Journal of Remote Sensing
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

# EcoVision: Submeter Land Cover Map over China's 42 Major Cities Derived by an Innovative Artificial Data Annotation Engine

**Encheng Zhang[1], Xin Huang[1*], Jiayi Li[1], and Jiawei Zhou[2]**

[1]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China. [2]Zhejiang Mingzhou Surveying and Mapping Institute, Ningbo 315100, China.

*Address correspondence to: xhuang@whu.edu.cn

**Citation:** Zhang E, Huang X, Li J, Zhou J. EcoVision: Submeter Land Cover Map over China's 42 Major Cities Derived by an Innovative Artificial Data Annotation Engine. *J. Remote Sens.* 2025;5:Article 0811. https://doi.org/10.34133/remotesensing.0811

Submitted 11 April 2025
Revised 24 July 2025
Accepted 24 July 2025
Published 4 September 2025

Land cover map can accurately characterize the spatial distribution of natural and artificial surface features. However, large-scale land cover products with submeter resolution are still scarce. To address this gap, this study proposes an innovative data annotation engine, called initial and expanded labeling, to generate reliable labels for high-resolution imagery. The engine takes imagery and historical products as input, generates a small number of labels using weight voting in the first stage, and iteratively expands the labels in the second stage. The proposed method can effectively deal with the insufficiency of training labels in large-scale submeter land cover mapping. Based on the datasets generated by this engine, we have produced the first large-scale submeter land cover map covering the urban areas of 42 major cities in China, called EcoVision. It has a spatial resolution of about 0.5 m with 8 representative urban land cover categories. The product has been validated with 23,850,000 randomly sampled validation pixels in 42 cities and has an overall accuracy of 83.6%. Compared with 5 existing land cover maps, EcoVision shows superior performance in spatial resolution, accuracy, and details. The product has been made public, providing high-precision data support for urban sustainable development research and territorial spatial planning.

## Introduction

Accurate land use and land cover (LULC) information is crucial for various studies, including environmental science, climate monitoring, food security, urban planning, disaster management, and ecosystem protection [1–3]. With the advancement of satellite technology, the spatial resolution of remote sensing imagery has been continuously increasing, providing data support for improving the spatial resolution of LULC mapping. Compared to low/medium-spatial-resolution imagery, high-resolution imagery contains rich information on land cover texture, shape, and spatial distribution, which has unique advantages for accurate mapping in highly heterogeneous regions such as densely populated megacities [4]. However, these detailed features also bring more complex land structures and patterns, posing challenges for LULC mapping using high-resolution imagery [5]. For medium/low-resolution LULC classification, people mainly rely on pixel-based machine learning algorithms, including decision trees [6], support vector machines [7], and random forest algorithms [8]. These classification methods mainly rely on spectral and textural features of pixels. However, in high-resolution imagery, land patterns contain limited low-level features in the spectral and spatial domains, making it difficult for traditional classification methods to identify fine land patterns, especially in urban areas [9].

In recent years, with the development of deep learning technology, neural networks have provided a more advanced and reliable solution for LULC classification. Compared to traditional methods, the major advantage of deep learning for image classification lies in its powerful feature representation capability, enabling it to learn and adaptively extract a large number of discriminative features for classification [10–12]. Although various deep learning-based LULC classification algorithms have been developed, the training process of common deep learning methods highly depends on a large number of high-precision labels, which are usually generated by tedious and time-consuming manual annotation for specific needs, severely limiting their extension to large-scale products [13]. The quality and diversity of training samples, which typically consist of raw images and corresponding labels, are key factors for large-scale model training. Previous studies have emphasized the vulnerability of deep learning models when dealing with data outside the distribution of training samples [14,15]. Notably, underperformance of the models is often attributed to the inadequacy of training data [14–16], including issues with the number, quality, representativeness, or diversity of samples. Although manually annotated samples can often provide more precise boundary and category information, their high time and labor costs limit their application in large-scale mapping. This also explains why the number of large-scale LULC products is relatively small, despite the increasing abundance of remote sensing imagery [17].

In order to reduce the workload of labeling for high-resolution LULC mapping, many efforts have been made in weakly supervised learning. Weakly supervised methods typically include incomplete supervised methods (only a subset of training samples have labels), inexact supervised methods (some training samples have no fine-grained labels), and inaccurate supervised methods (some training samples have wrong labels) [18,19]. Since weakly supervised methods do not require detailed manual annotation of the original images in the entire dataset, they can reduce the cost and time of manual labeling and have been adopted in LULC mapping. Cui et al. [17] designed a weakly supervised learning framework and achieved relatively good results in some areas of Hubei Province, China. Li et al. [9] produced a 1-m spatial resolution LULC map of Maryland, USA. Tong et al. [20] achieved land cover mapping results for 10 cities in different regions around the world. Although these weakly supervised learning studies have reduced the dependence on labels, they all only used one LULC product without incorporating higher-resolution products and products in other formats, resulting in models lacking precise supervised information. Therefore, compared with fully supervised models, these models still have difficulties in practical applications [17]. Meanwhile, researchers are also exploring the remote sensing application potential of foundation models. In Ref. [21], the zero-shot learning on the basis of Segment Anything Model (SAM) [22] was applied to remote sensing image segmentation tasks. In Ref. [23], the advantages of Contrastive Language-Image Pre-training [15] and SAM were combined for investigating the fine-grained semantic alignment and text-vision consistency. However, in general, foundation models trained in specific scenarios are difficult to transfer directly. They need retraining or fine-tuning with data from new scenarios. Although prompt learning can reduce dependence on annotated data, its effectiveness can be limited by the quality and quantity of prompts. In cases where annotation is scarce on a large dataset, the performance of the model can be affected. In addition, training labels are still crucial for improving model performance when dealing with complex and diverse remote sensing images.

Currently, most LULC studies have focused on the development of deep learning networks and the construction of loss functions [24–26], while the importance of training labels has been somehow overlooked [27]. In most of the current large-scale mapping, the labels usually rely on existing historical products rather than manual annotation, which improves the convenience of obtaining labels. However, a number of problems and challenges exist when applying the historical LULC products, such as the differences in data formats, mismatched spatial resolutions, inconsistent classification systems, pixel misalignment caused by image offset, different reference time, and classification differences between products. These uncertainties can lead to adverse effects on the generated labels as well as the subsequent modeling [27]. Therefore, there is an urgent need for new technologies to reduce the cost of collecting LULC labels for high-resolution imagery.

Based on the above considerations, (a) we proposed an innovative artificial data annotation engine called initial and expanded labeling (IEL), which can generate high-quality and diverse labels with the same resolution as the images, using a large number of available crowdsourced data as prompts. This alleviates the problems encountered in the joint use of crowdsourced data and the challenge of scarce labels in large-scale high-resolution LULC mapping. (b) Using the dataset generated by the IEL data annotation engine, we created a product called EcoVision. It is the first large-scale submeter-level LULC product covering the urban areas of 42 major cities in China, with a spatial resolution of 0.5 m and an overall accuracy of 83.6%.

## Materials and Methods

### Classification system

Based on the landscape style of urban areas in China and both domestic [28] and foreign research [29], EcoVision has defined a classification system consisting of 8 representative LULC categories for the urban areas of 42 cities in China: other impervious surface areas (OISAs), grass/shrubs, trees, soil, buildings, water, roads, and agriculture. Compared to Ref. [28], we added the category of agriculture to support detailed urban-ecological studies, and meanwhile, our land cover systems are consistent with Ref. [29], involving basic urban land cover categories.

Firstly, compared to medium/low-resolution imagery, high-resolution remote sensing imagery has richer information on land cover texture, shape, and spatial distribution, having the potential for more detailed classification. Therefore, as a common LULC type in urban areas, impervious surfaces are further divided into buildings, roads, and OISAs in EcoVision to facilitate more in-depth urban studies. Among them, buildings include various types of structures such as residential, commercial, and industrial buildings, roads include urban roads of all levels and transportation facilities such as bridges, and OISAs refer to hardened surfaces other than roads and buildings, such as parking lots and squares. These features are clearly visible on remote sensing imagery, are greatly affected by human activities, and are also areas of frequent human activity.

Secondly, LULC types such as grass/shrubs, trees, soil, agriculture, and water have important ecological and landscape values in urban areas. Grass/shrubs include natural pasture and artificial grassland, mainly areas dominated by small woody or herbaceous plants with a height of usually less than 2 m; trees include tree stands or tree patches of woody vegetation with a height of usually more than 2 m; soil refers to barren soil or sandy land without vegetation cover, such as construction sites and quarries; agriculture mainly includes land used for crop cultivation; and water includes rivers, reservoirs, lakes, and other areas with water throughout the year.

### Study areas

As shown in Fig. 1, we focused on 42 major cities in China, including 4 municipalities, 26 provincial capitals, and 12 large cities, which represent different urban scales, landscape characteristics, and urbanization intensities. We used the 2020 global urban boundary products [30] to extract the urban areas of the 42 cities [31] as the mapping areas. The selection of cities is the same as the Hi-ULCM product [28], aiming to generate an LULC product with a wider coverage area, higher resolution, and richer details based on it. In addition, the category of agriculture is added to support more detailed urban ecological studies.

### Imagery and crowdsourced geographic information

We collected images with a spatial resolution of 0.5 m from Google images, which were acquired approximately between 2019 and 2020. Due to their diverse sources, the images exhibit a variety of imaging conditions such as different seasons, angles, and tones. To mitigate the impact of color variations, all images were pre-processed using standardization and normalization techniques [32]. Additionally, the input patch size for the deep

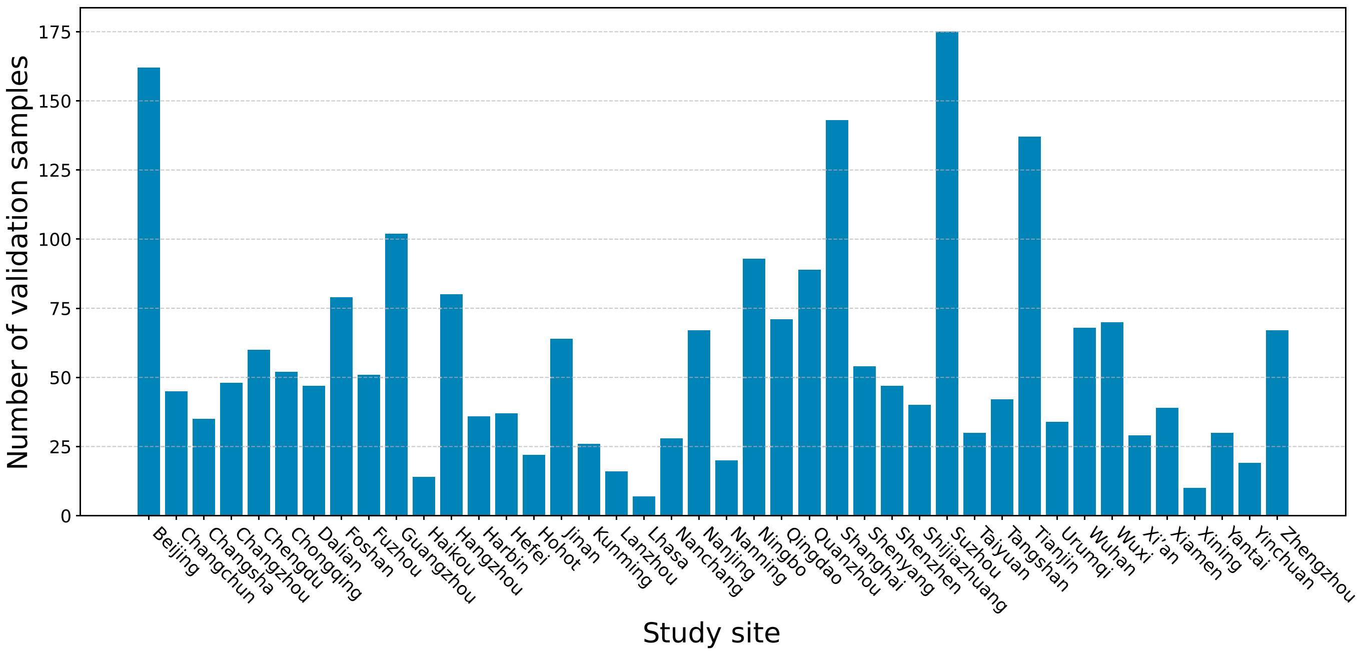

**Fig. 1.** The number of validation samples for the 42 study sites in China, including 4 municipalities, 26 provincial capitals, and 12 large cities.

learning network was defined as 1,024 × 1,024 pixels. In the end, about 270,000 remote sensing images were collected for these 42 major cities in China.

