al., 2021). Second, urban land use types often exhibit high similarities between classes and certain variations within a class, making it difficult to differentiate urban functions solely on image interpretation (Zhu et al., 2022). To ensure sample reliability, the integration of multiple data sources or even onsite investigation is required, which can largely reduce annotation efficiency. Third, the collection of urban land use samples comes with specific prerequisites, requiring annotators with relevant expertise in urban studies and data interpretation (Guzder-Williams et al., 2023). Moreover, a consistent and comprehensive annotation guideline is essential to mitigate label ambiguity and personal biases. Apart from the manual labeling, open data portals with LULC labels and land use references from government agencies are also feasible ways to collect samples.

4.2.4. Barriers to generating large-scale consistent urban land use classification products

Many studies leveraging deep learning methods for urban land use mapping often restrict their focus to a few selected cities or regions. The efficacy of these proposed models on a broader scale remains unverified. One critical factor that underpins the reliability of mapping methods for untested cities is the models' generalizability. A model that excels not just in the study area but also exhibits strong adaptability elsewhere is ideal for mapping urban land use on a regional and global scale (Srivastava et al., 2019). On the other hand, given that the current research continues to enhance model performance on local scales, one possible alternative might be to combine these localized results to form a global output. A challenge, though, is the lack of consistency in the definitions or classification schemes for urban land use categories across different studies, which will further impede data aggregation (Yang et al., 2021). While categories with a hierarchical structure can be grouped into a broader superclass, this will also sacrifice the detailed granularity.

4.3. Future directions

4.3.1. Establishing a global sample library

Establishing a global sample library is of great significance for mapping multi-scale urban land use categories across various regions or countries. The diversity and representativeness of these samples critically affect the generalizability of trained models (Ma et al., 2018; Su et al., 2020). However, collecting samples aligning with these requirements is laborious and time-consuming. We here suggest three potential strategies to expedite this process. First, existing land use information, ranging from VGI samples to self-organized sampling campaigns, and government-sponsored land-use maps, can be sourced and streamlined into expansive sample libraries through crowdsourcing. Second, time-series analysis techniques, such as inter-calibration and change detection, can be used to expand sample sizes by identifying consistent and stable data segments over time (Gong et al., 2019; Huang et al., 2020a). Third, given the advent of well-trained land use classification models over the past few years, these models can be utilized to swiftly generate land use annotation in areas lacking samples. Though some errors might be evident, they still provide certain references, making the labeling tasks less labor-intensive. Moreover, generative models can simulate realistic data, proving effective in supplementing sample sizes and enriching data diversity, as evidenced in studies of building and road classifications (Chen et al., 2022a; Lv et al., 2021).

4.3.2. Modeling with limited samples

Several opportunities can further improve land use classification performance by leveraging advanced deep learning models tailored for diverse situations. For example, semi-supervised learning (SSL), which acts as a bridge between supervised learning and unsupervised learning, can integrate both a vast pool of unlabeled data and a smaller set of unlabeled data for specific tasks (van Engelen and Hoos, 2020). Weakly supervised learning (WSL) aims to construct predictive models by learning with weak supervision (Zhou, 2018). It can undertake fine-grained tasks with broad labels, such as image-based, point-based, line-based, and polygon-based annotations (Yue et al., 2022). This approach simplifies the task of land use sample annotation and paves the way for facilitating multi-scale land use mappings. Moreover, many existing labeled samples, especially those sourced from crowdsourcing, suffer from issues of inconsistent quality and uncertainty. With manual verification of each sample being impractical, there is a pressing need to develop deep learning strategy with noisy labels. This would empower deep learning models to minimize the effects of such noisy labels and converge to the optimal parameters. Transfer learning aims at transferring knowledge learned from the source domain to the target domain to enhance models' generalization (Ma et al., 2024). The application of transfer learning can span across different tasks, spatial scales, and time periods, offering a promising way for mapping large-scale urban land use dynamics by taking advantage of knowledge from specific regions and temporal snapshots. Last but not least, drawing inspiration from the remarkable success of large language models, a series of foundation models for geospatial data were developed and proposed recently, such as Prithvi (Jakubik et al., 2023) by NASA and IBM, and Ringmo (Sun et al., 2023). These models that are trained on large amounts of unlabeled datasets via self-supervised learning will present a significant direction for deepening our understanding of urban environments when fine-tuned for specific tasks.

4.3.3. *Interpretable AI*

Despite the impressive accuracy across different tasks, deep learning models have often been criticized for their black-box nature, which obscures their interpretation and decision-making value (Hosseiny et al., 2022; Montavon et al., 2018). Therefore, increasing the interpretability of models is not only crucial for generating reliable land use classification products but also central to collective collaborations among different sectors. Moreover, delving deeper into the mechanisms by which a model can get its corresponding predictions, can help scientists and researchers to unveil new insights (Chen et al., 2023a; Reichstein et al., 2019). These discoveries can inform the development of more effective models, further enhancing the performance in land use classification tasks.

5. Conclusions

The emerging deep learning methods have proven to be powerful tools for understanding land use information within urban areas in recent years. Considering the current void of reviews placing focus on deep learning-based urban land use classification, we undertook a comprehensive survey on the advances of research by literature review and empirical analysis. This study examined the models, data, mapping units, and parameter settings and analyzed their impacts on the classification performance, aiming to provide in-depth and exhaustive guidance for mapping practices. Extensive assessments of the Shenzhen dataset, which investigated the impact of various factors on performance, revealed several key insights into urban land use mapping practices using deep learning. Higher spatial resolution and additional spectral bands significantly enhance classification accuracy. Integrating complementary data can improve accuracy when added to lower-resolution data, yet may reduce performance in very high-resolution datasets due to the potential introduction of noise. Although deep learning models perform better than traditional machine learning methods, complex deep learning models do not always outperform simpler ones. The characteristics of foundational model architectures should match the nature of the data being processed to maximize effectiveness. The resampling scale of input data affects both accuracy and computing time. Samples of different scales exhibit their own optimal resampling size. Scale-robust architectures and multi-scale training strategies should be carefully considered when using deep learning to analyze samples of diverse image scales. Larger training sample sizes effectively enhance deep learning model classification accuracy with diminishing benefits, while larger test samples ensure a more stable and reliable estimation of model performance with decreasing uncertainty in classification evaluation. A better trade-off between the sizes of training and the test set is the key to ensuring the model efficacy and evaluation uncertainty. Spatial transferability remains a significant challenge in urban land use classification. The deep learning model still exhibited significant performance degradation in new regions without fine-tuning. A good transferability practice of land use classification should consider the similarity between regions. Due to the issue of mixed land use, models trained on high-purity data can achieve the best performance, highlighting the importance of prioritizing high-purity samples for training to effectively enhance mapping accuracy. These findings collectively advance our understanding of the factors for a successful urban land use classification, providing a clear framework for future practices. Furthermore, we summarized the remaining key challenges from four aspects: the imbalanced development of data, computational costs, sample collection, and barriers to generating large-scale consistent products. As an evolving field in urban studies, more efforts are expected in the compilation of a global sample library, modeling with limited samples, and exploring the model's interpretability and transparency. Overall, this review will hopefully inform researchers and practitioners of guiding regional to global mapping practices, facilitating global environmental changes and sustainable land management.

