npj | urban sustainability

Perspective

Published in partnership with RMIT University

https://doi.org/10.1038/s42949-024-00188-3

How will ai transform urban observing, sensing, imaging, and mapping?

Check for updates

Qihao Weng [1,2] ✉, Zhiwei Li [1,2], Yinxia Cao [1,2], Xiaoyan Lu[1,2], Paolo Gamba [3], Xiaoxiang Zhu [4], Yonghao Xu [5], Fan Zhang[6], Rongjun Qin[7], Micheal. Y. Yang[8], Peifeng Ma[9], Wei Huang[10], Tiangang Yin[1,2], Qiming Zheng[1,2,11], Yuhan Zhou[1,2] & Greg Asner [12,13] ✉

Advances in artificial intelligence (AI) and Earth observation (EO) have transformed urban studies. This paper provides a commentary on how the AI-EO integration offers advancements in urban studies and applications. We conclude that AI will provide a deeper interpretation and autonomous identification of urban issues and the creation of customized urban designs. Open issues remain, especially in integrating diverse geospatial big data, data security, and developing a general analytical framework.

The need for monitoring and managing urban areas is amplified by the concern over global climate change. Cities are among the most complex of human settlements, and urban areas may be more vulnerable than rural settlements to the impacts of global climate change[1]. Most concerns, including health, water and infrastructure, severe weather events, energy requirements, urban metabolism, sea level rise, economic competitiveness, opportunities and risks, social and political structures, and the United Nation's Sustainable Development Goals (SDGs) can be better understood with Earth observation (EO) technology. In fact, EO techniques, in conjunction with in situ data collection, have been used to observe, monitor, measure, and model many of the components that comprise natural and human ecosystem cycles for decades[2].

Since the beginning of the 21st century, we have witnessed a great increase in EO-related research and development, technology transfer, and engineering activities worldwide. Commercial satellites acquire imagery at spatial resolutions previously only possible to aerial, with additional advantages for producing stereo image pairs conveniently for three-dimensional (3D) mapping[3]. Hyperspectral imaging affords the potential for detailed identification of materials and better estimates of their abundance on the Earth's surface, while light detection and ranging (LiDAR) technology provides high-accuracy height and other geometric information for urban structures and vegetation. Radar technology has been re-invented since the 1990s due greatly to the increase of spaceborne radar programs[3] and its images emphasize humidity, relief, and morphological structure of the observed terrain. Recently, nighttime light imagery and street-level imagery has emerged as additional important data sources in urban areas, particularly from a human perspective[4]. Moreover, many government agencies and companies have collected GIS (geographic information system) data sets along with remote sensing imagery for civic and environmental applications, such as Google Earth and Virtual Globe. These virtual "worlds", in conjunction with GPS, social media, and modern telecommunication technologies, have sparked much interest in the public for urban observing, sensing, imaging, and mapping. However, a great deal has yet to be learnt about the integrated use of these systems in understanding urban issues[5]. At the meantime, artificial intelligence (AI) is rapidly changing the field of remote sensing and mapping[6] and enables research and applications on previously inconceivable topics and at unprecedented scales. AI techniques, such as deep learning, have been proven to be both a science breakthrough and a powerful technical toolbox in many fields. Early success in EO digital image processing has been demonstrated via image pre-processing, classification, target recognition, and 3D reconstruction[7], but it remains a challenge to expand AI application in EO due largely to the complexity of urban landscapes and the existence of mixed pixels in urban areas[8].

The history of EO technology has revealed that three stages can be discerned in image processing, analysis, understanding, and pattern recognition. The first stage, from the 1970s to the beginning of the 21st century, focused on addressing the overall question of "What Is

[1]JC STEM Lab of Earth Observations, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China. [2]Research Centre for Artificial Intelligence in Geomatics, The Hong Kong Polytechnic University, Hung Hom, Hong Kong, China. [3]Department of Electrical, Biomedical and Computer Engineering, University of Pavia, Pavia, Italy. [4]Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany. [5]Department of Electrical Engineering, Linköping University, Linköping, Sweden. [6]Institute of Remote Sensing and Geographical Information System, School of Earth and Space Sciences, Peking University, Beijing, China. [7]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA. [8]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands. [9]Department of Geography and Resource Management, Institute of Space and Earth Information Science, Shenzhen Research Institute, The Chinese University of Hong Kong, Shatin, Hong Kong, China. [10]College of Surveying and Geo-informatics, Tongji University, Shanghai, 200092, China. [11]Department of Geography and Resource Management, The Chinese Univeristy of Hong Kong, Hong Kong, China. [12]Center for Global Discovery and Conservation Science, Arizona State University, Tempe, AZ, 85287, USA. [13]School of Ocean Futures, Arizona State University, Tempe, AZ, 85287, USA. ✉e-mail: qihao.weng@polyu.edu.hk; gregasner@asu.edu

Within a Pixel?"[9]. This stage was characterized by pixel-based and sub-pixel analysis and utilizes essentially the tone and color of pixels. The second stage focused on "larger than one pixel" representations and spanned over the first fifteen years of this century. This stage was characterized by object-based image analysis, emphasizing the importance of image texture in image analysis and pattern recognition. The third stage, from circa 2015 to the present, focused on human recognition, and was characterized by the ubiquitous use of AI techniques to mimic how human beings extract information at multiple spatial scales from an image. The human brain is organized in a deep architecture, and its perception and recognition are manifested at multiple levels of abstraction with non-linearity and feedback at different stages. Emerging trends in AI tend to respond to the question of how human beings perceive, recognize, and understand the world, instead of a machine view of the world through "data grids". AI uses EO data grids as building blocks to make sense, to facilitate, or to revolutionize, our understanding of the world.

In this paper, we will provide a review and synthesis on how AI reshapes the research paradigm of EO. In addition to assessing progress and problems in the frontiers of AI in urban studies, we are especially interested in new research directions, emerging trends, and advances across multiple sub-fields and beyond. We look at how EO and AI technologies integrate to offer the profound potential for advancements in every aspect of urban studies, including observation, sensing, imaging, mapping, and interpretation of urban challenges (Fig. 1).