We collected currently popular large-scale products related to LULC as our crowdsourced data:

1. Huang et al. [33] used a semi-supervised learning method, combined with incomplete labels from OpenStreetMap (OSM), and used Google imagery from 2019 to 2020 to extract China's 0.5-m-resolution building footprint data (CBF). The overall F1-score [34] of CBF reached 83.71%.
2. Li et al. [35] used Google imagery from 2019 to 2020 and deep learning technology to produce the first open map of urban construction sites (UCS) in China at 0.5 m resolution, with a classification accuracy of 94.2%.
3. Huang et al. [28] used 2015 ZY-3 satellite imagery and a semi-automatic process to produce a 2-m-resolution urban land cover map covering 42 major cities in China, called Hi-ULCM, with an overall accuracy of 88.6%. Its classification system includes 7 categories: OISAs, grass/shrubs, trees, bare soil, buildings, water, and roads.
4. Li et al. [25] used a semi-supervised learning method to create the first national land cover map of China at 1 m resolution using a deep learning framework and open-access data based on Google imagery from December 2022, with an overall accuracy of 73.61%. Its classification system includes 11 categories: tree cover, shrubland, grassland, cropland, building, traffic route, barren and sparse vegetation, snow and ice, water, wetland, and moss and lichen.
5. Huang et al. [36] used Sentinel-1 and Sentinel-2 data from 2016 and machine learning algorithms to produce the global impervious surface area (GISA) map at 10 m resolution, with an overall accuracy of 86%.
6. OSM data are a product widely used in global urban research [37]. They are a series of user-generated maps collaboratively edited by a large number of volunteers. The data quality and completeness of OSM vary greatly across different regions, and the data in some areas may not be detailed or accurate enough. We downloaded OSM data related to LULC categories (OSM_Polygon) and road-related data (OSM_Polyline) in June 2019.

The crowdsourced geographic data information we used are shown in Table 1.

## Annotation of validation samples

It should be noted that although the 42 cities we focused on are all densely populated regions in China, they cover a large geographical area and exhibit diverse natural landscapes. The differences in these natural landscapes have an important impact on urban LULC patterns, especially as some cities display unique urban landscapes or surface materials [28]. Therefore, to evaluate the effectiveness of IEL in large-scale LULC mapping, we implemented a random sampling scheme. We randomly collected image patches in each study city and marked them as validation samples through manual interpretation. Specifically, 2,385 image patches were collected in the 42 study cities, with each patch being 100 × 100 pixels, covering an area of approximately one-ten-thousandth of the study area. All 23,850,000 pixels were marked by manual visual interpretation. The patches in each city were selected randomly, and the number of patches depended on the urban area of each city (Fig. 1). It should be pointed out that the selected cities have unique and diverse landscape characteristics, which not only ensures the breadth of the validation samples but also increases the representativeness and reliability of the samples. The annotation of the validation samples was completed by a team of 3 interpreters within 1 month. To ensure consistency among the interpreters, they interpreted the ground objects in the images synchronously as a team. The proportions of LULC categories in the validation samples are as follows: OISAs: 20.8%, grass/

**Table 1.** Crowdsourced geographic information

| Product | Format | Resolution (m) | Time | Category |
|---|---|---|---|---|
| CBF [33] | Raster | 0.5 | 2019–2020 | Buildings |
| UCS [35] | Raster | 0.5 | 2019–2020 | Urban construction sites |
| Hi-ULCM [28] | Raster | 2.0 | 2015 | OISAs, grass/shrubs, trees, bare soil, buildings, water, and roads |
| SinoLC-1 [25] | Raster | 1.0 | 2022 | Tree cover, shrubland, grassland, cropland, building, traffic route, barren and sparse vegetation, snow and ice, water, wetland, and moss and lichen |
| GISA [36] | Raster | 10.0 | 2016 | Impervious surface area |
| OSM | Vector | | 2019 | OISAs, grass/shrubs, trees, soil, buildings, water, roads, and agriculture |

shrubs: 8.8%, trees: 11.1%, soil: 7.5%, buildings: 24.5%, water: 7.1%, roads: 11.9%, and agriculture: 8.3%.

## IEL data annotation engine

To address the issues of low efficiency in manual labeling and the difficulty in constructing costly label sets for large-scale LULC mapping, this study developed a data annotation engine called IEL (Fig. 2), which consists of 2 stages: (a) Priority-based weighted voting stage (stage 1): Prompt data (prompt) matching the image resolution are generated by pre-processing crowdsourced data. A small number of high-confidence pixels of determined categories (trusted pixels) are obtained through weighted voting based on product quality and spatiotemporal information, forming the initial label map ($\text{Label\_Map}_0$). (b) Iterative expansion stage (stage 2): An end-to-end semantic segmentation model is trained based on $\text{Label\_Map}_0$ to predict the imagery and generate new prompt data. The predicted results are combined with other products through a second voting process to build an expanded label. The model and label are iteratively optimized ($\text{Label\_Map}_i \rightarrow \text{Label\_Map}_{i+1}$) until the model's performance stabilizes, ultimately obtaining a high-precision mapping model and label.

Overall, this method includes 2 stages, aiming to obtain a large number of accurate and diverse labels from a large number of available crowdsourced data, and ultimately achieve the production of submeter-level LULC products for 42 major cities in China. The technical details of each stage of IEL are described in the following text.

## Priority-based weighted voting

As introduced in the "Imagery and crowdsourced geographic information" section, the crowdsourced data related to LULC include vector data, LULC products, and thematic products. However, directly generating labels from these data would face 3 major challenges: First, the data formats are inconsistent and the spatial resolutions are mismatched. Second, the classification systems are inconsistent. Third, there are label conflicts, mainly manifested as pixel misalignment due to image shifting, inconsistent reference times, and classification differences between products. To address these 3 issues, this study designed a weighted voting strategy to achieve low-cost automated label production in 3 steps.

The first step is to address the differences in data formats and the mismatches in spatial resolution in crowdsourced data. Polygon data can be converted to raster data through rasterization, and polyline data can be transformed to raster data by expanding a buffer zone around their centerlines. The resolution of the raster data obtained from this process is set to the same as that of the target imagery, thus unifying all data into raster format. Subsequently, the nearest-neighbor interpolation method can be used to resample the remaining products with mismatched resolutions. Given that the LULC values in the products are discrete data, the nearest-neighbor interpolation can effectively ensure the stability of the LULC information while generating more detailed new values.

The second step is to address the inconsistency of product classification systems within the crowdsourced data. In order to avoid the subjectivity caused by human operations, we introduced large language models, specifically GPT-4o [38], to objectively establish classification mapping relationships across different LULC products. This was achieved through the following prompt-based approach: "Given your expertise and extensive experience in the field of LULC mapping, please convert the classification system used in <an existing product (reference)> into the classification system of <the target system>." In this study, the target classification system comprises 8 categories: OISAs, grass/shrubs, trees, soil, buildings, water, roads, and agriculture. GPT-4o can achieve the mapping of different classification systems by learning the knowledge from the reference materials. For example, for the SinoLC-1 used in this study, GPT-4o successfully mapped several categories, such as tree cover to trees, shrubland and grassland to grass/shrubs, cropland to agriculture, building to buildings, traffic route to roads, barren and sparse vegetation to soil, and water to water. Furthermore, it identified that there are no corresponding mapping relationships for the categories of snow and ice, wetland, and moss and lichen within the SinoLC-1 classification system relative to the target system. Similarly, the categories of all the LULC products can be mapped according to the reasonable relationships provided by GPT-4o, transforming them into the classification system of the final product. The same method can be applied to the thematic products for category mapping. It should be noted that, due to differences in category definitions, the mapping relationships may not always be simple one-to-one or many-to-one mappings, and there may also be complex mappings such as one-to-many or many-to-many. For example, GISA records

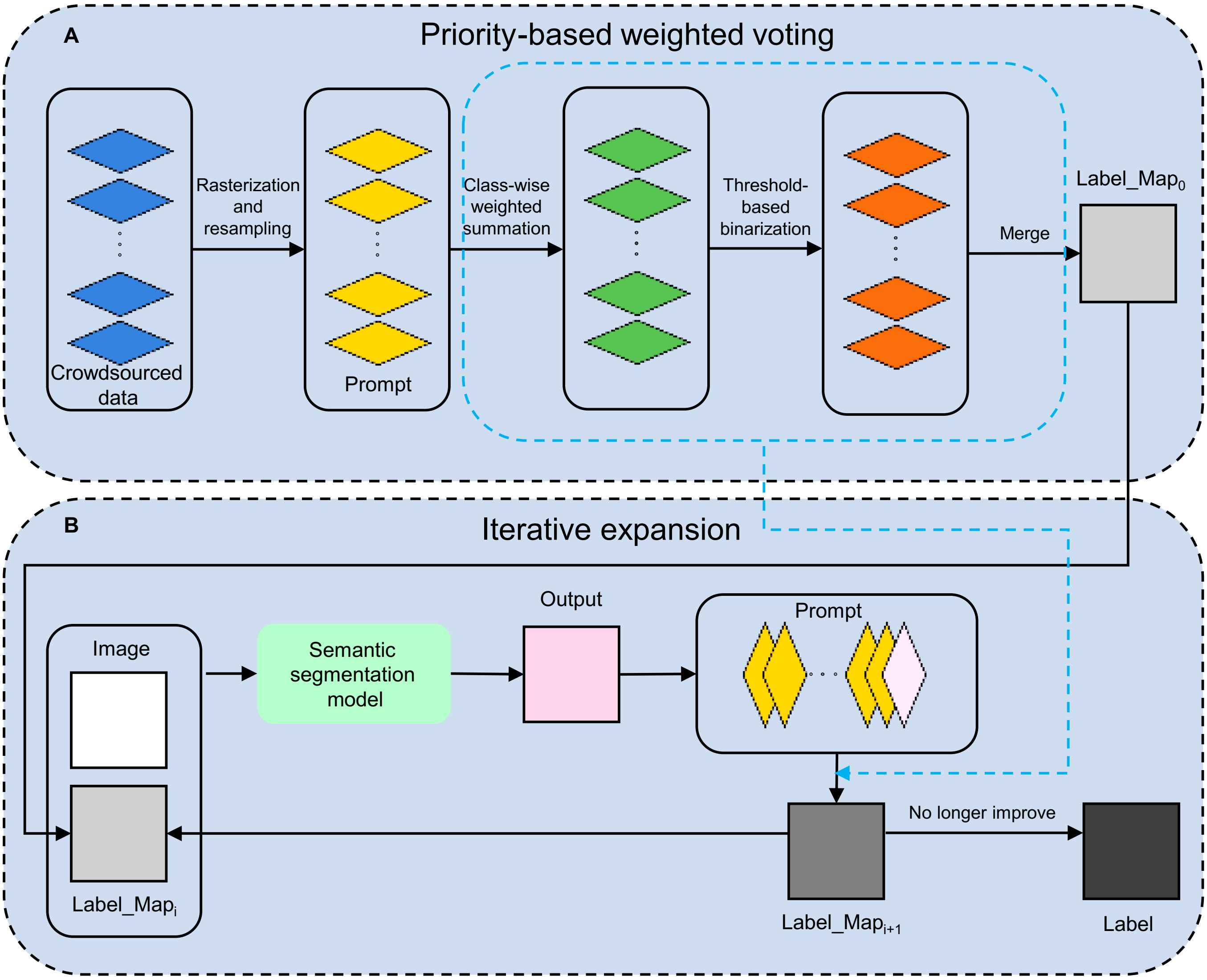
A
Priority-based weighted voting
Crowdsourced data
Rasterization and resampling
Prompt
Class-wise weighted summation
Threshold-based binarization
Merge
Label_Map$_0$
B
Iterative expansion
Image
Semantic segmentation model
Output
Prompt
Label_Map$_i$
Label_Map$_{i+1}$
No longer improve
Label