CRediT authorship contribution statement

Ziming Li: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation. **Bin Chen:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization. **Shengbiao Wu:** Writing – review & editing. **Mo Su:** Writing – review & editing, Data curation. **Jing M. Chen:** Writing – review & editing. **Bing Xu:** Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This study was supported by the National Natural Science Foundation of China Young Scientists Fund (42201373), National Key Research and Development Program of China (2022YFB3903703), The University of Hong Kong HKU-100 Scholars Fund, Seed Fund for Basic Research, Strategic Interdisciplinary Research Scheme Fund, the International Research Center of Big Data for Sustainable Development Goals (CBAS2022GSP04), and the Croucher Foundation (CAS22902/CAS22HU01). The authors would like to thank Professor Peng Gong for his constructive suggestions and insightful discussion on this work. The authors would also thank the editor and four anonymous reviewers for providing valuable comments and suggestions, which have greatly improved this work.

References

Abbas, S., Peng, Q., Wong, M.S., Li, Z., Wang, J., Ng, K.T.K., Kwok, C.Y.T., Hui, K.K.W., 2021. Characterizing and classifying urban tree species using bi-monthly terrestrial hyperspectral images in Hong Kong. ISPRS J. Photogramm. Remote Sens. 177, 204–216.

Aleissaee, A.A., Kumar, A., Anwer, R.M., Khan, S., Cholakkal, H., Xia, G.S., Khan, F.S., 2023. Transformers in remote sensing: a survey. Remote Sens. 15, 1860.

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., Hasan, M., Van Essen, B.C., Awwal, A.A.S., Asari, V.K., 2019. A state-of-the-art survey on deep learning theory and architectures. Electronics 8, 292.

Ansith, S., Bini, A.A., 2022. Land use classification of high resolution remote sensing images using an encoder based modified GAN architecture. Displays 74, 102229.

Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning - a new frontier in artificial intelligence research. IEEE Comput. Intell. Mag. 5, 13–18.

Arino, O., Gross, D., Ranera, F., Leroy, M., Bicheron, P., Brockman, C., Defourny, P., Vancutsem, C., Achard, F., Durieux, L., Bourg, L., Latham, J., Gregorio, A.D., Witt, R., Herold, M., Sambale, J., Plummer, S., Weber, J.L., 2007. GlobCover: ESA service for global land cover from MERIS. In: 2007 IEEE International Geoscience and Remote Sensing Symposium, pp. 2412–2415.

Atkinson, P.M., Tatnall, A.R.L., 1997. Introduction neural networks in remote sensing. Int. J. Remote Sens. 18, 699–709.

Barnsley, M.J., Barr, S.L., 1996. Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. Photogramm. Eng. Remote. Sens. 62, 949–958.

Bartholomé, E., Belward, A.S., 2005. GLC2000: a new approach to global land cover mapping from earth observation data. Int. J. Remote Sens. 26, 1959–1977.

Biljecki, F., Ito, K., 2021. Street view imagery in urban analytics and GIS: a review. Landsc. Urban Plan. 215, 104217.

Bin, J., Gardiner, B., Li, E., Liu, Z., 2020. Multi-source urban data fusion for property value assessment: a case study in Philadelphia. Neurocomputing 404, 70–83.

Bischof, H., Schneider, W., Pinz, A.J., 1992. Multispectral classification of Landsat-images using neural networks. IEEE Trans. Geosci. Remote Sens. 30, 482–490.

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

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., Rampini, A., 2014. A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. Inf. Sci. 258, 312–327.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Campos-Taberner, M., García-Haro, F.J., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilabert, M.A., 2020. Understanding deep learning in land use classification based on Sentinel-2 time series. Sci. Rep. 10, 17188.

Chan, C.S., Anderson, D.T., Ball, J.E., 2017. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. J. Appl. Remote. Sens. 11, 042609.

Chen, Z., Zhang, Y., Guindon, B., Esch, T., Roth, A., Shang, J., 2013. Urban land use mapping using high resolution SAR data based on density analysis and contextual information. Can. J. Remote. Sens. 38, 738–749.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30m resolution: a POK-based operational approach. ISPRS J. Photogramm. Remote Sens. 103, 7–27.

Chen, B., Chen, L., Lu, M., Xu, B., 2017a. Wetland mapping by fusing fine spatial and hyperspectral resolution images. Ecol. Model. 353, 95–106.

Chen, Y., Liu, X., Li, X., Liu, X., Yao, Y., Hu, G., Xu, X., Pei, F., 2017b. Delineating urban functional areas with building-level social media data: a dynamic time warping (DTW) distance based k-medoids method. Landsc. Urban Plan. 160, 48–60.

Chen, T., Hui, E.C.M., Wu, J., Lang, W., Li, X., 2019. Identifying urban spatial structure and urban vibrancy in highly dense cities using georeferenced social media data. Habitat Int. 89, 102005.

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 5, 410–441.

Chen, H., Li, W., Shi, Z., 2022a. Adversarial instance augmentation for building change detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Chen, H., Qi, Z., Shi, Z., 2022b. Remote sensing image change detection with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Chen, M., Qian, Z., Boers, N., Jakeman, A.J., Kettner, A.J., Brandt, M., Kwan, M.-P., Batty, M., Li, W., Zhu, R., Luo, W., Ames, D.P., Barton, C.M., Cuddy, S.M., Koirala, S., Zhang, F., Ratti, C., Liu, J., Zhong, T., Liu, J., Wen, Y., Yue, S., Zhu, Z., Zhang, Z., Sun, Z., Lin, J., Ma, Z., He, Y., Xu, K., Zhang, C., Lin, H., Lü, G., 2023a. Iterative integration of deep learning in hybrid earth surface system modelling. Nat. Rev. Earth Environ. 4, 568–581.

Chen, Y., He, C., Guo, W., Zheng, S., Wu, B., 2023b. Mapping urban functional areas using multisource remote sensing images and open big data. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 16, 7919–7931.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In (p. arXiv:1406.1078).

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D.B., Ermon, S., 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In (p. arXiv:2207.08051).

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Process. Mag. 35, 53–65.

Dai, D., Yang, W., 2011. Satellite image classification via two-layer sparse coding with biased image representation. IEEE Geosci. Remote Sens. Lett. 8, 173–176.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (pp. 886-893).

Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. DeepGlobe 2018: a challenge to parse the earth through satellite images. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 172–181.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J., 2014. Dynamic population mapping using mobile phone data. Proc. Natl. Acad. Sci. 111, 15888–15893.

DeVries, W., 1928. The Michigan land economic survey. J. Farm Econ. 10.

Dixon, B., Candade, N., 2007. Multispectral landuse classification using neural networks and support vector machines: one or the other, or both? Int. J. Remote Sens. 29, 1185–1206.