AI in earth observations: progress and problems

Theoretical basis of AI in urban systems

AI can assist in addressing many issues in urban systems by detailed and extensive sensing of urban environments[10]. Deep learning[11] is a branch of machine learning that utilizes deep neural networks to learn and represent complex patterns in data, which can be employed for tasks such as fine object recognition. Natural language processing[12] focuses on how computers understand and process human language, which can be applied to analyze and extract insights from urban-related textual data, such as social media data. Reinforcement learning[13] is a learning paradigm that aims to train intelligent agents by interacting with the environment to learn optimal action strategies, which can be utilized to optimize decision-making in such areas as urban transportation systems and energy management. These theoretical bases enable the use of AI technology to analyze and address problems in urban research, providing deeper insights and better knowledge to support decision-making.

The power of AI in urban research lies in its ability to process multiple types of data, analyze complex patterns, and make informed predictions, which is crucial for understanding complex urban systems. The application of AI methods and techniques usually considers many factors, including research tasks (e.g., image classification, object detection, etc.), the modality of data (e.g., optical, radar images, etc.), the hardware (e.g., graphics processing unit) and platform (e.g., local, distributed, or cloud computing)[14], the selection of models, the construction of networks, and the validation of results. The joint use of multimodal data should be carefully considered in the construction of networks. The criteria for model selection depend on the specific task, data, and the desired output. The construction of networks is not yet unified and explainable. Therefore, a general framework and guidelines for selecting models, constructing networks, and validating results are needed to fully leverage the potential of AI in urban studies.

Digital image processing

In digital image processing, Convolutional Neural Network (CNN) is widely used due to its powerful ability of local feature extraction[15]. For sequential or time-series data, Recurrent Neural Network (RNN) is popular[16]. Recently, a new AI model, transformer, has achieved great advances in natural language processing, and has been successfully transferred into the image processing field. Compared to CNN and RNN, the transformer entirely consists of attention mechanisms and can model long-range dependency between input and output at much lower training cost[17]. Fueled by the rapid development of hardware, big data, and AI techniques, foundation models based on transformers have been successfully proposed for general purposes and can be readily applied to various downstream tasks[18]. In the field of remote sensing, foundation models have increasingly received wide concerns, e.g., Prithvi[19] and RemoteCLIP[20]. These foundation models point out a promising direction for dealing with multi-modal data and general tasks. This can be outlined in a general framework with three components: model, input, and output (Fig. 2).

Model. AI models often operate as a "black box"[21], neglecting the underlying physical mechanisms. To address this issue, we propose a general AI framework, which consists of three parts: the encoder for feature extraction; the feature fusion module for the fusion of diverse features; and the decoder for the reconstruction of output features. The key innovation lies in the utilization of prior knowledge and the integration of major cutting-edge AI models, such as CNN, transformer, RNN, graph neural network (GNN), and generative adversarial network (GAN). Different AI models can serve as encoders or decoders based on

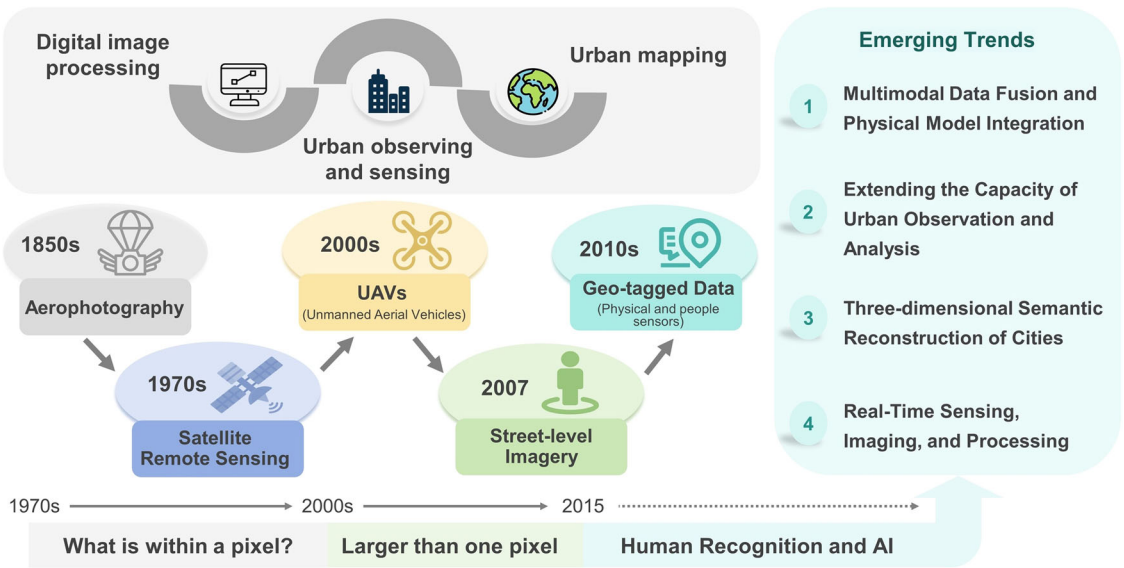

Fig. 1 | Topics covered in this paper and the logic for their arrangement (image created by the authors).