**Fig. 2.** The initial and expanded labeling (IEL) data annotation engine for LULC mapping using semantic segmentation model and crowdsourced data. (A) Priority-based weighted voting. (B) Iterative expansion.

all impervious surfaces, which correspond to the 3 more specific impervious surface categories (OISAs, roads, and buildings) in the EcoVision classification system; Hi-ULCM does not have the agriculture category, and GPT-4o suggests that agriculture may be mixed into the bare soil and grass/shrubs categories, resulting in a 2-to-3 mapping relationship between bare soil and grass/shrubs in Hi-ULCM and the agriculture, grass/shrubs, and soil in the EcoVision classification system. After the above pre-processing, each crowdsourced data are converted into raster data, representing a single category or multiple categories, which we call "prompt".

The third step aims to solve the problem of conflicting label classification (Fig. 3) by proposing a priority-based weighted voting strategy to generate initial labels for image patches. The strategy is based on each image-patch pixel and centers on 3 key hyperparameters: priority sequence, weights, and confidence thresholds. The specific implementation process is as follows: For each image patch, all corresponding prompts are collected. Then, based on these prompts, the priority sequence is established by using the principle of "easy-to-judge first". Prompts mapped to the target category are collected in priority order, and the confidence for each pixel in the image patch is computed by weighted summation based on reference time, spatial resolution, and accuracy of each prompt. Once the threshold for that category is reached by a pixel's confidence, the pixel is considered to be a trusted pixel, which is classified accordingly, and excluded from further calculations. The remaining pixels continue to participate in the evaluation of the next priority category until all category determinations have been completed. The labeling results generated in this stage are defined as Label_Map$_0$ in this study.

The priority sequence for each image patch follows the "easy-to-judge first" principle. Specifically, for each image patch, categories supported by single-category thematic products with the same or higher spatial resolution are given priority, as these data can meet the requirements

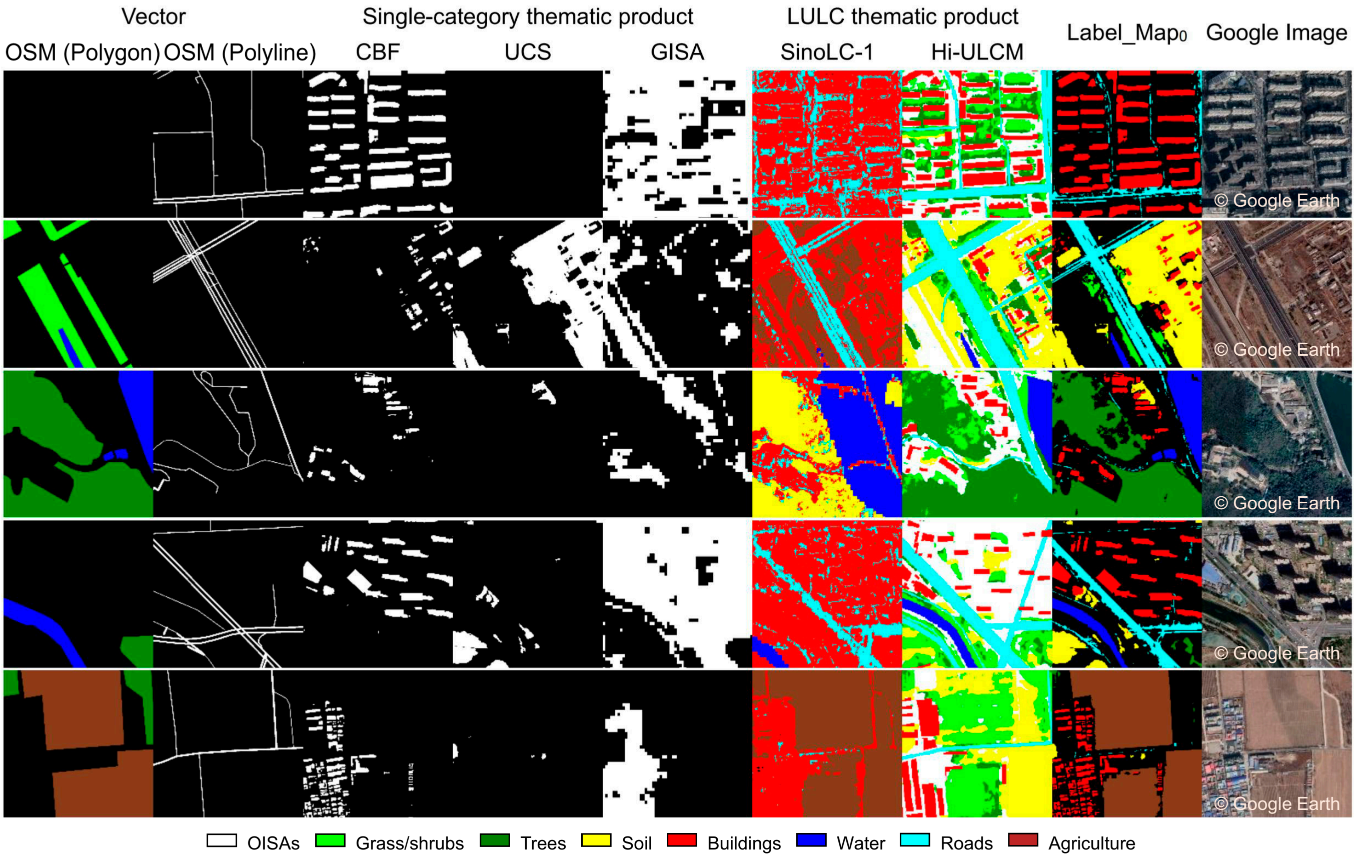

**Fig. 3.** The classification differences between various prompts, as well as the results of priority-based weighted voting.

for spatial refinement. Secondly, categories are sorted in descending order of the number of prompts, aiming to reduce the risk of misjudgment caused by insufficient prompts. Finally, categories are prioritized based on the validation accuracy of each prompt reported in its literature, ensuring that categories with high reliability are determined first.

The weights ($k$) are determined in terms of the spatiotemporal consistency and validation accuracy. As shown in Eq. 1, level 1 weight ($k_1$) is given to the thematic products with superior spatiotemporal consistency and accuracy; level 2 weight ($k_2$) is given to the products with slightly inferior but still acceptable spatiotemporal consistency and accuracy; and level 3 weight ($k_3$) is given to products that have poor spatiotemporal consistency and accuracy (the definitions of superior, slightly inferior, and poor in this study are given in the "Framework implementation" section). In this way, the confidence of each pixel corresponding to the category can be calculated using Eq. 2, where $j = 0, 1, \ldots, m$ represents the index of prompts, $i = 0, 1, \ldots, n$ represents the category of the final product, and $k_i^j$ and $a_i^j$, respectively, indicate the weight and opinion of the corresponding $\text{Prompt}_j$. If $\text{Prompt}_j$ considers the current pixel as the foreground of that category, $a_i^j$ is 1; otherwise, it is 0.

$$k = \begin{cases} k_1, & \text{if superior} \\ k_2, & \text{if slightly inferior} \\ k_3, & \text{if poor} \end{cases} \tag{1}$$

$$\text{Confidence}_i = \sum\nolimits_{j=0}^{m} k_i^j a_i^j$$
$$\text{where } a_i^j = \begin{cases} 0, \text{if Prompt}_j = 0 \\ 1, \text{other} \end{cases} \tag{2}$$

The process of determining the confidence thresholds for each category in each image patch is divided into 2 steps: (a) For categories with simple mapping relationships (i.e., one-to-one or many-to-one), determine whether there is a level 1 weight prompt under that category in the image patch. If so, its threshold is set as the level 1 weight to fully exploit the potential of high-quality data; otherwise, the thresholds are set as the sum of $n$ associated level 2 and level 3 weight, aiming to avoid an excessive mixing of low-confidence pixels, thereby ensuring the accuracy of the screening results. (b) For categories with complex mapping relationships (one-to-many or many-to-many), the thresholds are set by adding the sum of the weights of all complex mapping relationship prompts' weights to the threshold determined in the first step, to avoid misclassification. We summarized the implementation process of the proposed priority-based weighted voting strategy for an image patch using the pseudo code (Algorithm 1), where H and W represent its height and width, respectively, and K represents the number of prompts corresponding to the currently processed category:

**Algorithm 1.** Pseudo code for priority-based weighted voting strategy.

```
# Ps: the prompts corresponding to the current image patch, each sized H × W
# map: find prompts that have a mapping relationship with the current class
# ps: rules for setting the priority sequence of target product classes
# w: rules for setting weights
# t: rules for setting thresholds
# Out: label after the first stage
Cs = ps(Ps) # priority sequence, sized N, represents the n categories of the target product
for C in Cs do
  P = map(C, Ps) # the prompts mapped to the C category, sized K × H × W
  Ws = w(P) # Weights, sized K × 1 × 1
  Ts = t(P) # Thresholds, sized N
  Conf = sum(P × Ws, axis=0) # confidence of the current category, sized H × W
  Out = where(Conf > Ts[C] and Out is Null, Conf, Out) # only set categories for undetermined pixels
```

It is worth noting that considering that all prompts have the same spatial resolution as the image after pre-processing, and the above process is performed pixel by pixel, the spatial resolution of the obtained labels Label_Map$_0$ is the same as that of the image. For our submeter-level Google imagery, Label_Map$_0$ is still at the submeter level. Figure 3 shows the results of priority based weighted voting.