Dong, L., Du, H., Mao, F., Han, N., Li, X., Zhou, G., Zhu, D.E., Zheng, J., Zhang, M., Xing, L., Liu, T., 2020. Very high resolution remote sensing imagery classification using a fusion of random Forest and deep learning technique—subtropical area for example. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 13, 113–128.

Dong, R., Fang, W., Fu, H., Gan, L., Wang, J., Gong, P., 2022. High-resolution land cover mapping through learning with noise correction. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

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. In (p. arXiv:2010.11929).

Du, S., Du, S., Liu, B., Zhang, X., Zheng, Z., 2020. Large-scale urban functional zone mapping by integrating remote sensing images and open social data. GIScience Remote Sens. 57, 411–430.

Du, S., Du, S., Liu, B., Zhang, X., 2021. Mapping large-scale and fine-grained urban functional zones from VHR images using a multi-scale semantic segmentation network and object based approach. Remote Sens. Environ. 261, 112480.

Du, S., Zheng, M., Guo, L., Wu, Y., Li, Z., Liu, P., 2024. Urban building function classification based on multisource geospatial data: a two-stage method combining unsupervised and supervised algorithms. Earth Sci. Inf. 17, 1179–1201.

Fan, Y., Ding, X., Wu, J., Ge, J., Li, Y., 2021. High spatial-resolution classification of urban surfaces using a deep learning method. Build. Environ. 200.

Fan, Z., Zhang, F., Loo, B.P.Y., Ratti, C., 2023. Urban visual intelligence: uncovering hidden city profiles with street view images. Proc. Natl. Acad. Sci. USA 120, e2220417120.

Fang, F., Zeng, L., Li, S., Zheng, D., Zhang, J., Liu, Y., Wan, B., 2022. Spatial context-aware method for urban land use classification using street view images. ISPRS J. Photogramm. Remote Sens. 192, 1–12.

Findell, K.L., Berg, A., Gentine, P., Krasting, J.P., Lintner, B.R., Malyshev, S., Santanello, J.A., Shevliakova, E., 2017. The impact of anthropogenic land use and land cover change on regional climate extremes. Nat. Commun. 8, 989.

Flanagin, A.J., Metzger, M.J., 2008. The credibility of volunteered geographic information. GeoJournal 72, 137–148.

Foley, J.A., Costa, M.H., Delire, C., Ramankutty, N., Snyder, P., 2003. Green surprise? How terrestrial ecosystems could affect earth's climate. Front. Ecol. Environ. 1, 38–44.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science 309, 570–574.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. Remote Sens. Environ. 252, 112128.

Frias-Martinez, V., Frias-Martinez, E., 2014. Spectral clustering for sensing urban land use using twitter activity. Eng. Appl. Artif. Intell. 35, 237–245.

Friedl, M.A., McIver, D.K., Hodges, J.C.F., Zhang, X.Y., Muchoney, D., Strahler, A.H., Woodcock, C.E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sens. Environ. 83, 287–302.

Galesic, M., Bruine de Bruin, W., Dalege, J., Feld, S.L., Kreuter, F., Olsson, H., Prelec, D., Stein, D.L., van der Does, T., 2021. Human social sensing is an untapped resource for computational social science. Nature 595, 214–222.

Gao, M., Guo, H., Liu, L., Zeng, Y., Liu, W., Liu, Y., Xing, H., 2024. Integrating street view imagery and taxi trajectory for identifying urban function of street space. Geo-spat. Inf. Sci. 1–23.

Garg, R., Kumar, A., Bansal, N., Prateek, M., Kumar, S., 2021. Semantic segmentation of PolSAR image data using advanced deep learning model. Sci. Rep. 11, 15365.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E.L., Fei-Fei, L., 2017. Using deep learning and Google street view to estimate the demographic makeup of neighborhoods across the United States. Proc. Natl. Acad. Sci. USA 114, 13108–13113.

Gong, P., Howarth, P.J., 1992. Land-use classification of SPOT HRV data using a cover-frequency method. Int. J. Remote Sens. 13, 1459–1471.

Gong, P., Liang, S., Carlton, E.J., Jiang, Q., Wu, J., Wang, L., Remais, J.V., 2012. Urbanisation and health in China. Lancet 379, 843–852.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., Niu, Z., Huang, X., Fu, H., Liu, S., Li, C., Li, X., Fu, W., Liu, C., Xu, Y., Wang, X., Cheng, Q., Hu, L., Yao, W., Zhang, H., Zhu, P., Zhao, Z., Zhang, H., Zheng, Y., Ji, L., Zhang, Y., Chen, H., Yan, A., Guo, J., Yu, L., Wang, L., Liu, X., Shi, T., Zhu, M., Chen, Y., Yang, G., Tang, P., Xu, B., Giri, C., Clinton, N., Zhu, Z., Chen, J., Chen, J., 2013. Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. Int. J. Remote Sens. 34, 2607–2654.

Gong, P., Liu, H., Zhang, M., Li, C., Wang, J., Huang, H., Clinton, N., Ji, L., Li, W., Bai, Y., Chen, B., Xu, B., Zhu, Z., Yuan, C., Ping Suen, H., Guo, J., Xu, N., Li, W., Zhao, Y., Yang, J., Yu, C., Wang, X., Fu, H., Yu, L., Dronova, I., Hui, F., Cheng, X., Shi, X., Xiao, F., Liu, Q., Song, L., 2019. Stable classification with limited sample: transferring a 30-m resolution sample set collected in 2015 to mapping 10-m resolution global land cover in 2017. Sci. Bull. 64, 370–373.

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.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks. In (p. arXiv: 1406.2661).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. Pattern Recogn. 77, 354–377.

Guan, Q., Cheng, S., Pan, Y., Yao, Y., Zeng, W., 2021. Sensing mixed urban land-use patterns using municipal water consumption time series. Ann. Am. Assoc. Geogr. 111, 68–86.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., Lew, M.S., 2016. Deep learning for visual understanding: a review. Neurocomputing 187, 27–48.

Guo, H., Dou, C., Chen, H., Liu, J., Fu, B., Li, X., Zou, Z., Liang, D., 2023a. SDGSAT-1: the world's first scientific satellite for sustainable development goals. Sci Bull (Beijing) 68, 34–38.

Guo, Z., Wen, J., Xu, R., 2023b. A shape and size free-CNN for urban functional zone mapping with high-resolution satellite images and POI data. IEEE Trans. Geosci. Remote Sens. 61, 1–17.

Guo, Y., Tang, J., Liu, H., Yang, X., Deng, M., 2024. Identifying up-to-date urban land-use patterns with visual and semantic features based on multisource geospatial data. Sustain. Cities Soc. 101.

Gupta, A., Kale, L.V., Gioachin, F., March, V., Suen, C.H., Lee, B.-S., Faraboschi, P., Kaufmann, R., Milojicic, D., 2013. The who, what, why, and how of high performance computing in the cloud. In: 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, pp. 306–314.

Guzder-Williams, B., Mackres, E., Angel, S., Blei, A.M., Lamson-Hall, P., 2023. Intra-Urban Land Use Maps for a Global Sample of Cities from Sentinel-2 Satellite Imagery and Computer Vision. Computers, Environment and Urban Systems, p. 100.