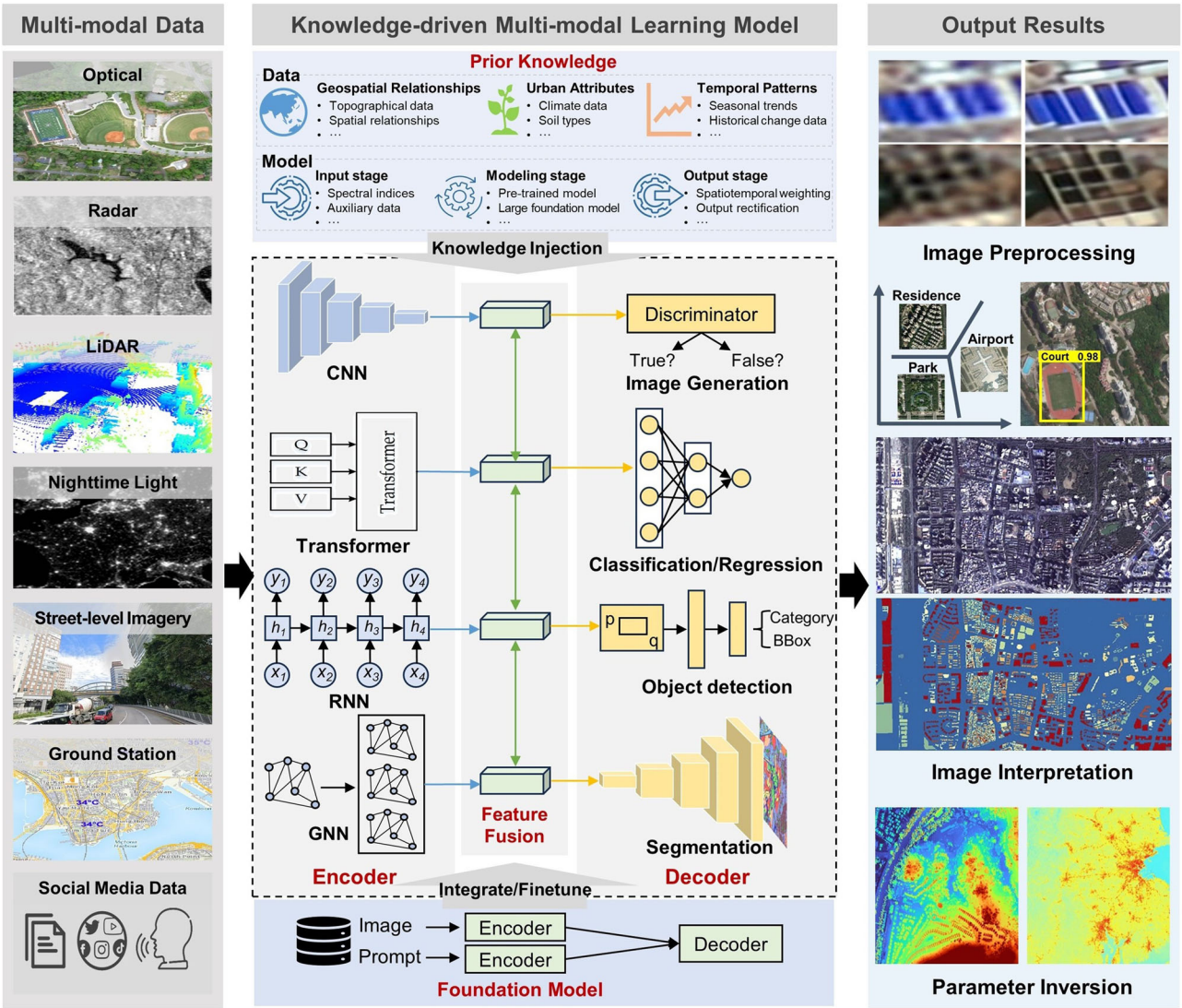

Fig. 2 | The general AI framework in Earth observations (image created by the authors).

their strengths in feature representation. Prior knowledge can be integrated at different stages. At the input stage, it can enrich and integrate prominent features, reducing redundancy, such as remotely sensed spectral indices. During the modeling, prior knowledge can be accounted for in network weights through model pre-training or fine-tuning. At the output stage, it can guide the learning process and provide more reliable outputs, e.g., by the addition of spatial-temporal weighted terms. This knowledge-driven approach enhances the model interpretability and generalization and compensates for limited training data.

Input data. These EO data are characterized by diverse spectral, spatial, and temporal resolutions and broad spatial coverage, enabling long-term urban monitoring. Within the AI framework, prior knowledge complements raw data, especially when the available input data is limited. The type of prior knowledge to be incorporated depends mainly on research objectives, geospatial relationships, urban attributes, and temporal patterns (see Fig. 2).

Output data. The output is application-specific, ranging from image preprocessing and interpretation to parameter estimation. Utilizing an appropriate model informed by practical urban knowledge yields more accurate and comprehensive insights, contributing to more effective urban sensing and imaging.

Urban mapping

AI can handle different types of data, including text, audio, image, and video, and can integrate them to produce more accurate results than traditional methods. It enhances data interpretation capabilities and helps make informed decisions in various fields. It has revolutionized the field of urban mapping by processing and analyzing various types of data. In this section, we will discuss three applications of AI in urban mapping: land use and land cover (LULC) mapping, building detection, and road extraction.

LULC mapping has long been a hot topic and is evolving with deep learning[22]. The exceptional performance of deep learning in LULC mapping is due to several factors. First, deep learning eliminates the need for manual feature engineering due to the inherent ability of the models to learn directly from data. Second, deep learning enhances the ease of incorporating heterogeneous multi-modal data into the mapping process. Third, deep learning can generate diverse output types, such as point-level categories, segmented objects, and bounding boxes[23]. Nevertheless, deep learning is data-driven and relies heavily on labeled data. In addition, although diverse LULC products have been developed for local or global regions, there exist considerable uncertainties and inconsistencies. Urban green spaces (UGS), as a special type of land cover, play an important role in understanding urban ecosystems, climate, environment, public health concerns, and the SDGs at various spatial scales. Mapping of UGS with remote sensing is challenging due to the existence of mixed pixels and the cost and time of

collecting quality training data. CNN and other deep learning methods have been employed for UGS mapping and found them effective[24].

Building detection is one of the most profoundly advanced areas of EO-based deep learning. Historically, building feature-based methods have been developed to advance automated building detection[25], but they rely on domain-specific knowledge to manually design building- related features to be detected and mapped. Deep learning, trained using existing open-source databases obtained by citizen science, have become a mainstream for building detection[26]. For instance, Microsoft has released a global building footprints dataset generated by deep learning networks, which was almost impossible to achieve in the past, yet the completeness of this dataset still needs attention.

Similarly, AI has made it possible for automatic extraction of roads. For example, the foundation model has been utilized to extract road networks by employing autoencoders and contrastive learning for self-supervised training on large-scale unlabeled remote sensing images[27]. Parameter-efficient fine-tuning methods were used to apply these general foundation models for road extraction tasks. Because self-supervised training learns the distribution of vast amounts of data, the model's feature representation capabilities are significantly enhanced, thereby improving the performance of road extraction. Cross-modal learning has also been applied to road extraction tasks[28]. For instance, GPS data is used to address the issue of insufficient road data labels to some extent. AI methods are still constrained in road detection and mapping in several aspects. First, there is a lack of an accurate and diverse training dataset for global-scale road mapping[29]. Second, the generalization ability of AI models remains limited for global applications. Third, the lack of inductive reasoning ability for AI models leads to disconnected roads, which may lead to inaccurate conclusions in road network-based urban studies. AI methods focus mainly on recognizing individual pixels as roads, rather than inferring road connectivity according to the cognitive process applied by human beings[30].