## Iterative expansion

So far, there are still some pixels whose LULC category has not been determined, mainly due to the limitations of each prompt, which result in the confidence levels for all categories corresponding to these pixels not reaching the preset threshold. In addition, Label_Map$_0$ is largely constrained by the coverage area of historical products, making it difficult to apply in regions where historical products are scarce. Therefore, we further introduced a semantic segmentation network to fully ensure the number of prompts corresponding to each image patch and gradually classify the remaining undetermined pixels. The semantic segmentation model is a deep learning model that automatically learns the category of each pixel based on the contextual information of the image patch by training on a large amount of labeled data, achieving pixel-by-pixel classification of the image patch.

Specifically, the Label_Map$_0$ generated in the first stage was used as the label for the image patches to construct a dataset. A semantic segmentation network was trained based on this dataset, ignoring the loss propagation of pixels with undetermined categories in the labels, thus obtaining a network model that has learned LULC classification knowledge. Unlike Label_Map$_0$, which has pixels with undetermined categories, the trained semantic segmentation model can classify all pixels of each image patch based on the classification knowledge in the dataset, generating a predicted result of the same size. Although the model's prediction results can be directly used as the final mapping results, to make more rational use of its information, the model's prediction results were regarded as a new type of prompt. By combining historical products, we continued to vote for all pixels with undetermined categories in the image patches, generating labels with more trusted pixels, denoted as Label_Map$_1$. Compared to directly using the model output for mapping, the introduction of prompts aims to further utilize information from historical products, alleviate bias during model training, improve label accuracy, and enhance the model's ability to classify complex land features.

The difference of the voting process between the first and the second stage lies in the addition of a new prompt, "model prediction", for each image. According to the setting principles of priority sequence, weights, and confidence thresholds described in the "Priority-based weighted voting" section, for each image patch, the priority sequence remains unchanged. Meanwhile, the one-to-one mapping relationship between the new prompt and the target product also keeps the confidence thresholds unchanged and the weights of the other prompts remain the same, while the weight of the new prompt is determined only by the validation accuracy analysis. In this way, "model prediction" is involved in the priority-based weighted voting strategy to generate Label_Map$_1$ with more trusted pixels.

It can be seen that the new label Label_Map$_1$ involved in the semantic segmentation model can be used to construct a new dataset with the original image patches, and the new dataset can be used to train a new semantic segmentation model. Therefore, to obtain the best model and labels, as shown in Fig. 2B, we incorporated model training and label expansion into a loop process. Specifically, in each round, the dataset generated by the initial labels and the original image patches was used to train the model. When the accuracy of the model did not improve on the validation samples, the model outputs for the image patches were fed into the priority-based weighted voting to generate new labels with more trusted pixels. This iterative process continued until the accuracy of the model on the validation samples reached a saturation point, resulting in the final mapping model and label. It is worth noting that in this iterative process, the priority sequence, weights, and confidence thresholds of each image patch did not change, and the trusted pixels of each round were saved to the next loop, aiming to make full use of the pixel labels generated in the previous steps to assist model learning. Among them, Label_Map$_i$ is iterated to obtain Label_Map$_{i+1}$, and the final label is Label_Map$_x$. Figure 4 shows the iterative process and final results of the labels of several of our Google image patches in the second stage. In addition, we have summarized the implementation process of the proposed iterative expansion method for image patches using pseudo code (Algorithm 2):

**Algorithm 2.** Pseudo code for iterative expansion.

```
# Is, Iv, Ls, Lv, Ps: all images, validation images, their labels and all prompts
# train: train model with all images and labels until validation loss no longer decreases
# val: validate model with validation images and labels
# infer: perform model inference
# pwv: priority-based weighted voting
# Out: the final label
V_max = -inf
while True:
  train(model, Is, Ls, Iv, Lv)
  V = val(model, Iv, Lv)
  if V > V_max :
    V_max = V
    P = infer(model, Is)
    Ls = pwv(P, Ps, Ls)
  else:
    break
```

In this way, by fully considering the historical product information corresponding to each image patch and using the semantic segmentation model with submeter-level image LULC classification knowledge, IEL generated high-quality and diverse labels. The final large-scale mapping result is based on the final label Label_Map$_x$, with the mapping model supplementing the small number of undetermined pixels.

## Framework implementation

The IEL data annotation engine was implemented using PyTorch 2.1.1 and Python 3.9, and runs on a single NVIDIA GeForce RTX3090 TI graphics processing unit (GPU) accelerator equipped with a memory capacity of 24 GB. The deep learning model involves different semantic segmentation models in the second stage of IEL. We selected several state-of-the-art semantic segmentation networks for our experiments. These models, encompassing CNN (convolutional neural network)-based and transformer-based semantic segmentation approaches, included UNet [39], HRNet [40], PSPNet [41], DeepLab V3+ [42], ViT [43], Swin transformer [44], and SegFormer [45]. The first 4 models are CNN-based: UNet has a symmetric encoder–decoder structure with skip connections, HRNet maintains high-resolution features throughout the model, PSPNet incorporates a pyramid pooling module, and DeepLab V3+ combines depthwise separable convolution and atrous convolution. The last 3 networks are transformer-based: ViT applies transformers directly to image patches, Swin transformer proposes a hierarchical structure with shifted windows, and SegFormer integrates multiscale features in a simple yet effective manner for semantic segmentation. In the comparative experiments, Label_Map$_0$ was uniformly used as the label for Google image patches to train these 7 models. After validation, the OA of the above models were 79.92%, 80.06%, 78.89%, 80.12%, 78.92%, 79.85%, and 78.22%, respectively. DeepLab V3+ showed the best performance during training. Therefore, the trained DeepLab V3+ was used in the iterative expansion stage of IEL. In the training stage, the models were uniformly trained for 10 epochs using the Adam optimizer with a batch size of 1, and only the model with the best performance on the validation samples was retained. All models were supervised by cross-entropy, with a learning rate of 0.01. The momentum parameter was set at 0.9, complemented by a weight decay of 0.0005.

Detailed mapping relationships between various products and the EcoVision classification system are presented in Table 2. In this study, level 1 weight ($k_1 = 0.3$) was assigned to thematic products with a time difference of no more than 1 year, a spatial resolution no lower than that of the target imagery, and an accuracy higher than 85%; level 2 weight ($k_2 = 0.2$) was given to products with a time difference of no more than 5 years, a resolution no lower than 4 times the spatial resolution of the target imagery, and an accuracy higher than 80%; and level 3 weight ($k_3 = 0.1$) was applied to the remaining products. The threshold was set as the sum of associated 3 level 3 prompts' weights. The priority sequence, weights, and confidence thresholds in Table 3 can be used to convert the products used in this study into labels, and their accuracy assessment is presented in "Performance of each stage of IEL" section.

## Accuracy assessment

We compared EcoVision with the validation samples using 4 popular metrics—overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), and F1-score. In addition, we further visually compared EcoVision with 2 sets of LULC products, including 2 high-resolution products: SinoLC-1 and Hi-ULCM, and 3 open-access medium-resolution LULC products: ESRI_2020 [46], ESA_WorldCover [47], and FROM-GLC10 [48].

# Results

## Performance of EcoVision

Figure 5 shows the overall visual illustration of EcoVision for some cities, as well as the mapping results and details in the representative city of Nanjing. Figure 5 shows that buildings, roads, and OISAs are distributed in the city center, while roads crisscross between various blocks, forming a dense network. The main river meanders from the suburbs through the city, and other water bodies such as ponds, reservoirs, and creeks are distributed in a patchy or linear pattern. Although agriculture, soil, grass/shrubs, and trees are also sparsely distributed in the urban area, they are mainly found in the suburbs. Therefore, the overall visual results of EcoVision accurately reflect the spatial distribution of multiple LULC categories and demonstrate the actual layout and development characteristics of Chinese cities.

Using 2,385 validation samples of size 100 × 100 generated from 42 cities, the OA of EcoVision is 83.63%. Table 4 shows the confusion matrix calculated from the validation samples, which is used to represent the OA of EcoVision, as well as the corresponding PA, UA, and F1-score for each category. Considering that, to our knowledge, this is the first submeter-level LULC product covering a large area of China, and given the proposed IEL engine, which uses a weakly supervised approach to eliminate the tedious task of manually drawing labels, the mapping results are generally satisfactory.

As seen from Table 4, the F1-scores of all LULC categories except grass/shrubs (68.38%) and soil (78.00%) are above 80%.

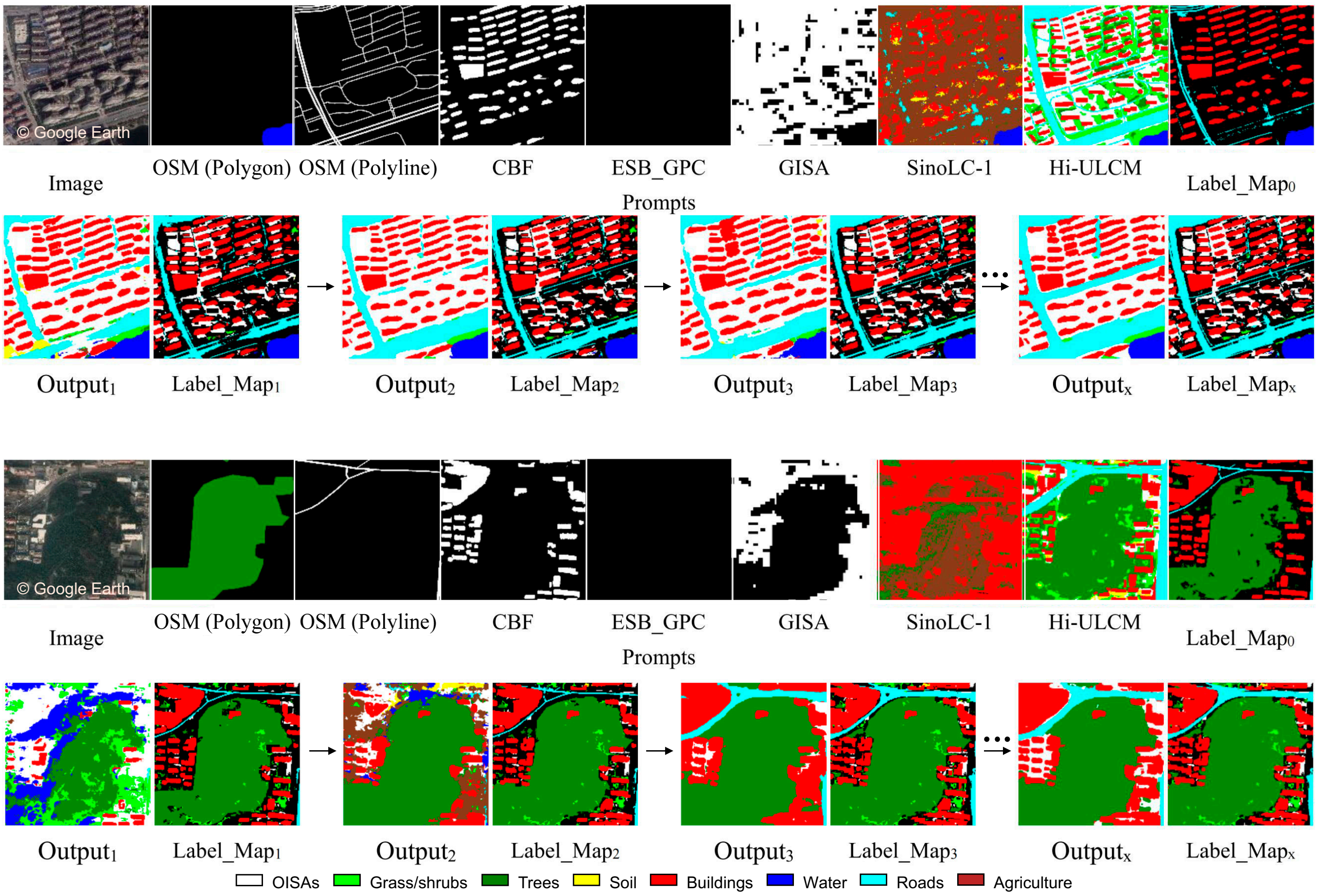

**Fig. 4.** The generation process of label within the iterative expansion.

**Table 2.** Category relations between various products and the proposed EcoVision. The road polyline is uniformly expanded by a buffer zone of 1.5 m. Italic indicates product names.