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.

Han, W., Wang, L., Feng, R., Gao, L., Chen, X., Deng, Z., Chen, J., Liu, P., 2020. Sample generation based on a supervised Wasserstein generative adversarial network for high-resolution remote-sensing scene classification. Inf. Sci. 539, 177–194.

Han, K., Wang, Y., Guo, J., Tang, Y., Wu, E., 2024. Vision GNN: an image is worth graph of nodes. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (p. Article 603). Curran Associates Inc, New Orleans, LA, USA.

Hang, R., Liu, Q., Hong, D., Ghamisi, P., 2019. Cascaded recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57, 5384–5394.

Hansen, M.C., Defries, R.S., Townshend, J.R.G., Sohlberg, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. Int. J. Remote Sens. 21, 1331–1364.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. SMC-3, 610–621.

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, Y., Lee, E., Warner, T.A., 2017. A time series of annual land use and land cover maps of China from 1982 to 2013 generated using AVHRR GIMMS NDVI3g data. Remote Sens. Environ. 199, 201–217.

He, C., Fang, P., Zhang, Z., Xiong, D., Liao, M., 2019. An end-to-end conditional random fields and skip-connected generative adversarial segmentation network for remote sensing images. Remote Sens. 11, 1604.

He, H., Lin, X., Yang, Y., Lu, Y., 2020. Association of street greenery and physical activity in older adults: a novel study using pedestrian-centered photographs. Urban For. Urban Green. 55, 126789.

Heiden, U., Heldens, W., Roessner, S., Segl, K., Esch, T., Mueller, A., 2012. Urban structure type characterization using hyperspectral remote sensing and height information. Landsc. Urban Plan. 105, 361–375.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Hoffmann, E.J., Abdulahhad, K., Zhu, X.X., 2023. Using social media images for building function classification. Cities 133, 104107.

Hofmann, T., 1999. Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR coNference on Research and Development in Information Retrieval, pp. 50–57.

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2022. SpectralFormer: rethinking hyperspectral image classification with transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2, 359–366.

Hosseiny, B., Abdi, A.M., Jamali, S., 2022. Urban Land Use and Land Cover Classification with Interpretable Machine Learning – A Case Study Using Sentinel-2 and Auxiliary Data. Remote Sensing Applications: Society and Environment, p. 28.

Hou, Y., Biljecki, F., 2022. A comprehensive framework for evaluating the quality of street view imagery. Int. J. Appl. Earth Obs. Geoinf. 115, 103094.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In (p. arXiv:1704.04861).

Hu, F., Xia, G.-S., Hu, J., Zhang, L., 2015. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. Remote Sens. 7, 14680–14707.

Hu, S., Gao, S., Wu, L., Xu, Y., Zhang, Z., Cui, H., Gong, X., 2021. Urban function classification at road segment level using taxi trajectory data: a graph convolutional neural network approach. Comput. Environ. Urban. Syst. 87, 101619.

Hua, Y., Mou, L., Zhu, X.X., 2019. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. ISPRS J. Photogramm. Remote Sens. 149, 188–199.

Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pp. 2261–2269.

Huang, B., Zhao, B., Song, Y., 2018. 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, H., Wang, J., Liu, C., Liang, L., Li, C., Gong, P., 2020a. The migration of training samples towards dynamic global land cover mapping. ISPRS J. Photogramm. Remote Sens. 161, 27–36.

Huang, Z., Qi, H., Kang, C., Su, Y., Liu, Y., 2020b. An ensemble learning approach for urban land use mapping based on remote sensing imagery and social sensing data. Remote Sens. 12, 3254.

Huang, X., Yang, J., Li, J., Wen, D., 2021. Urban functional zone mapping by integrating high spatial resolution nighttime light and daytime multi-view imagery. ISPRS J. Photogramm. Remote Sens. 175, 403–415.

Huang, H., Huang, J., Chen, B., Xu, X., Li, W., 2024. Recognition of functional areas in an Old City based on POI: a case study in Fuzhou, China. J. Urban Plan. Dev. 150.

Ilieva, R.T., McPhearson, T., 2018. Social-media data for urban sustainability. Nat. Sustain. 1, 553–565.

Jafarbiglu, H., Pourreza, A., 2022. A comprehensive review of remote sensing platforms, sensors, and applications in nut crops. Comput. Electron. Agric. 197, 106844.

Jakubik, J., Roy, S., Phillips, C.E., Fraccaro, P., Godwin, D., Zadrozny, B., Szwarcman, D., Gomes, C., Nyirjesy, G., Edwards, B., Kimura, D., Simumba, N., Chu, L., Karthik Mukkavilli, S., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Hanxi, Li, Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R., 2023. Foundation Models for Generalist Geospatial Artificial Intelligence. In (p. arXiv:2310.18660.

Jarrahi, M.H., Memariani, A., Guha, S., 2023. The principles of data-centric AI. Commun. ACM 66, 84–92.

Javali, A., Gupta, J., Sahoo, A., 2021. A review on synthetic aperture radar for earth remote sensing: challenges and opportunities. In: 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 596–601.

Ji, S., Wang, D., Luo, M., 2021. Generative adversarial network-based full-space domain adaptation for land cover classification from multiple-source remote sensing images. IEEE Trans. Geosci. Remote Sens. 59, 3816–3828.

Johnson, N., Treible, W., Crispell, D., 2022. OpenSentinelMap: A large-scale land use dataset using OpenStreetMap and Sentinel-2 imagery. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE Computer Society, pp. 1332–1340.

Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M.R., Kuemmerle, T., Meyfroidt, P., Mitchard, E.T.A., Reiche, J., Ryan, C.M., Waske, B., 2016. A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. Remote Sens. 8, 70.

Jozdani, S., Chen, D., Pouliot, D., Alan Johnson, B., 2022. A review and meta-analysis of generative adversarial networks and their applications in remote sensing. Int. J. Appl. Earth Obs. Geoinf. 108, 102734.

Kalluri, H.R., Prasad, S., Bruce, L.M., 2010. Decision-level fusion of spectral reflectance and derivative information for robust hyperspectral land cover classification. IEEE Trans. Geosci. Remote Sens. 48, 4047–4058.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. 173, 24–49.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 3149–3157.

Kong, B., Ai, T., Zou, X., Yan, X., Yang, M., 2024. A Graph-Based Neural Network Approach to Integrate Multi-Source Data for Urban Building Function Classification. Computers, Environment and Urban Systems, p. 110.

Koukoletsos, T., Haklay, M., Ellul, C., 2012. Assessing data completeness of VGI through an automated matching procedure for linear data. Trans. GIS 16, 477–498.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Leng, J., Li, T., Bai, G., Dong, Q., Dong, H., 2016. Cube-CNN-SVM: a novel hyperspectral image classification method. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1027–1034.

Leung, D., Newsam, S., 2010. Proximate sensing: Inferring what-is-where from georeferenced photo collections. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2955–2962.

Levin, N., Johansen, K., Hacker, J.M., Phinn, S., 2014. A new source for high spatial resolution night time images — the EROS-B commercial satellite. Remote Sens. Environ. 149, 1–12.