Urban observing and sensing

Following the discussion on the three widest applications in urban mapping, where optical remote sensing methods are primarily utilized, this chapter focuses the discussion on urban observation and sensing with other sensing systems and platforms, such as LiDAR, Synthetic Aperture Radar (SAR), street-level imagery, as well as people as virtual sensors.

LiDAR technology offers exceptional 3D data acquisition capabilities for urban landscapes, structures and infrastructure, as well as monitoring changes over time. Small-footprint airborne LiDAR delivers high-resolution topographic data, excelling at generating detailed 3D urban environment models[31]. Integrating AI with LiDAR data processing enables sophisticated classification and analysis of urban features. For instance, AI models trained on CNNs have improved interpreting and merging information from these diverse sensor modalities, thereby enhancing point semantic labeling and classification accuracy[32]. Nonetheless, the fusion of LiDAR with other sensors to improve information retrieval with the existence of occlusion from LiDAR viewing geometry poses significant challenges for urban applications. To address these challenges, cross-modal learning strategies leverage LiDAR data combined with visual and thermal imagery to compensate for areas where LiDAR data is incomplete or obstructed, thereby enriching the dataset[33]. In addition, self-supervised learning models have been utilized, autonomously predicting missing or noisy data sections based on patterns identified in complete and clean sections[34]. This approach enhances data quality and facilitates learning from the intrinsic structure of LiDAR data without relying on manually labeled examples, which is particularly beneficial for managing large datasets and standardizing data quality across different systems.

SAR, featuring all-weather capability, rapid revisit, and multi-angle observations, is an important EO technology. Increasing accessibility of SAR has significantly enabled the application of AI for urban sensing and mapping[35]. Additionally, interferometric SAR (InSAR) techniques are used to process and analyze multitemporal SAR, enabling accurate measurements of urban surface and infrastructure deformation. Compared to optical images, SAR exhibits distinct characteristics, including speckle noise, multipath scattering, and geometrical distortions, which negatively impact their interpretation. These issues also pose challenges for AI-based analysis of SAR images in conjunction with optical ones.

The potential of street-level imagery has been advanced with AI for data mining and knowledge discovery[10]. For example, it is possible to evaluate the conditions of urban infrastructure through semantic segmentation methodologies[36]. Deeper insights, including safety, architectural age and style, and the urban socio-economic environment are also available through AI[37]. Despite these advancements, challenges remain, such as in the integration of street-level with satellite/airborne data. Satellite/airborne sensing provides a large-scale perspective but is limited to top-down or oblique observations, while street-level imagery offers a ground-based observation from a human's perspective.

The aforementioned EO technologies have traditionally been applied to study static objects such as LULC. Recently, massive geo-tagged data on dynamic objects (e.g., human behaviors) have been generated by physical and people sensors (people as virtual sensors). These data, such as GPS trajectories, surveillance data, urban environment data (e.g., temperature and air quality data) and human-generated data, are mostly associated with geo-locations, capturing urban dynamics (e.g., human movements, urban events and processes) from different angles. They provide multi-dimensional EO data in a granular manner, which have greatly catalyzed the application of AI techniques to urban sensing[38]. For instance, AI has been widely applied to GPS trajectories and urban environment data, which has significantly improved human movement prediction[39] and urban environment change forecasting[40]. Nevertheless, challenges such as fusing the geo-tagged data with the EO data for effective AI modeling remained due to their spatiotemporal scale differences and the qualities of measurement.

Human-generated data can be categorized into passive sensing (e.g., data generated from social media, location-based services, and mobile devices) and active sensing (e.g., Public Participation Geographic Information Systems (PPGIS), Volunteered Geographic Information Systems (VGIS), and surveys). These data sources offer diverse insights into human activities and behaviors, as well as other human factors related to social sustainability, such as environmental experiences, perceptions, and needs. Social media data, for instance, can be used to analyze social phenomena such as segregation, while PPGIS data can help identify community needs and preferences regarding urban planning.

Emerging trends

To explore research directions and emerging trends, new and directed research questions should be considered. Here, we will focus on two of them: How will AI transform urban observing, sensing, imaging, and mapping? How can urban landscapes, phenomena, and events be better perceived and recognized with AI models using EO and geospatial big data?

Multimodal data fusion and physical model integration

There is a growing interest in the remote sensing community for multi-modal data, acquired from a variety of platforms, including satellites, aircraft, unmanned aerial vehicles, autonomous vehicles, ground sensor networks, social media, and by different sensors, such as optical, radar, LiDAR, and human sensors (individuals as virtual sensors). Considering the differences in imaging mechanisms, the fusion of data from diverse modalities can be conducted at the feature level and decision level. Particularly, the multimodal data fusion strategy must be carefully designed to ensure high fidelity in feature representation and enhanced accuracy in data interpretation.

Current challenges mainly include the following. The first is the differences in data structure. The second is the imbalanced number of labeled samples across modalities, which can lead to a significant gap in performance when models are individually trained. The third is the need for hundreds of millions of labeled samples, which are costly and often unavailable publicly. Deep learning models appear well suited to accommodate different data sources, due largely to the fact that they can directly learn the nonlinear relationship between input and output

representations and do not need any extra steps, e.g., hand-crafted feature extraction in traditional methods. Developing a unified foundation model for complex urban areas is crucial for building a unified urban mapping framework and promoting urban applications. To this end, multi-modal data can be incorporated together to enhance the sensing of urban areas as well as human activities. Moreover, prior knowledge and the physical rules of urban areas can be merged into the foundation model to construct knowledge-driven or physically constrained models for promoting the comprehensive sensing of cities.