| Vector product | LULC thematic product | | Single-category thematic product | Output product |
|---|---|---|---|---|
| *OSM* | *Hi-ULCM* | *SinoLC-1* | | *EcoVision* |
| Apron, dam, fuel, parking, pitch, sports center, track, stadium | OISAs | | *GISA* | OISAs |
| Grass, meadow | Grass/shrub | Shrubland, grassland | | Grass/shrub |
| Forest, orchard, scrub | Trees | Tree cover | | Trees |
| Quarry | Soil | Barren and sparse vegetation | *UCS* | Soil |
| Car dealership, bank, commercial | Buildings | Building | *CBF, GISA* | Buildings |
| Water, reservoir, dock, riverbank | Water | Water | | Water |
| Living street, motorway, motorway link, path, primary, primary link, residential, secondary, secondary link, service, tertiary, tertiary link, track, trunk, trunk link | Roads | Traffic route | *GISA* | Roads |
| Farmland | Grass/shrub, Soil | Cropland | | Agriculture |

**Table 3.** Voting parameter settings of different prompts for Label_Map$_0$ in EcoVision

| Category | CBF | UCS | Hi-ULCM | SinoLC-1 | GISA | OSM | Threshold |
|---|---|---|---|---|---|---|---|
| Buildings | 0.3 |  | 0.2 | 0.1 | 0.1 | 0.1 | 0.4 |
| Soil |  | 0.3 | 0.2 | 0.1 |  | 0.1 | 0.5 |
| Roads |  |  | 0.2 | 0.1 | 0.1 | 0.1 | 0.4 |
| Water |  |  | 0.2 | 0.1 |  | 0.1 | 0.3 |
| Trees |  |  | 0.2 | 0.1 |  | 0.1 | 0.3 |
| Grass/shrub |  |  | 0.2 | 0.1 |  | 0.1 | 0.4 |
| OISAs |  |  | 0.2 |  | 0.1 | 0.1 | 0.4 |
| Agriculture |  |  | 0.2 | 0.1 |  | 0.1 | 0.4 |

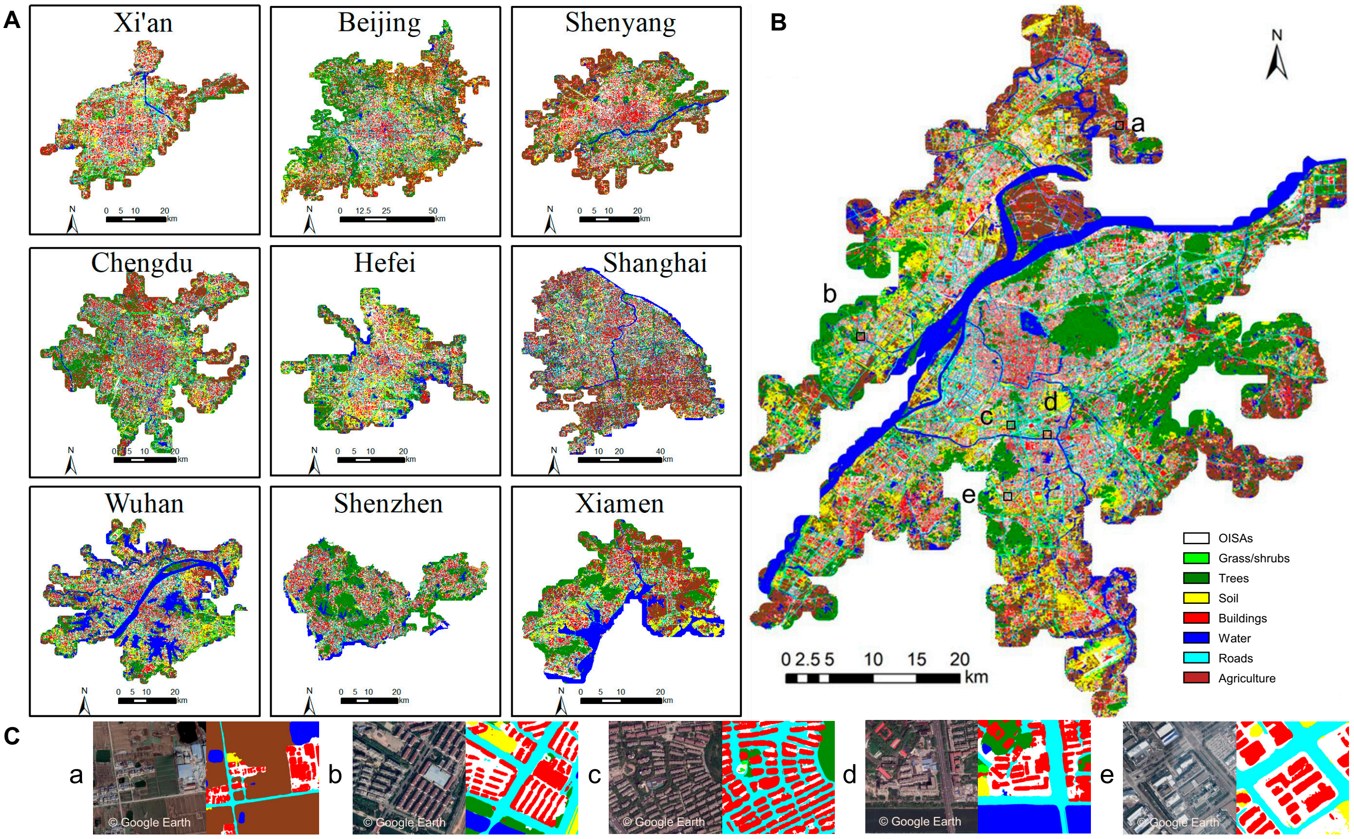

**Fig. 5.** An illustration of EcoVision for some cities (A), the results for Nanjing (B), and the close-up maps of Nanjing and the corresponding Google images (C).

Among them, the F1-score of buildings is the highest (90.00%), which reflects the efficacy of the high-quality product CBF in label production in this study. In addition, the results of F1-score indicate that the performance of EcoVision for the basic urban scenes—buildings (90.00%), roads (83.04%), and OISAs (81.09%)—can meet the needs of urban spatial analysis. For each LULC category, except for OISAs, grass/shrubs, and trees, the PA of the remaining 5 LULC categories is higher than the UA. The performance of the final product is satisfactory in most categories, as the accuracy of most categories is above 75%. As shown in Fig. 6, the UA, PA, and F1-score of the 8 land cover categories in the 42 cities are generally stable. Categories like buildings, water, and agriculture have stably high accuracy, with median F1-scores above 85%. Overall, the performance of EcoVision is satisfactory and has the potential to provide more detailed data for the study of urban environments.

## City-level LULC in EcoVision

We summarized the area of each LULC category in the urban areas of all 42 studied cities. Figure 7 shows the geographical regions and the percentages of LULC categories for each city, facilitating comparisons of the proportions of various LULC categories between or within cities. Among the 42 cities, the area occupied by buildings has the largest proportion, averaging

19.17%, followed by agriculture (17.29%), OISAs (13.37%), trees (12.80%), roads (11.18%), soil (11.09%), water (9.44%), and grass/shrubs (5.66%). There are no obvious geographical differences in the area percentages of artificial surfaces (i.e., buildings, roads, and OISAs). Cities in the north and west (including N, NE, NW, and SW regions) generally have a lower proportion of water areas, while some cities in the south and east (including E and S regions) have a relatively higher proportion. A similar trend is also observed in the central region, where Changsha and Wuhan in the south have larger water areas, while Zhengzhou has a relatively smaller proportion. In addition, the proportion of green spaces (including grass/shrubs and trees) is relatively high in the E, S, and SW regions, exceeding 10% in most cities, except for Shanghai and Suzhou. This may be related to the generally more suitable climate for plant growth in these areas. Bare soil tends to be more common in N and E, and less so in W and S.

## Discussion

### Performance of each stage of IEL

In this section, we introduced the quantitative results obtained from the intermediate processes of IEL, aiming to illustrate the value of each stage. First, the performance of $\text{Label\_Map}_0$ and $\text{Label\_Map}_1$ is shown in Table 5.

Overall, both $\text{Label\_Map}_0$ and $\text{Label\_Map}_1$ achieved good OA performance. The excellent performance of $\text{Label\_Map}_0$ demonstrates the potential of using existing products to provide labels, which is also the main source of initial labels for the data annotation engine we proposed. In comparison, the results of $\text{Label\_Map}_1$ show that the supplementation through semantic segmentation models can also bring additional benefits to the labels. Especially for grass/shrubs, the PA in $\text{Label\_Map}_0$ was only 32.32%, due to the lack of suitable grassland or

**Table 4.** Mapping accuracy based on validation samples for EcoVision over the 42 cities

| Accuracy | OISAs | Grass/shrubs | Trees | Soil | Buildings | Water | Roads | Agriculture |
|---|---|---|---|---|---|---|---|---|
| PA (%) | 76.29 | 55.09 | 80.09 | 90.52 | 92.30 | 88.31 | 90.94 | 90.20 |
| UA (%) | 86.53 | 90.12 | 88.93 | 68.53 | 87.82 | 82.54 | 76.41 | 80.50 |
| F1-score (%) | 81.09 | 68.38 | 84.28 | 78.00 | 90.00 | 85.33 | 83.04 | 87.79 |