Li, X., Zhang, C., Li, W., 2017. Building block level urban land-use information retrieval based on Google street view images. GIScience Remote Sens. 54, 819–835.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019a. Deep learning for hyperspectral image classification: An overview. IEEE Trans. Geosci. Remote Sens. 57, 6690–6709.

Li, T., Leng, J., Kong, L., Guo, S., Bai, G., Wang, K., 2019b. DCNR: deep cube CNN with random forest for hyperspectral image classification. Multimed. Tools Appl. 78, 3411–3433.

Li, W., Sun, K., Li, W., Wei, J., Miao, S., Gao, S., Zhou, Q., 2023a. Aligning semantic distribution in fusing optical and SAR images for land use classification. ISPRS J. Photogramm. Remote Sens. 199, 272–288.

Li, Z., He, W., Cheng, M., Hu, J., Yang, G., Zhang, H., 2023b. SinoLC-1: the first 1 m resolution national-scale land-cover map of China created with a deep learning framework and open-access data. Earth System Science Data 15, 4749–4780.

Lipton, Z.C., Berkowitz, J., Elkan, C., 2015. A Critical Review of Recurrent Neural Networks for Sequence Learning. In (p. arXiv:1506.00019).

Liu, X., Long, Y., 2016. Automated identification and characterization of parcels with OpenStreetMap and points of interest. Environ. Plan. B Plan Des. 43, 341–360.

Liu, Y., Wang, F., Xiao, Y., Gao, S., 2012. Urban land uses and traffic 'source-sink areas': evidence from GPS-enabled taxi data in Shanghai. Landsc. Urban Plan. 106, 73–87.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L., 2015. Social sensing: a new approach to understanding our socioeconomic environments. Ann. Assoc. Am. Geogr. 105, 512–530.

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, Q., Hang, R., Song, H., Li, Z., 2018a. Learning multiscale deep features for high-resolution satellite image scene classification. IEEE Trans. Geosci. Remote Sens. 56, 117–126.

Liu, X., Hu, G., Chen, Y., Li, X., Xu, X., Li, S., Pei, F., Wang, S., 2018b. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google earth engine platform. Remote Sens. Environ. 209, 227–239.

Liu, W., Wu, W., Thakuriah, P., Wang, J., 2020. The geography of human activity and land use: a big data approach. Cities 97.

Liu, X., Xu, Y., Engel, B.A., Sun, S., Zhao, X., Wu, P., Wang, Y., 2021a. The impact of urbanization and aging on food security in developing countries: the view from Northwest China. J. Clean. Prod. 292, 126067.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, pp. 9992–10002.

Liu, R., Zhang, H., Ling, J., 2022. Hybrid transformer networks for urban land use classification from optical and SAR images. In: IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, pp. 707–710.

Long, Y., Shen, Z., 2015. Discovering functional zones using bus smart card data and points of interest in Beijing. In: Geospatial Analysis to Support Urban Planning in Beijing. Springer International Publishing, pp. 193–217.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pp. 3431–3440.

Loveland, T.R., Reed, B.C., Brown, J.F., Ohlen, D.O., Zhu, Z., Yang, L., Merchant, J.W., 2000. Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. Int. J. Remote Sens. 21, 1303–1330.

Lowe, D.G., 2004. Distinctive image features from scale-invariant Keypoints. Int. J. Comput. Vis. 60, 91–110.

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.

Luus, F.P.S., Salmon, B.P., van den Bergh, F., Maharaj, B.T.J., 2015. Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. 12, 2448–2452.

Lv, N., Ma, H., Chen, C., Pei, Q., Zhou, Y., Xiao, F., Li, J., 2021. Remote sensing data augmentation through adversarial training. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 14, 9318–9333.

Lyu, H., Lu, H., Mou, L., 2016. Learning a transferable change rule from a recurrent neural network for land cover change detection. Remote Sens. 8, 506.

Ma, L., Ma, F., Ji, Z., Gu, Q., Wu, D., Deng, J., Ding, J., 2015a. Urban land use classification using LiDAR geometric, spatial autocorrelation and Lacunarity features combined with Postclassification processing method. Can. J. Remote. Sens. 41, 334–345.

Ma, Y., Wu, H., Wang, L., Huang, B., Ranjan, R., Zomaya, A., Jie, W., 2015b. Remote sensing big data computing: challenges and opportunities. Futur. Gener. Comput. Syst. 51, 47–60.

Ma, L., Fu, T., Li, M., 2018. Active learning for object-based image classification using predefined training objects. Int. J. Remote Sens. 39, 2746–2765.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177.

Ma, Y., Chen, S., Ermon, S., Lobell, D.B., 2024. Transfer learning in environmental remote sensing. Remote Sens. Environ. 301.

Man, Q., Dong, P., Guo, H., 2015. Pixel- and feature-level fusion of hyperspectral and lidar data for urban land-use classification. Int. J. Remote Sens. 36, 1618–1644.

Manandhar, R., Odeh, I., Ancev, T., 2009. Improving the accuracy of land use and land cover classification of Landsat data using post-classification enhancement. Remote Sens. 1, 330–344.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using ImageNet Pretrained networks. IEEE Geosci. Remote Sens. Lett. 13, 105–109.

McClellan, DeWitt, Hemmer, Matheson, Moe, 1989. Multispectral image-processing with a three-layer backpropagation network. In: International 1989 Joint Conference on Neural Networks, vol.151, pp. 151–153.

Mo, Y., Guo, Z., Zhong, R., Song, W., Cao, S., 2024. Urban functional zone classification using light-detection-and-ranging point clouds, aerial images, and point-of-interest data. Remote Sens. 16.

Montavon, G., Samek, W., Müller, K.-R., 2018. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K.P., 2013. A tutorial on synthetic aperture radar. IEEE Geosci. Remote Sens. Magaz. 1, 6–43.

Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 55, 3639–3655.

Myint, S.W., Gober, P., Brazel, A., Grossman-Clarke, S., Weng, Q., 2011. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. Remote Sens. Environ. 115, 1145–1161.

Ng, M.K., 2011. Strategic planning of China's first special economic zone: Shenzhen City master plan (2010–2020). Plan. Theory Pract. 12, 638–642.

Nijhawan, R., Joshi, D., Narang, N., Mittal, A., Mittal, A., 2019. A futuristic deep learning framework approach for land use-land cover classification using remote sensing imagery. In: Mandal, J.K., Bhattacharyya, D., Auluck, N. (Eds.), Advanced Computing and Communication Technologies. Springer Singapore, Singapore, pp. 87–96.

Nogueira, K., Penatti, O.A.B., dos Santos, J.A., 2017. Towards better exploiting convolutional neural networks for remote sensing scene classification. Pattern Recogn. 61, 539–556.

Oliva-Santos, R., Maciá-Pérez, F., Garea-Llano, E., 2014. Ontology-based topological representation of remote-sensing images. Int. J. Remote Sens. 35, 16–28.

Ouyang, S., Du, S., Zhang, X., Du, S., Bai, L., 2023. MDFF: a method for fine-grained UFZ mapping with multimodal geographic data and deep network. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 16, 9951–9966.