A promising research direction is the development of diverse learning schemes, e.g., self- supervised learning and cross-modal learning. Self-supervised learning deals with unlabeled data and underpins deep learning's advances in large natural language models trained on web-scale corpora of unlabeled text, such as Generative Pre-trained Transformer (GPT)[41]. In computer vision, self-supervised learning can match and, in some cases, surpass models trained on labeled data[42], even on highly competitive benchmarks like ImageNet. How to leverage these recent advances in the AI field to facilitate the transformation of the remote sensing field will become a hot spot in the coming years. Cross-modal learning allows learning in any individual sensory modality to be enhanced with information from other modalities. The current AI systems have taken only tiny steps in the cross-modal learning direction. We can expect that more models will emerge as backbone methods in the remote sensing field.

Extending the capacity of urban observation and analysis

AI can construct mathematical models between big data and labels without human intervention. This section selected three cases, i.e., building detection, road detection, and human behaviors analysis, to discuss the influence of AI, considering the complexity of urban observation and analysis and the diversity of urban elements (visible objects or invisible interactions). Moreover, these cases have long been studied in remote sensing and can serve as demos.

AI gradually extends the boundary of building detection to more fine-grained and complex tasks, e.g., roof type identification, building function classification, and 3D building reconstruction. Currently, the accuracy of building edges is relatively low and building vectorization needs dedicated post-processing to replace manual work. Moreover, the reconstruction of building roofs and walls at city-to-regional scales is still underexplored. Many impervious area data products at 10–30 m resolution can serve as a spatial constraint to speed up large area building detection. Furthermore, geospatial big data, such as crowdsourced buildings from Open Street Map (OSM), should be fully explored, since they provide the diversity and representativeness of training data, but their quality must be assessed. OSM can be a viable alternative to official data sources even in data-scarce regions, because AI models can learn correct knowledge from high-quality data and apply it to correct low-quality data by using noisy label learning techniques[43]. The inclusion of data-scarce regions with or without crowd-sourced labels could be useful for improving the generalization ability of AI models, facilitating the construction of large models[18,44].

With respect to road extraction, a promising direction is to build a global-scale road training dataset using existing open-source big data. Besides, it will be beneficial to develop a cyclic AI framework, which can continually adjust the network parameters with the feed of new input data and effectively adapt the trained model to new data distribution. A data-driven and knowledge- guided AI framework would be more desirable, due to the combination of the advantages of visual perception and human cognition. Remote sensing imagery reflects physical processes, and physics-based models can provide important priori knowledge for AI models.

Furthermore, street-level imagery, nighttime light, human sensors, and geotagged data can provide a wide range of information about urban forms and underlying socio-economic dynamics. AI shows a promising potential in integrating these multi-source data and a few new directions can be identified. The first one is about effective data integration. A standardized AI model handling different datasets in terms of scale, format, and resolution would be key for improving the training performance. The second one is about down-to-earth AI model development.

Application-specific AI models would be more essential for urban sensing and intelligence with minimal fine-turning. The third one is about high-resolution map updating in an effective way. The final one relates to the tradeoff between data privacy and AI development. Allowing AI to access human behavior related data as much as possible should be considered with security issues in mind.

Three-dimensional semantic reconstruction of cities

The 3D semantic reconstruction of cities refers to accurately reconstructing the 3D scene geometry and simultaneously interpreting the scene to semantic object classes, such as individual buildings, trees, and roads. Figure 3 shows an example of semantic reconstruction. This information is the bedrock of digital infrastructures that fuel contemporary EO, digital twin, and smart city applications[45].

For many years, semantic 3D models have been created through manual or semi-automated approaches[46]. Today, 3D semantic reconstruction has entered an AI era. High accuracy and full automation of AI methods mean better estimates of human settlements and stronger spatio-temporal coverage that were previously not possible. For example, real-time semantic segmentation and depth estimation from video feeds are becoming basic components in major driving-assistant systems in autonomous vehicles and robots. In EO, the increased automation and accuracy mean the ability to interpret large volumes of data across different scales. Thus, contemporary 3D semantic reconstruction aims to break the data scale boundaries, producing landscape 3D semantic models at an extremely high spatial resolution (Fig. 3 depicts the desired outcome). In addition to semantics on the physical properties of the urban objects, those related to higher-level human perception of the scene can be very valuable. Examples of such high-level "semantics" include human understanding of building and community functions of a city, urban forms and planning[47]. However, studies on such topics with AI are still limited, which, by making use of current 3D semantic reconstruction methods, can be further sought using advanced co-learning of visual semantics and large language models[48,49] and multimodal data fusion approaches (Section "Multi-modal data fusion and physical model integration").

This field suffers from common AI problems. Good AI models require high-quality training data in considerable quantities. For example, most of the known public datasets focus on developed countries, which may highly likely lead AI models to generate results that may not be desirable in less developed regions. Future efforts can be put into unsupervised/weakly supervised training (for noisy or sparse data)[50] and domain adaptation[51] to alleviate the absence of sufficient training data. Further, there are some technical issues in 3D reconstruction. For instance, the construction of networks for multi-view images or LiDAR needs to consider the imaging geometry model to obtain the spatial location of each pixel, which is more complex than single-view images. Besides, the texture mapping and simulation for different objects require sophisticated computation and optimal parameter searching.

Real-time sensing, imaging, and processing

Real-time sensing, imaging, and processing are crucial for detecting, assessing, and managing various types of urban risks, including both natural and anthropogenic risks. Figure 4 illustrates GeoAI for real-time urban sensing, imaging, and processing. Satellite imagery provides abundant data for urban monitoring, and the incorporation of multiple satellites and the introduction of video satellites enables real-time observation[52]. IoT-integrated sensors and the advent of the Fifth-Generation technology facilitate faster data collection and transmission, bolstering real-time sensing and imaging. While advanced AI models offer robust real-time data handling capabilities, improving efficiency in tasks like traffic flow prediction and flood forecasting, it is essential to recognize the complexities involved. These models can effectively integrate diverse data from satellites, unmanned aerial vehicles, ground sensors, social media, and other platforms, providing a comprehensive and real-time view of urban environments and effective urban management, especially during emergencies and disasters. It is crucial to ensure that AI serves as a tool guided by human decision-makers, maintaining transparency and accountability. Integrating