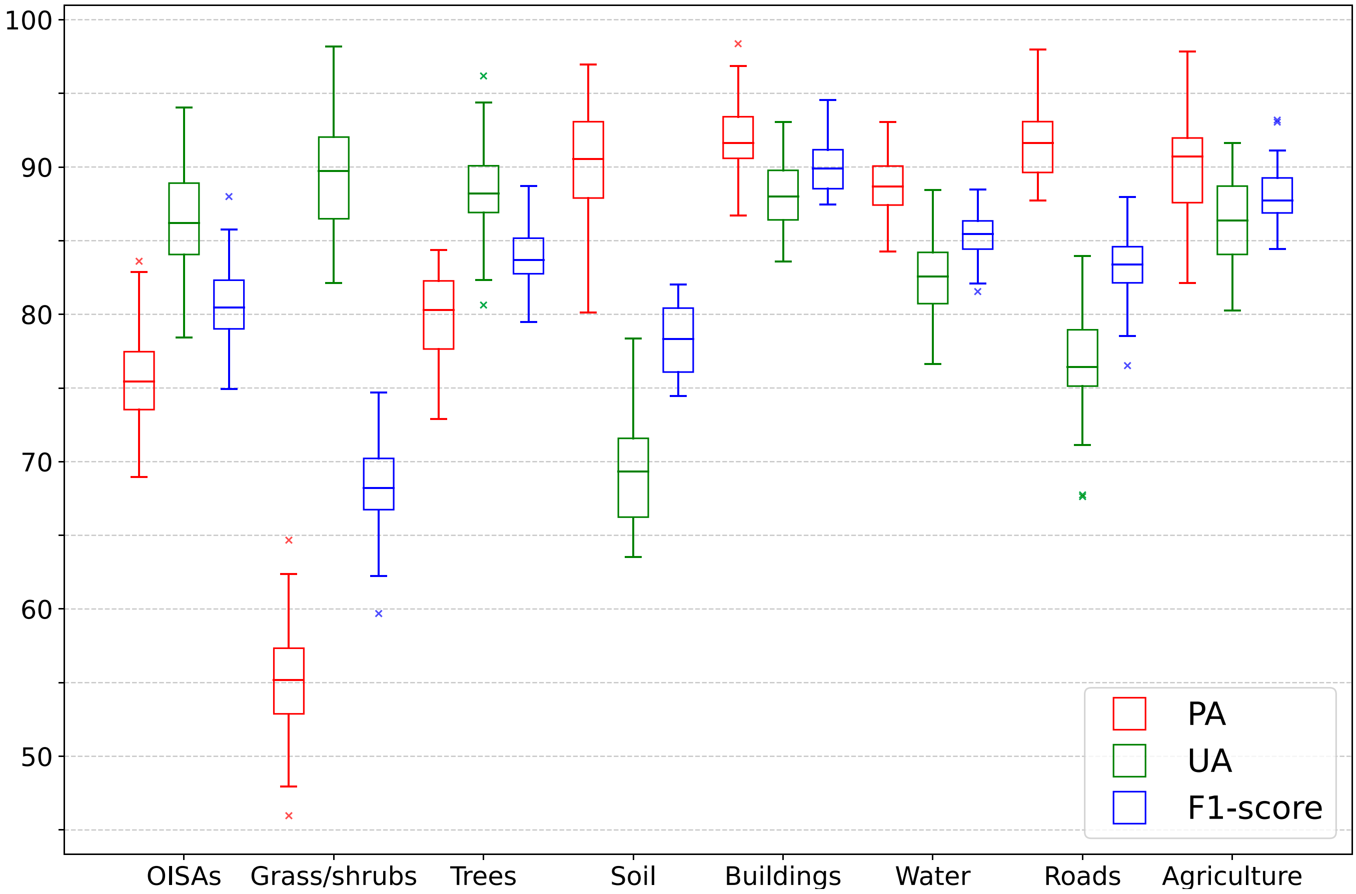

**Fig. 6.** Statistical results of EcoVision's user's accuracy, producer's accuracy, and F1-scores in 42 cities.

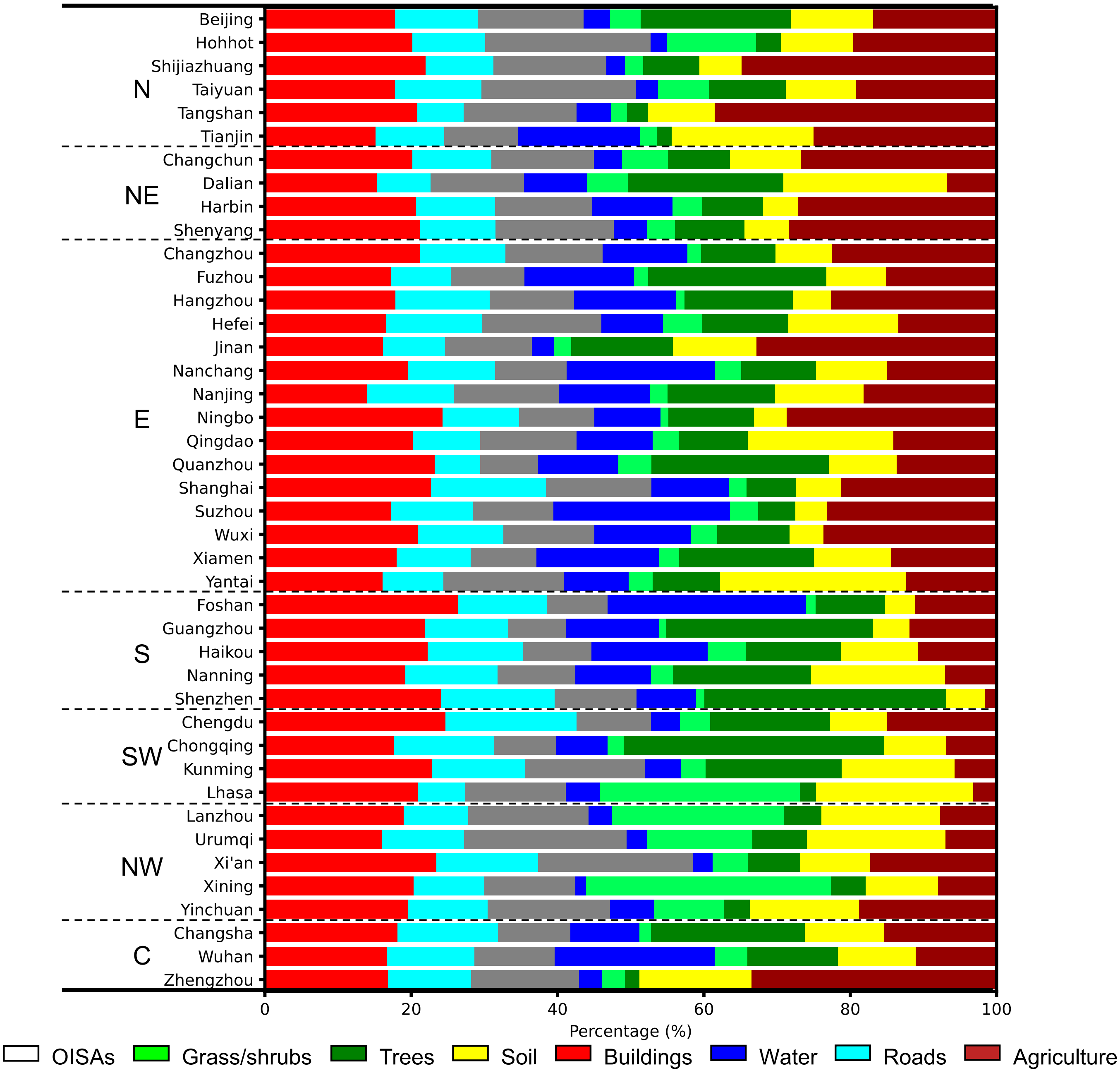

**Fig. 7.** The LULC composition of 42 major Chinese cities in EcoVision, categorized by geographical region.

shrub products currently available. Through the model's supplementation, its PA increased to 45.31%. It should be noted that, as previously introduced, both Label_Map$_0$ and Label_Map$_1$ include unlabeled pixels. To estimate the proportion of trusted pixels for the 2 labels within the entire dataset, we systematically counted their numbers and proportions in the validation samples. As shown in Table 5, the "Ratio" column reflects that the count of trusted pixels exhibits a progressive increase through iterative optimization.

This study incorporates the relatively low-resolution product GISA (10 m), and its positive role in EcoVision is elaborated as follows. On the one hand, high-resolution impervious surface products are scarce. As one of the highest-resolution large-scale relevant thematic products, the GISA (10 m) product serves as a vital reference for the OISAs category, and avoids uncertainty from relying on a single data source. As shown in Table 2, only 2 raster products (GISA and Hi-ULCM) provide data for the OISAs category. On the other hand, by assigning a low weight to the GISA product and placing the OISAs category in a later order during the voting process, the data engine can effectively reduce classification errors. Figure 8 visually demonstrates the positive role of the GISA product. The first row shows how our engine uses GISA to automatically correct the classification errors in the Hi-ULCM product for the top-left part, hence preventing error accumulation. The second row illustrates that even with resolution differences, the use of low-resolution GISA product in the high-resolution land cover classification does not lead to coarse-grained misclassifications in the final product.

**Table 5.** Mapping accuracy and number of trusted pixels based on validation samples for Label_Map$_0$ and Label_Map$_1$

| Label | Accuracy | OISAs | Grass/ shrubs | Trees | Soil | Buildings | Water | Roads | Agriculture | OA (%) | Ratio (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Label_Map0 | PA (%) | 76.93 | 32.32 | 88.08 | 91.23 | 93.12 | 91.49 | 86.98 | 87.82 | 88.28 | 45.43 |
| | UA (%) | 88.82 | 56.83 | 89.13 | 78.55 | 91.36 | 90.01 | 89.21 | 88.38 | | |
| | F1-score (%) | 82.45 | 41.21 | 88.60 | 84.42 | 92.23 | 90.74 | 88.08 | 88.10 | | |
| Label_Map1 | PA (%) | 74.25 | 45.31 | 82.41 | 94.32 | 93.88 | 92.58 | 92.28 | 93.88 | 87.59 | 79.12 |
| | UA (%) | 90.88 | 89.03 | 88.56 | 79.22 | 92.84 | 89.18 | 83.69 | 86.02 | | |
| | F1-score (%) | 81.73 | 60.06 | 85.37 | 86.11 | 93.36 | 90.85 | 87.78 | 89.78 | | |

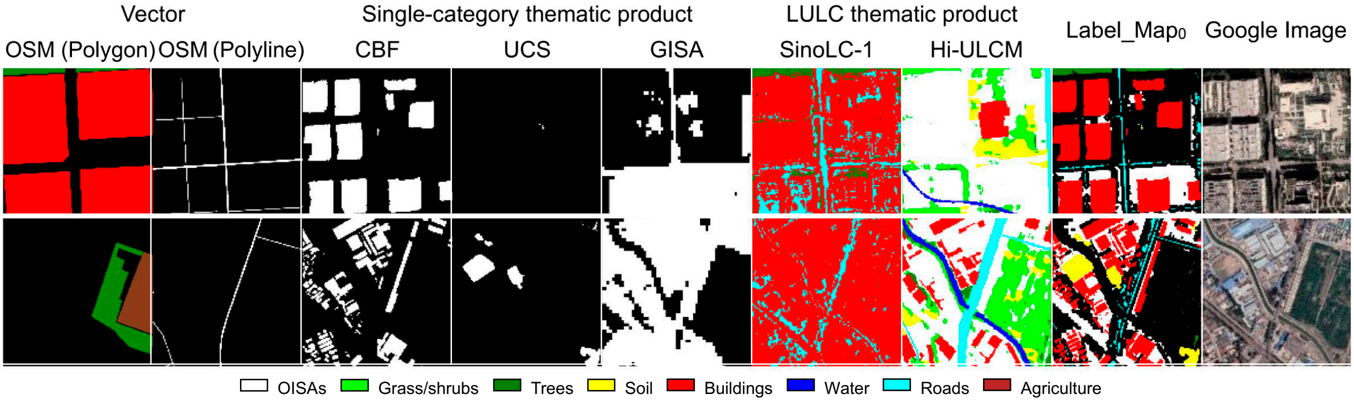

**Fig. 8.** The data engine IEL can effectively utilize low-resolution GISA and provides valuable support for producing EcoVision.