Paisitkriangkrai, S., Sherrah, J., Janney, P., van den Hengel, A., 2016. Semantic labeling of aerial and satellite imagery. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 9, 2868–2881.

Pan, G., Qi, G.D., Wu, Z.H., Zhang, D.Q., Li, S.J., 2013. Land-use classification using taxi GPS traces. IEEE Trans. Intell. Transp. Syst. 14, 113–123.

Pan, E., Mei, X., Wang, Q., Ma, Y., Ma, J., 2020a. Spectral-spatial classification for hyperspectral image based on a single GRU. Neurocomputing 387, 150–160.

Pan, S., Guan, H., Chen, Y., Yu, Y., Nunes Gonçalves, W., Marcato Junior, J., Li, J., 2020b. Land-cover classification of multispectral LiDAR data using CNN with optimized hyper-parameters. ISPRS J. Photogramm. Remote Sens. 166, 241–254.

Pan, Y., Zeng, W., Guan, Q., Yao, Y., Liang, X., Yue, H., Zhai, Y., Wang, J., 2020c. Spatiotemporal dynamics and the contributing factors of residential vacancy at a fine scale: a perspective from municipal water consumption. Cities 103, 102745.

Paola, J.D., Schowengerdt, R.A., 1995. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. IEEE Trans. Geosci. Remote Sens. 33, 981–996.

Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A., 2019. Deep learning classifiers for hyperspectral imaging: a review. ISPRS J. Photogramm. Remote Sens. 158, 279–317.

Pathan, S.K., Jothimahi, P., Kumar, D.S., Pendharkar, S.P., 1989. Urban land use mapping and zoning of Bombay metropolitan region using remote sensing data. J. Indian Soc. Remote Sens. 17, 11–22.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., Zhou, C., 2014. A new insight into land use classification based on aggregated mobile phone data. Int. J. Geogr. Inf. Sci. 28, 1988–2007.

Penatti, O.A.B., Nogueira, K., Santos, J.A.D., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 44–51.

Qiao, Z., Yuan, X., 2021. Urban land-use analysis using proximate sensing imagery: a survey. Int. J. Geogr. Inf. Sci. 35, 2129–2148.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204.

Ríos, S.A., Muñoz, R., 2017. Land use detection with cell phone data using topic models: case Santiago, Chile. Comput. Environ. Urban. Syst. 61, 39–48.

Rozenstein, O., Karnieli, A., 2011. Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. Appl. Geogr. 31, 533–544.

Rußwurm, M., Körner, M., 2017. Temporal vegetation modelling using Long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1496–1504.

Sak, H., Senior, A., Beaufays, F., 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. In (p. arXiv:1402.1128).

Sanlang, S., Cao, S., Du, M., Mo, Y., Chen, Q., He, W., 2021. Integrating aerial LiDAR and very-high-resolution images for urban functional zone mapping. Remote Sens. 13.

Scheibenreif, L., Hanna, J., Mommert, M., Borth, D., 2022. Self-supervised vision transformers for land-cover segmentation and classification. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1421–1430.

Schilling, K.E., Jha, M.K., Zhang, Y.-K., Gassman, P.W., Wolter, C.F., 2008. Impact of land use and land cover change on the water balance of a large agricultural watershed: historical effects and future directions. Water Resour. Res. 44, W00A09.

Schneider, A., Friedl, M.A., Potere, D., 2010. Mapping global urban areas using MODIS 500-m data: new methods and datasets based on 'urban ecoregions'. Remote Sens. Environ. 114, 1733–1746.

Schulz, D., Yin, H., Tischbein, B., Verleysdonk, S., Adamou, R., Kumar, N., 2021. Land use mapping using Sentinel-1 and Sentinel-2 time series in a heterogeneous landscape in Niger, Sahel. ISPRS J. Photogramm. Remote Sens. 178, 97–111.

Senaratne, H., Mobasheri, A., Ali, A.L., Capineri, C., Haklay, M., 2017. A review of volunteered geographic information quality assessment methods. Int. J. Geogr. Inf. Sci. 31, 139–167.

Shang, C., Li, X., Foody, G.M., Du, Y., Ling, F., 2022. Superresolution land cover mapping using a generative adversarial network. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Sharma, A., Liu, X., Yang, X., 2018. Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks. Neural Netw. 105, 346–355.

Shen, Y., Karimi, K., 2016. Urban function connectivity: characterisation of functional urban streets with social media check-in data. Cities 55, 9–21.

Simkin, R.D., Seto, K.C., McDonald, R.I., Jetz, W., 2022. Biodiversity impacts and conservation implications of urban land expansion projected to 2050. Proc. Natl. Acad. Sci. 119, e2117297119.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015), pp. 1–14.

Solberg, A.H.S., Jain, A.K., Taxt, T., 1994. Multisource classification of remotely sensed data: fusion of Landsat TM and SAR images. IEEE Trans. Geosci. Remote Sens. 32, 768–778.

Soliman, A., Soltani, K., Yin, J., Padmanabhan, A., Wang, S., 2017. Social sensing of urban land use based on analysis of twitter users' mobility patterns. PLoS One 12, e0181657.

Song, P., Li, J., An, Z., Fan, H., Fan, L., 2023. CTMFNet: CNN and transformer multiscale fusion network of remote sensing urban scene imagery. IEEE Trans. Geosci. Remote Sens. 61, 1–14.

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.

Steiger, E., de Albuquerque, J.P., Zipf, A., 2015. An advanced systematic literature review on spatiotemporal analyses of twitter data. Trans. GIS 19, 809–834.

Su, M., Guo, R., Chen, B., Hong, W., Wang, J., Feng, Y., Xu, B., 2020. Sampling strategy for detailed urban land use classification: a systematic analysis in Shenzhen. Remote Sens. 12, 1497.

Su, C., Hu, X., Meng, Q., Zhang, L., Shi, W., Zhao, M., 2024. A multimodal fusion framework for urban scene understanding and functional identification using geospatial data. Int. J. Appl. Earth Obs. Geoinf. 127.

Sun, J., Wang, H., Song, Z., Lu, J., Meng, P., Qin, S., 2020. Mapping essential urban land use categories in Nanjing by integrating multi-source big data. Remote Sens. 12.

Sun, Z., Jiao, H., Wu, H., Peng, Z., Liu, L., 2021. Block2vec: An approach for identifying urban functional regions by integrating sentence embedding model and points of interest. ISPRS Int. J. Geo Inf. 10.

Sun, Z., Peng, Z., Yu, Y., Jiao, H., 2022. Deep convolutional autoencoder for urban land use classification using mobile device data. Int. J. Geogr. Inf. Sci. 36, 2138–2168.

Sun, X., Wang, P., Lu, W., Zhu, Z., Lu, X., He, Q., Li, J., Rong, X., Yang, Z., Chang, H., He, Q., Yang, G., Wang, R., Lu, J., Fu, K., 2023. RingMo: a remote sensing foundation model with masked image modeling. IEEE Trans. Geosci. Remote Sens. 61, 1–22.