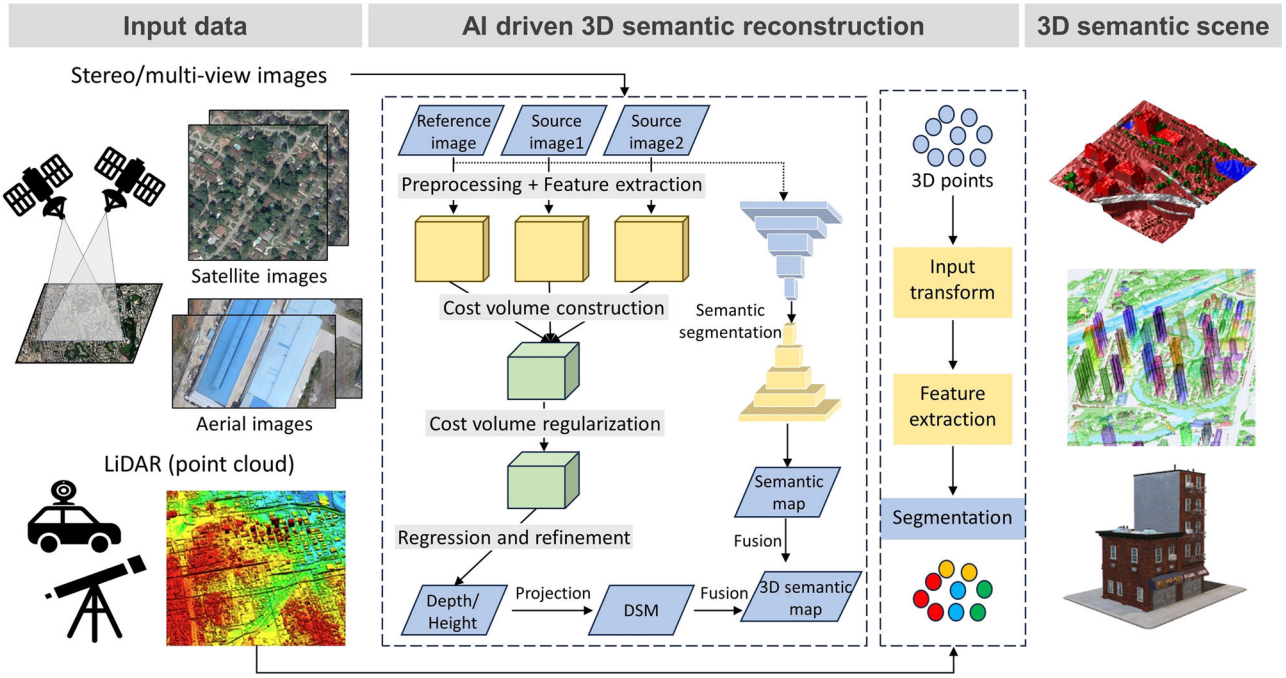

Fig. 3 | An example of semantic reconstruction of cities (image created by the authors).

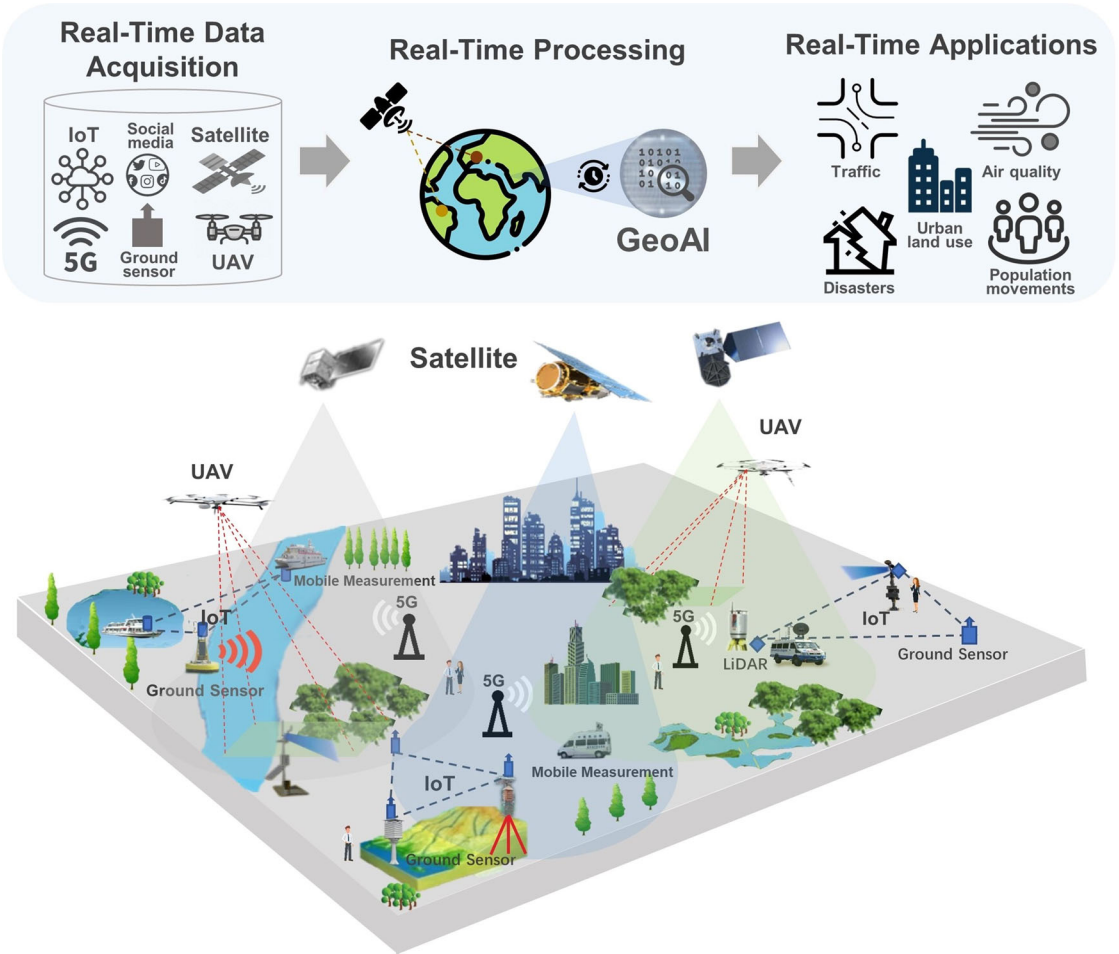

Fig. 4 | GeoAI for real-time urban sensing, imaging, and processing (image created by the authors).

human judgment is necessary to address complex urban challenges, and considerations of data access, ethics, and standardization must be prioritized.