In this study, we proposed an iterative expansion method aimed at fully exploiting the advantages of various products and semantic segmentation models. To verify the effectiveness of this method and determine the best number of iterations, we conducted experiments in 42 cities. Using the same Google imagery data, we trained the model with labels generated from different numbers of iterations and validated each set of generated labels. With the increase number of iterations, the OA of the model gradually increased and reached the highest point. Specifically, the first iteration brought the greatest improvement, increasing the accuracy from 80.12% to 82.26%. In subsequent iterations, the accuracy further improved to 83.13% and 83.60%, with the increments becoming progressively smaller. After 4 iterations, the model's accuracy stabilized at 83.79%.

Figure 9 displays several Google images of different styles and the visual results of the model outputs during each iteration, further illustrating the benefits brought by iterative expansion. Specifically, roads vary in size (e.g., small countryside pathways versus big highways), and the surface material could be asphalt, concrete, or gravel. In urban areas, they are easily obscured by tall buildings, resulting in large areas of shadow and further presenting different styles. The initial labels had low diversity, which made the model somewhat inadequate in dealing with road extraction with high intra-class differences, leading to a considerable omission of roads (Fig. 9A and B). As the number of iterations increased and the labels became more diverse, the model's performance in road extraction showed a stable and marked improvement trend. Similar to roads, OISAs also have different styles, and in the visible spectral information, they are very similar to the exposed rock surfaces in the soil category, which is why OISAs and soil are easily confused in the initial iterations (Fig. 9B). Similar situations also occur with farmland and muddy water bodies both appearing brownish (Fig. 9C), dark water bodies and dense forests, as well as shadows caused by tall buildings (Fig. 9A and E). However, the distinction among these features showed a stable improvement trend with iteration, thanks to our proposed iterative expansion, which enriched the label and further increased intra-class diversity. Meanwhile, to filter out inappropriate pixels and avoid the contamination of the label by the model's incorrect output (such as the phenomenon of large areas of water or forest appearing in urban commercial districts during initial iterations), the role of other prompts should not be overlooked. They can effectively filter out unreliable information and prevent cumulative errors in the model. It is worth noting that we were pleasantly surprised to find that the advantage of iteration is also quite evident for the agriculture extraction. Unlike the other LULC classes, the agriculture category has fewer prompts, only from SinoLC-1 and OSM_Polygon. However, the amount of agriculture in OSM_Polygon is relatively small, and the accuracy of agriculture in SinoLC-1 is relatively low (reportedly, its PA is 71.07% and UA is 67.86%). Therefore, directly introducing them as labels would inevitably

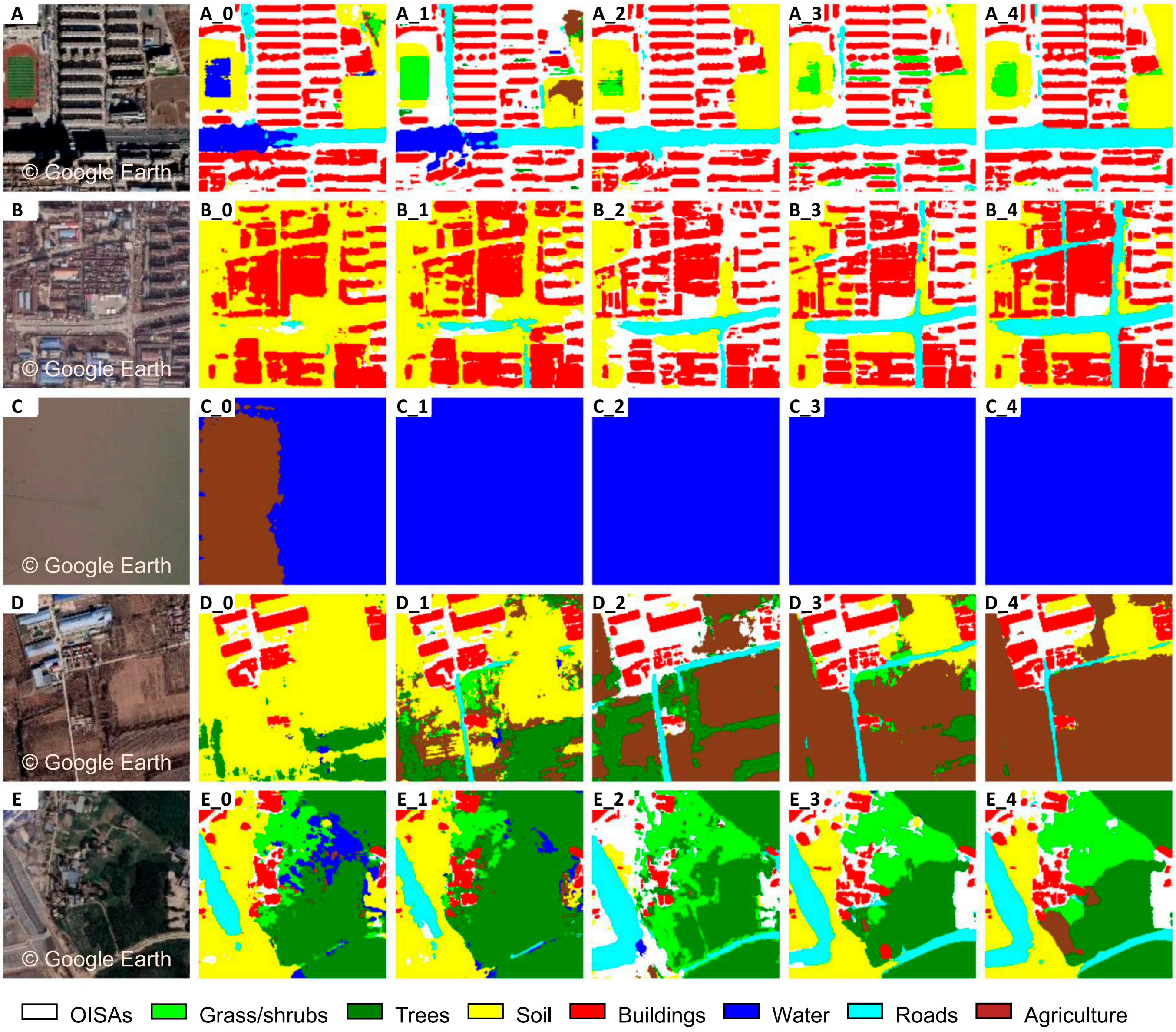

**Fig. 9.** Various styles of Google imagery (e.g., "A", "B", and "C") and their corresponding iterative results (e.g., "A_1" signifies the results of the model trained with Label_map$_0$), demonstrating the advantages gained through the iterative process.

affect the final agriculture extraction. However, through iteration, we overcame the limitations of the original products and filtered out unreliable and uncertain agricultural pixels from SinoLC-1, which gradually improved the model's ability to extract agricultural land (Fig. 9D and E). As for the building category, there was not much change throughout the iteration, thanks to the high-quality China's building footprint product CBF.

## Comparison with state-of-the-art LULC datasets

In this section, we compared the EcoVision product with 2 high-resolution LULC products and 3 medium-resolution LULC products, in terms of both visualization and quantitative results. The spatial resolution, LULC types, and their corresponding relationships are shown in Table 6. Figure 10 displays a visual comparison of the 6 products in several local areas. Despite the differences in spatial resolution or geographical coverage among these products, through comparative analysis, we can still intuitively feel that EcoVision, with its higher-resolution imagery and high-quality labels generated by the data annotation engine IEL, is able to capture more detailed urban spatial features and more accurately distinguish various LULC classes in heterogeneous areas.

The city center is covered by a large amount of impervious surface. In Fig. 10, this LULC type is represented rather roughly in the 3 medium-resolution products, only able to be roughly identified as impervious surfaces or buildings, and it is difficult to further subdivide them into roads, buildings, and OISAs (Fig. 10). Compared with Hi-ULCM, which is also a high-resolution product, EcoVision has a higher spatial resolution and more detailed classification, especially with the addition of the agriculture category, thus providing more comprehensive urban ecological information. In the Hi-ULCM product, the data for water, buildings, and roads are directly sourced from OSM and

**Table 6.** Category relations among FROM_GLC10, ESA_WorldCover, ESRI_2020, Hi-ULCM, SinoLC-1, and EcoVision

| Product | FROM_GLC10 (10 m) | ESRI_GLC10 (10 m) | ESA_GLC10 (10 m) | Hi-ULCM (2 m) | SinoLC-1 (1 m) | EcoVision (0.5 m) |
|---|---|---|---|---|---|---|
| LULC types | Impervious area | Built area | Built-up | OISAs | Building | OISAs |
| | | | | Buildings | | Buildings |
| | | | | Roads | Traffic route | Roads |
| | Forest | Trees | Tree cover | Trees | Tree cover | Trees |
| | Shrubland | Scrub | Shrubland | Grass/shrubs | Shrubland | Grass/shrubs |
| | Grassland | Grass | Grassland | | Grassland | |
| | Bare land | Bare | Barren/sparse vegetation | Soil | Barren and sparse vegetation | Soil |
| | Water body | Water | Permanent water bodies | Water | Water | Water |
| | Cropland | Crops | Cropland | | Cropland | Agriculture |

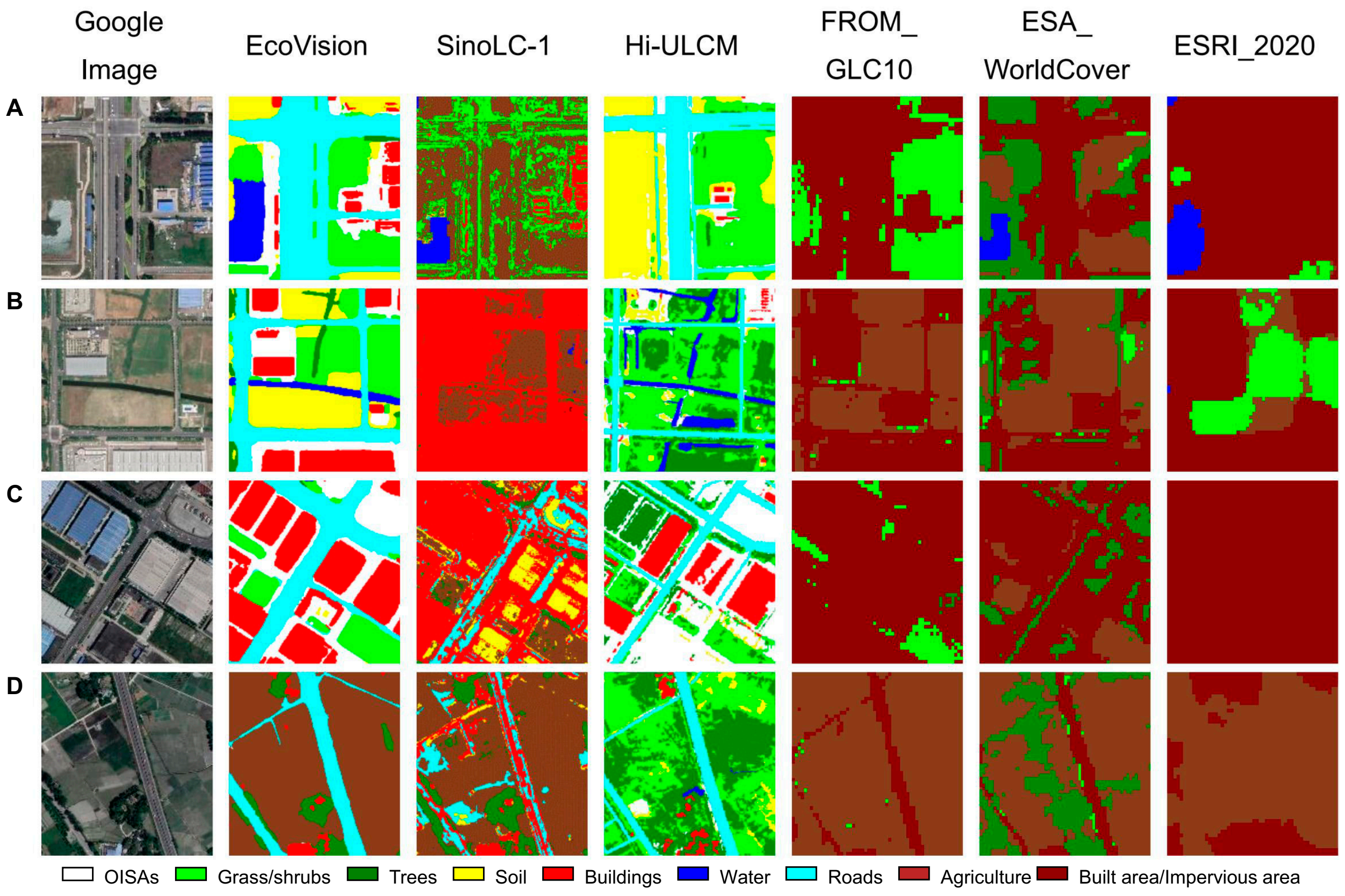

**Fig. 10.** The visual comparison demonstrations of 6 products across 4 local regions show that EcoVision exhibits the best performance. (A) to (D) show Google images and corresponding land cover maps from different products.