Szabo, S., 2016. Urbanisation and food insecurity risks: assessing the role of human development. Oxf. Dev. Stud. 44, 28–48.

Tang, X., Woodcock, C.E., Olofsson, P., Hutyra, L.R., 2021. Spatiotemporal assessment of land use/land cover change and associated carbon emissions and uptake in the Mekong River basin. Remote Sens. Environ. 256, 112336.

Theobald, D.M., 2014. Development and applications of a comprehensive land use classification and map for the US. PLoS One 9, e94628.

Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322.

Toole, J.L., Ulm, M., González, M.C., Bauer, D., 2012. Inferring land use from mobile phone activity. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. Association for Computing Machinery, Beijing, China, pp. 1–8.

Tu, W., Cao, J., Yue, Y., Shaw, S.-L., Zhou, M., Wang, Z., Chang, X., Xu, Y., Li, Q., 2017. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. Int. J. Geogr. Inf. Sci. 31, 2331–2358.

Tu, Y., Chen, B., Zhang, T., Xu, B., 2020. Regional mapping of essential urban land use categories in China: a segmentation-based approach. Remote Sens. 12, 1058.

United Nations Department of Economic Social Affairs, 2019. World Urbanization Prospects: The 2018 Revision. United Nations.

Vaglio Laurin, G., Liesenberg, V., Chen, Q., Guerriero, L., Del Frate, F., Bartolini, A., Coomes, D., Wilebore, B., Lindsell, J., Valentini, R., 2013. Optical and SAR sensor synergies for forest and land cover mapping in a tropical site in West Africa. Int. J. Appl. Earth Obs. Geoinf. 21, 7–16.

van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109, 373–440.

Vargas-Munoz, J.E., Srivastava, S., Tuia, D., Falcão, A.X., 2021. OpenStreetMap: challenges and opportunities in machine learning and remote sensing. IEEE Geosci. Remote Sens. Magaz. 9, 184–199.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Long Beach, California, USA, pp. 6000–6010.

Verpoorter, C., Kutser, T., Tranvik, L., 2012. Automated mapping of water bodies using Landsat multispectral data. Limnol. Oceanogr. Methods 10, 1037–1050.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55, 881–893.

Wang, J., Li, X., Zhou, S., Tang, J., 2017. Landcover Classification Using Deep Fully Convolutional Neural Networks. In (pp. IN11E-02).

Wang, D., Szymanski, B.K., Abdelzaher, T., Ji, H., Kaplan, L., 2019. The age of social sensing. Computer 52, 36–45.

Wang, Z.J., Liu, H.X., Zhu, Y.D., Zhang, Y.R., Basiri, A., Buttner, B., Gao, X., Cao, M.Q., 2021. Identifying urban functional areas and their dynamic changes in Beijing: using multiyear transit smart card data. J. Urban Plan. Dev. 147, 04021002.

Wang, J., Bretz, M., Dewan, M.A.A., Delavar, M.A., 2022. Machine learning in modelling land-use and land cover-change (LULCC): current status, challenges and prospects. Sci. Total Environ. 822, 153559.

Wang, J., Feng, C.-C., Guo, Z., 2023. A novel graph-based framework for classifying urban functional zones with multisource data and human mobility patterns. Remote Sens. 15.

Wang, S., Hu, T., Xiao, H., Li, Y., Zhang, C., Ning, H., Zhu, R., Li, Z., Ye, X., 2024. GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review. Int. J. Digital Earth 17.

Whiteside, T.G., Boggs, G.S., Maier, S.W., 2011. Comparing object-based and pixel-based classifications for mapping savannas. Int. J. Appl. Earth Obs. Geoinf. 13, 884–893.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C., 2015. Discovering urban activity patterns in cell phone data. Transportation 42, 597–623.

Wilkinson, G.G., 2005. Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Trans. Geosci. Remote Sens. 43, 433–440.

Williams, R.J., Zipser, D., 1989. A learning algorithm for continually running fully recurrent neural networks. Neural Comput. 1, 270–280.

Wu, C., Murray, A.T., 2003. Estimating impervious surface distribution by spectral mixture analysis. Remote Sens. Environ. 84, 493–505.

Wu, X., Liu, X., Zhang, D., Zhang, J., He, J., Xu, X., 2022. Simulating mixed land-use change under multi-label concept by integrating a convolutional neural network and cellular automata: a case study of Huizhou, China. GIScience Remote Sens. 59, 609–632.

Wu, M., Huang, Q., Gao, S., Zhang, Z., 2023. Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning. Int. J. Appl. Earth Obs. Geoinf. 125.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: a benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens. 55, 3965–3981.

Xia, B., Kong, F., Zhou, J., Wu, X., Xie, Q., 2022. Land resource use classification using deep learning in ecological remote sensing images. Comput. Intell. Neurosci. 2022, 7179477.

Xiao, B., Liu, J., Jiao, J., Li, Y., Liu, X., Zhu, W., 2022. Modeling dynamic land use changes in the eastern portion of the hexi corridor, China by cnn-gru hybrid model. GIScience Remote Sens. 59, 501–519.

Xiao, F., Zhou, Y., Huang, Y., 2023. Old wine in a new bottle: understanding the expansion of the Shenzhen special economic zone in China. J. Urban Plan. Dev. 149.

Xie, Y., Sha, Z., Yu, M., 2008. Remote sensing imagery in vegetation mapping: a review. J. Plant Ecol. 1, 9–23.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Proces. Syst. 34, 12077–12090.

Xie, X., Xu, Y., Feng, B., Wu, W., 2024. Multiscale urban functional zone recognition based on landmark semantic constraints. ISPRS Int. J. Geo Inf. 13.

Xu, H., 2008. A new index for delineating built-up land features in satellite imagery. Int. J. Remote Sens. 29, 4269–4276.

Xu, B., Gong, P., 2007. Land-use/land-cover classification with multispectral and hyperspectral EO-1 data. Photogramm. Eng. Remote Sens. 73, 955–965.

Xu, S., Mu, X., Chai, D., Zhang, X., 2018. Remote sensing image scene classification based on generative adversarial networks. Remote Sens. Lett. 9, 617–626.

Xu, C., Du, X., Fan, X., Giuliani, G., Hu, Z., Wang, W., Liu, J., Wang, T., Yan, Z., Zhu, J., Jiang, T., Guo, H., 2022a. Cloud-based storage and computing for remote sensing big data: a technical review. Int. J. Digital Earth 15, 1417–1445.

Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., He, N., 2022b. A Framework for Urban Land Use Classification by Integrating the Spatial Context of Points of Interest and Graph Convolutional Neural Network Method. Computers, Environment and Urban Systems, p. 95.

Yan, W.Y., Shaker, A., El-Ashmawy, N., 2015. Urban land cover classification using airborne LiDAR data: a review. Remote Sens. Environ. 158, 295–310.

Yan, B., Janowicz, K., Mai, G., Gao, S., 2017. From ITDL to Place2Vec. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1–10.