Security concerns

Despite the great success of AI in urban environmental science, the challenges posed by AI security issues should not be neglected[53]. Due to the intrinsic characteristics of machine learning algorithms, AI models usually exhibit high vulnerability to adversarial attacks and backdoor attacks. These attacks can seriously threaten the reliability of Internet Of Thing devices (e.g., drones) and intelligent applications (e.g., autonomous driving) in smart city systems, simply by imperceptible adversarial perturbations or backdoor triggers. Besides, as AI systems gather and process an ever-expanding volume of data, protecting the data privacy of individual users within smart city systems becomes a pressing concern[54]. There is an urgent demand to establish a collaborative framework that not only enhances the protection of personal data but also preserves the autonomy of data providers, enabling them to actively contribute to the collective intelligence of AI systems. Furthermore, uncertainty exists throughout the entire lifecycle of urban remote sensing. This uncertainty constantly propagates and accumulates, thereby affecting the accuracy and reliability of the ultimate output generated by the deployed AI model.

These challenges reveal the imperative for the advancement of secure AI models in EO. Specifically, advanced techniques should be developed to improve the intrinsic resistibility against adversarial/backdoor attacks, while simultaneously identifying and mitigating potential threats posed by adversaries within the urban systems. Data privacy strategies such as federated learning should be embraced to decentralize the learning process, enabling AI models to be trained across distributed data sources without compromising the confidentiality of sensitive information held by individual urban data providers. Advanced algorithms should be designed to further decrease uncertainty, ensuring that errors and risks remain highly controllable to achieve a truly trustworthy AI system in urban environments. We believe that AI security will play an increasingly important role in shaping the future of digital, smart, and sustainable cities.

Conclusions

Over the past several decades, the field of EO has been developing rapidly with the advent of new sensors and algorithms, the reinvention of "old" technology, and more computing power such as AI, and is gaining great interest in academia, governments, industries, and the public. This paper provides the most up-to-date knowledge on AI for urban areas, what trends are emerging and how AI technology can be applied to provide practical solutions for a sustainable urban future.

We envision three trends in the future research of AI for urban observation, imaging, and analysis. First, AI will provide a deeper and more comprehensible interpretation of the fundamental principles underlying urban issues. Multi-modal data can expand urban sensing, imaging, and mapping capabilities, and previously obscure information can be rendered visible. The incorporation of interpretable AI techniques can facilitate effective analysis in addressing urban challenges and enable a better understanding of the disparities between AI and human policymaking. Second, AI provides a diverse range of precise methodologies to enhance the field of urban studies. AI-generated content empowers researchers to generate practical solutions tailored to address specific challenges in various scenarios. For instance, generative AI may have the potential to act as an agent to simulate and forecast complex urban dynamics, offering a granular understanding of how cities evolve over time under various scenarios. This technology will further facilitate the creation of customized urban designs in alignment with the goals of sustainable urban development established by governments and the United Nations. By leveraging AI, a range of alternative approaches to urban planning and design processes can be introduced, resulting in more effective and targeted solutions to the challenges faced in current urban development.

Data availability

Data sharing not applicable – no new data generated.

Received: 21 March 2024; Accepted: 11 November 2024;
Published online: 28 November 2024

References

1. Gamble, J. L., Ebi, K. L., Grambsch, A. E., Sussman, F. G. & Wilbanks, T. J. *Analyses of the Effects of Global Change on Human Health and Welfare and Human Systems*. *U.S. Environmental* http://www.environmentportal.in/files/Jul08sap4-6-CC_IMP_Health-FIRE-Repo.pdf (2008).
2. Weng, Q. Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sens. Environ.* **117**, 34–49 (2012).
3. Weng, Q. *An introduction to contemporary remote sensing*. (McGraw-Hill Education, 2012).
4. Zhang, F. et al. Measuring human perceptions of a large-scale urban region using machine learning. *Landsc. Urban Plan.* **180**, 148–160 (2018).
5. Trinder, J. C. Extraction of parameters from remote sensing data for environmental indices for urban sustainability. *Remote Sens. Sustain*. 3–27 (2017).
6. Zhu, X. X. et al. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **5**, 8–36 (2017).
7. Zhang, L., Zhang, L. & Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geosci. Remote Sens. Mag.* **4**, 22–40 (2016).
8. Lu, D. & Weng, Q. Spectral mixture analysis of the urban landscape in Indianapolis with Landsat ETM+ imagery. *Photogramm. Eng. Remote Sens.* **70**, 1053–1062 (2004).
9. Cracknell, A. P. Review article Synergy in remote sensing-what's in a pixel? *Int. J. Remote Sens.* **19**, 2025–2047 (1998).
10. Fan, Z., Zhang, F., Loo, B. P. Y. & Ratti, C. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proc. Natl. Acad. Sci.* **120**, e2220417120 (2023).
11. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
12. Chowdhary, K. R. Natural language processing. *Fundam. Artif. Intell*. 603–649 (2020).
13. Mankowitz, D. J. et al. Faster sorting algorithms discovered using deep reinforcement learning. *Nature* **618**, 257–263 (2023).
14. Wang, Y., Wei, G.-Y. & Brooks, D. A systematic methodology for analysis of deep learning hardware and software platforms. *Proc. Mach. Learn. Syst.* **2**, 30–43 (2020).
15. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems* 1097–1105 (2012).
16. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. *Adv. Neural Inf. Process. Syst*. **27**, 3104–3112 (2014).
17. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv Prepr. arXiv2010.11929* (2020).
18. Kirillov, A. et al. Segment anything. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4015–4026 (2023).
19. Jakubik, J. et al. Prithvi-100M. at https://doi.org/10.57967/hf/0952 (2023).
20. Liu, F. et al. RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–16 (2024).
21. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **538**, 20 (2016).

22. Brown, C. F. et al. Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci. Data* **9**, 251 (2022).
23. Lu, X., Zhong, Y. & Zhang, L. Open-source data-driven cross-domain road detection from very high resolution remote sensing imagery. *IEEE Trans. Image Process.* **31**, 6847–6862 (2022).
24. Chen, Y. et al. Developing an intelligent cloud attention network to support global urban green spaces mapping. *ISPRS J. Photogramm. Remote Sens.* **198**, 197–209 (2023).
25. Huang, X. & Zhang, L. Morphological building/shadow index for building extraction from high-resolution imagery over urban areas. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**, 161–172 (2011).
26. Ji, S., Wei, S. & Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Trans. Geosci. Remote Sens.* **57**, 574–586 (2018).
27. Hetang, C. et al. Segment Anything Model for Road Network Graph Extraction. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2556–2566 (2024).
28. Li, B., Gao, J., Chen, S., Lim, S. & Jiang, H. DF-DRUNet: A decoder fusion model for automatic road extraction leveraging remote sensing images and GPS trajectory data. *Int. J. Appl. Earth Obs. Geoinf.* **127**, 103632 (2024).
29. Demir, I. et al. DeepGlobe 2018: A challenge to parse the earth through satellite images. in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* vols 2018-June 172–181 (2018).
30. Bastani, F. et al. RoadTracer: Automatic Extraction of Road Networks from Aerial Images. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 4720–4728. https://doi.org/10.1109/CVPR.2018.00496 (2018).
31. Wang, R., Peethambaran, J. & Chen, D. Lidar point clouds to 3-D urban models $: $ A review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **11**, 606–627 (2018).
32. Jaritz, M., Vu, T.-H., De Charette, R., Wirbel, É. & Pérez, P. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1533–1544 (2022).
33. Aiello, E., Valsesia, D. & Magli, E. Cross-modal learning for image-guided point cloud shape completion. *Adv. Neural Inf. Process. Syst.* **35**, 37349–37362 (2022).
34. Vats, A. et al. Terrain-Informed Self-Supervised Learning: Enhancing Building Footprint Extraction from LiDAR Data with Limited Annotations. *IEEE Trans. Geosci. Remote Sens*. pp. 1–10 (2024).
35. Rouet-Leduc, B., Jolivet, R., Dalaison, M., Johnson, P. A. & Hulbert, C. Autonomous extraction of millimeter-scale deformation in InSAR time series using deep learning. *Nat. Commun.* **12**, 6480 (2021).
36. Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M. & Teitler, J. O. Using Google Street View to audit neighborhood environments. *Am. J. Prev. Med.* **40**, 94–100 (2011).
37. Sun, J. et al. Automatic atmospheric correction for shortwave hyperspectral remote sensing data using a time-dependent deep neural network. *ISPRS J. Photogramm. Remote Sens.* **174**, 117–131 (2021).
38. Salcedo-Sanz, S. et al. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. *Inf. Fusion* **63**, 256–272 (2020).
39. Huang, W. & Li, S. Understanding human activity patterns based on space-time-semantics. *ISPRS J. Photogramm. Remote Sens.* **121**, 1–10 (2016).
40. Li, H., Yuan, Z., Novack, T., Huang, W. & Zipf, A. Understanding spatiotemporal trip purposes of urban micro-mobility from the lens of dockless e-scooter sharing. *Comput. Environ. Urban Syst.* **96**, 101848 (2022).
41. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. (2018).
42. Zhang, J., Zheng, Y. & Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. in *Proceedings of the AAAI conference on artificial intelligence* vol. 31 (2017).
43. Song, H., Kim, M., Park, D., Shin, Y. & Lee, J. G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Trans. Neural Networks Learn. Syst*. https://doi.org/10.1109/TNNLS.2022.3152527 (2022).
44. Yang, L. Depth anything: Unleashing the power of large-scale unlabeled data. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10371–10381 (2024).
45. Zheng, Y., Liu, F. & Hsieh, H.-P. U-air: When urban air quality inference meets big data. in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* 1436–1444 (2013).
46. Gröger, G. & Plümer, L. CityGML–Interoperable semantic 3D city models. *ISPRS J. Photogramm. Remote Sens.* **71**, 12–33 (2012).
47. Lin, A. et al. Identifying urban building function by integrating remote sensing imagery and POI data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14**, 8864–8875 (2021).
48. Radford, A. et al. Learning transferable visual models from natural language supervision. in *International conference on machine learning* 8748–8763 (PMLR, 2021).
49. Hegde, D., Valanarasu, J. M. J. & Patel, V. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2028–2038 (2023).
50. Qin, R. & Liu, T. A review of landcover classification with very-high resolution remotely sensed optical images—Analysis unit, model scalability and transferability. *Remote Sens* **14**, 646 (2022).
51. Tuia, D., Persello, C. & Bruzzone, L. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geosci. Remote Sens. Mag.* **4**, 41–57 (2016).
52. Li, D., Wang, M., Dong, Z., Shen, X. & Shi, L. Earth observation brain (EOB): An intelligent earth observation system. *Geo-spatial Inf. Sci.* **20**, 134–140 (2017).
53. Xu, Y. et al. AI security for geoscience and remote sensing: Challenges and future trends. *IEEE Geosci. Remote Sens. Mag.* **11**, 60–85 (2023).
54. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **1**, 389–399 (2019).

Acknowledgements

This research has received funding from Global STEM Professorship, Hong Kong SAR Government (P0039329), Hong Kong RGC (grant reference # 15300923), and Hong Kong Polytechnic University (P0046482 and P0038446). The authors are grateful to the editors and reviewers for their constructive comments and suggestions which helped improve the manuscript.

Author contributions

Q. Weng conceptualized and designed the paper. Z. Li, Y. Cao, X. Lu, P. Gamba, X. Zhu, Y. Xu, F. Zhang, R. Qin, M. Yang, P. Ma, W. Huang, T. Yin, G. Asner prepared the original paper draft. Q. Weng, Z. Li, Y. Cao, X. Lu, F. Zhang, W. Huang, T. Yin contributed to the review and editing of

the manuscript. Q. Zheng and Y. Zhou participated in discussion of an earlier draft of this manuscript. All authors approved the manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Qihao Weng or Greg Asner.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024