Yan, X., Jiang, Z., Luo, P., Wu, H., Dong, A., Mao, F., Wang, Z., Liu, H., Yao, Y., 2024. A multimodal data fusion model for accurate and interpretable urban land use mapping with uncertainty analysis. Int. J. Appl. Earth Obs. Geoinf. 129.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. Association for Computing Machinery, San Jose, California, pp. 270–279.

Yang, Z., Li, W., Chen, Q., Wu, S., Liu, S., Gong, J., 2018. A scalable cyberinfrastructure and cloud computing platform for forest aboveground biomass estimation based on the Google earth engine. Int. J. Digital Earth 12, 995–1012.

Yang, C., Rottensteiner, F., Heipke, C., 2021. A hierarchical deep learning framework for the consistent classification of land use objects in geospatial databases. ISPRS J. Photogramm. Remote Sens. 177, 38–56.

Yang, M., Kong, B., Dang, R., Yan, X., 2022. Classifying urban functional regions by integrating buildings and points-of-interest using a stacking ensemble method. Int. J. Appl. Earth Obs. Geoinf. 108.

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., Mai, K., 2016. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. Int. J. Geogr. Inf. Sci. 31, 825–848.

Yao, Y., Yan, X., Luo, P., Liang, Y., Ren, S., Hu, Y., Han, J., Guan, Q., 2022. Classifying land-use patterns by integrating time-series electricity data and high-spatial resolution remote sensing imagery. Int. J. Appl. Earth Obs. Geoinf. 106, 102664.

Yokoya, N., Grohnfeldt, C., Chanussot, J., 2017. Hyperspectral and multispectral data fusion: a comparative review of the recent literature. IEEE Geosci. Remote Sens. Magaz. 5, 29–56.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., Schirokauer, D., 2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. Photogramm. Eng. Remote. Sens. 72, 799–811.

Yu, M., Xu, H., Zhou, F., Xu, S., Yin, H., 2023. A deep-learning-based multimodal data fusion framework for urban region function recognition. ISPRS Int. J. Geo Inf. 12.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241.

Yuan, J., Wang, S., Wu, C., Xu, Y., 2022. Fine-grained classification of urban functional zones and landscape pattern analysis using hyperspectral satellite imagery: a case study of Wuhan. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 15, 3972–3991.

Yue, J., Fang, L., Ghamisi, P., Xie, W., Li, J., Chanussot, J., Plaza, A., 2022. Optical remote sensing image understanding with weak supervision: concepts, methods, and perspectives. IEEE Geosci. Remote Sens. Magaz. 10, 250–269.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In (p. arXiv: 1605.07146).

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., Arino, O., 2022. ESA WorldCover 10 m 2021 v200. Zenodo.

Zang, N., Cao, Y., Wang, Y., Huang, B., Zhang, L., Mathiopoulos, P.T., 2021. Land-use mapping for high-spatial resolution remote sensing image via deep learning: a review. IEEE J. Select. Top. Appl. Earth Observ. Remote Sens. 14, 5372–5391.

Zhang, C., Zhao, T., Li, W., 2015. Geospatial Semantic Web. Springer, Cham.

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, W., Villarini, G., Vecchi, G.A., Smith, J.A., 2018b. Urbanization exacerbated the rainfall and flooding caused by hurricane Harvey in Houston. Nature 563, 384–388.

Zhang, X., Li, W., Zhang, F., Liu, R., Du, Z., 2018c. Identifying urban functional zones using public bicycle rental records and point-of-interest data. ISPRS Int. J. Geo Inf. 7.

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., Harrison, P.A., Pan, X., Li, H., Sargent, I., Atkinson, P.M., 2020a. Scale sequence joint deep learning (SS-JDL) for land use and land cover classification. Remote Sens. Environ. 237, 111593.

Zhang, J., Li, X., Yao, Y., Hong, Y., He, J., Jiang, Z., Sun, J., 2020b. The Traj2Vec model to quantify residents' spatial trajectories and estimate the proportions of urban land-use types. Int. J. Geogr. Inf. Sci. 35, 193–211.

Zhang, Y., Chen, G., Myint, S.W., Zhou, Y., Hay, G.J., Vukomanovic, J., Meentemeyer, R. K., 2022. UrbanWatch: a 1-meter resolution land cover and land use database for 22 major cities in the United States. Remote Sens. Environ. 278, 113106.

Zhang, Y., Liu, P., Biljecki, F., 2023. Knowledge and topology: a two layer spatially dependent graph neural networks to identify urban functions with time-series street view image. ISPRS J. Photogramm. Remote Sens. 198, 153–168.

Zhao, W., Du, S., 2016. Scene classification using multi-scale deeply described visual words. Int. J. Remote Sens. 37, 4119–4131.

Zhao, B., Huang, B., Zhong, Y., 2017a. Transfer learning with fully Pretrained deep convolution networks for land-use classification. IEEE Geosci. Remote Sens. Lett. 14, 1436–1440.

Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017b. Contextually guided very-high-resolution imagery classification with semantic segments. ISPRS J. Photogramm. Remote Sens. 132, 48–60.

Zhao, K., Liu, Y., Hao, S., Lu, S., Liu, H., Zhou, L., 2022a. Bounding boxes are all we need: street view image classification via context encoding of detected buildings. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Zhao, W., Li, M., Wu, C., Zhou, W., Chu, G., 2022b. Identifying urban functional regions from high-resolution satellite images using a context-aware segmentation network. Remote Sens. 14.

Zheng, Q., Weng, Q., Huang, L., Wang, K., Deng, J., Jiang, R., Ye, Z., Gan, M., 2018. A new source of multi-spectral high spatial resolution night-time light imagery—JL1-3B. Remote Sens. Environ. 215, 300–312.

Zheng, Y., Zhang, X., Ou, J., Liu, X., 2024. Identifying building function using multisource data: a case study of China's three major urban agglomerations. Sustain. Cities Soc. 108.

Zhong, Y., Cao, Q., Zhao, J., Ma, A., Zhao, B., Zhang, L., 2017. Optimal decision fusion for urban land-use/land-cover classification based on adaptive differential evolution using hyperspectral and LiDAR data. Remote Sens. 9, 868.

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, 111838.

Zhong, Y., Yan, B., Yi, J., Yang, R., Xu, M., Su, Y., Zheng, Z., Zhang, L., 2023. Global urban high-resolution land-use mapping: from benchmarks to multi-megacity applications. Remote Sens. Environ. 298.

Zhou, Z.-H., 2018. A brief introduction to weakly supervised learning. Natl. Sci. Rev. 5, 44–53.

Zhou, X., Zhang, L., 2016. Crowdsourcing functions of the living city from twitter and foursquare data. Cartogr. Geogr. Inf. Sci. 43, 393–404.

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.

Zhou, W., Persello, C., Li, M., Stein, A., 2023. Building use and mixed-use classification with a transformer-based network fusing satellite images and geospatial textual information. Remote Sens. Environ. 297.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Magaz. 5, 8–36.

Zhu, Q., Lei, Y., Sun, X., Guan, Q., Zhong, Y., Zhang, L., Li, D., 2022. Knowledge-guided land pattern depiction for urban land use mapping: a case study of Chinese cities. Remote Sens. Environ. 272.