China's web map service providers A-map and Map World, rather than being extracted from imagery. Although this method reduces the cost of data acquisition, its reference time is unreliable, and the data quality still needs further verification [49]. For example, there are omissions of water in Fig. 10A, buildings in Fig. 10C, and roads in Fig. 10D. Compared with the SinoLC-1 product, which only contains building and road information, EcoVision further supplements other impervious surface information in urban areas and demonstrates superior performance in the extraction of building footprint information, by accurately identifying and extracting individual buildings. When addressing the issue of shadow caused by high-rise buildings in city

**Table 7.** Comparison of mapping accuracy based on test samples for SinoLC-1, Hi-ULCM, ESRI_2020, ESA_WorldCover, FROM_GLC10, and EcoVision

| LULC Type | Accuracy | OISAs | Buildings | Roads | Grass/shrubs | Soil | Trees | Water | Agriculture | OA (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| EcoVision | PA (%) | 76.29 | 55.09 | 80.09 | 90.52 | 92.30 | 88.31 | 90.94 | 90.20 | 83.63 |
| | UA (%) | 86.53 | 90.12 | 88.93 | 68.53 | 87.82 | 82.54 | 76.41 | 80.50 | |
| | F1-score (%) | 81.09 | 68.38 | 84.28 | 78.00 | 90.00 | 85.33 | 83.04 | 87.79 | |
| SinoLC-1 | PA (%) | 69.34 | | 43.83 | 22.84 | 34.58 | 40.11 | 63.84 | 70.83 | 51.13 |
| | UA (%) | 64.17 | | 60.88 | 32.35 | 71.66 | 46.28 | 86.25 | 36.16 | |
| | F1-score (%) | 66.05 | | 50.97 | 26.78 | 46.95 | 42.97 | 73.37 | 47.88 | |
| Hi-ULCM | PA (%) | 75.24 | 78.26 | 76.88 | 53.35 | 51.12 | 74.48 | 80.46 | | 72.27 |
| | UA (%) | 66.87 | 80.12 | 79.62 | 69.64 | 53.83 | 78.32 | 81.83 | | |
| | F1-score (%) | 70.81 | 79.18 | 78.23 | 60.42 | 52.46 | 76.35 | 81.14 | | |
| ESRI_2020 | PA (%) | | 88.32 | | 32.12 | 20.12 | 32.51 | 58.17 | 78.10 | 46.78 |
| | UA (%) | | 61.13 | | 29.33 | 34.69 | 68.13 | 71.93 | 32.84 | |
| | F1-score (%) | | 72.25 | | 30.66 | 25.47 | 44.02 | 64.32 | 46.24 | |
| ESA_WorldCover | PA (%) | | 73.10 | | 28.13 | 27.38 | 48.52 | 52.32 | 84.50 | 47.67 |
| | UA (%) | | 70.33 | | 36.52 | 36.81 | 73.56 | 88.31 | 36.18 | |
| | F1-score (%) | | 71.69 | | 31.78 | 31.40 | 58.47 | 65.71 | 50.67 | |
| FROM_GLC10 | PA (%) | | 70.64 | | 37.50 | 31.37 | 31.63 | 43.88 | 62.47 | 46.42 |
| | UA (%) | | 62.91 | | 28.66 | 29.36 | 66.90 | 89.17 | 29.64 | |
| | F1-score (%) | | 66.55 | | 32.49 | 42.95 | 58.32 | 58.82 | 40.20 | |

centers, the SinoLC-1 product often misclassifies shadows as other categories, resulting in phenomena such as road discontinuity and the appearance of large areas of natural vegetation in city centers, which do not match actual observations. In contrast, EcoVision can accurately reveal the linear morphology of roads, overcoming the interference of roadside vegetation and high-rise buildings on road extraction.

Based on the LULC mapping relationships in Table 6 and using the validation samples, we further quantitatively compared EcoVision with 5 LULC products, with the results presented in Table 7. Overall, EcoVision achieves the best result in terms of OA with 83.63%, followed by Hi-ULCM (72.27%), SinoLC-1 (51.13%), ESA_WorldCover (47.67%), ESRI_2020 (46.78%), and FROM_GLC10 (46.42%). In terms of the F1-score for each category, EcoVision also shows the best performance. From this, we have achieved the goal of generating an LULC product with a wider coverage area, higher resolution, and richer details based on Hi-ULCM.

In summary, compared to the current state-of-the-art LULC products over a large area, EcoVision, with its highest spatial resolution, richer detail representation, superior accuracy, and more detailed category classification, provides more accurate and comprehensive data support for research and applications in fields related to urban spatial analysis.

## Conclusion

The contributions of this study lie in both methodology and product aspects. On the one hand, we proposed a 2-stage data annotation engine called IEL. In the first stage, it generates a small number of labels through priority-based weighted voting, while in the second stage, it iteratively extends these labels using semantic segmentation models. This method can overcome challenges, such as data format and resolution differences, inconsistent classification systems, and label conflicts between different products, avoid tedious manual pixel annotation, and effectively alleviate the challenge of label scarcity in large-scale high-resolution LULC mapping.

Furthermore, we have created China's first large-scale submeter-level LULC product, EcoVision, using labels generated by IEL. The spatial resolution of this product is 0.5 m, covering 42 major cities in China. The product was validated with a total of 23,850,000 randomly selected pixels from the 42 cities, achieving an OA of 83.63%, providing high-quality reference for future related research. The superiority of EcoVision was also verified by comparing it with 5 other state-of-the-art LULC products, and it outperformed the only large-scale LULC product in China (SinoLC-1) in terms of OA, with higher resolution and richer details. The EcoVision dataset has been publicly released: https://doi.org/10.5281/zenodo.14921585.

## Acknowledgments

We gratefully acknowledge the free access of the SinoLC-1 products provided by Wuhan University, the ESA_WorldCover products provided by the European Space Agency, the ESRI_2020 products provided by ESRI, Inc. and IO, Inc., the FROM_GLC10 products provided by Tsinghua University, the LULC information provided by the OSM, and the VHR Google Earth images provided by Google Inc.

**Funding:** This work was supported by the National Key Research and Development Program of China (grant number

2024YFF1306102) and the National Natural Science Foundation of China (grant numbers 42271328 and 42471391).

**Author contributions:** All authors contributed to the writing of the manuscript. X.H. conceived the study. E.Z. designed the methodology. J.Z. provided data. E.Z. programmed the framework codes. E.Z. prepared the original draft. X.H., J.L., and J.Z. revised the draft.

**Competing interests:** The authors declare that they have no competing interests.

## Data Availability

The EcoVision product generated in this paper and corresponding user guidelines are available from https://doi.org/10.5281/zenodo.14921585. The product consists of 42 city tiles in the GeoTIFF format, with each city tile named "C.tif," where "C" explains the city name. For example, the 0.5 m LULC map of Wuhan is named "wuhan.tif". In addition, each tile contains an LULC label band ranging from 0 to 255, where the corresponding relationship between this value and the LULC type is shown in Fig. 5.

## References

1. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*. PMLR; 2020. p. 1597–1607.
2. Liu M, Zhang P, Shi Q, Liu M. An adversarial domain adaptation framework with KL-constraint for remote sensing land cover classification. *IEEE Geosci Remote Sens Lett*. 2022;19:1–5.
3. Zhang HK, Roy DP. Using the 500 m MODIS land cover product to derive a consistent continental scale 30 m Landsat land cover classification. *Remote Sens Environ*. 2017;197:15–34.
4. Tong X-Y, Xia G-S, Zhu XX. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS J Photogramm Remote Sens*. 2023;196:178–196.
5. Yuan Q, Shen H, Li T, Li Z, Li S, Jiang Y, Xu H, Tan W, Yang Q, Wang J, et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens Environ*. 2020;241:Article 111716.
6. Friedl MA, Brodley CE. Decision tree classification of land cover from remotely sensed data. *Remote Sens Environ*. 1997;61(3):399–409.
7. Bruzzone L, Marconcini M. Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. *IEEE Trans Geosci Remote Sens*. 2009;47(4):1108–1122.
8. Chan JC-W, Paelinckx D. Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sens Environ*. 2008;112(6):2999–3011.
9. Li Z, Zhang H, Lu F, Xue R, Yang G, Zhang L. Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. *ISPRS J Photogramm Remote Sens*. 2022;192:244–267.
10. Kussul N, Lavreniuk M, Skakun S, Shelestov A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci Remote Sens Lett*. 2017;14(5):778–782.
11. Li Y, Zhang H, Xue X, Jiang Y, Shen Q. Deep learning for remote sensing image classification: A survey. *Wiley Interdiscip Rev: Data Min Knowl Discovery*. 2018:8.
12. Zeiler MD, Fergus R. Visualizing and Understanding Convolutional Networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors. *Computer Vision – ECCV 2014*. Cham: Springer International Publishing; 2014. p. 818–833.
13. Zhu XX, Tuia D, Mou L, Xia G-S, Zhang L, Xu F, Fraundorfer F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci Remote Sens Mag*. 2017;5(4):8–36.
14. Lang N, Jetz W, Schindler K, Wegner JD. A high-resolution canopy height model of the earth. *Nat Ecol Evol*. 2023;7(11):1778–1789.
15. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. PMLR; 2021. p. 8748–8763.
16. Guzder-Williams B, Mackres E, Angel S, Blei AM, Lamson-Hall P. Intra-urban land use maps for a global sample of cities from Sentinel-2 satellite imagery and computer vision. *Comput Environ Urban Syst*. 2023;100:Article 101917.
17. Cui H, Zhang G, Chen Y, Li X, Hou S, Li H, Ma X, Guan N, Tang X. Knowledge evolution learning: A cost-free weakly supervised semantic segmentation framework for high-resolution land cover classification. *ISPRS J Photogramm Remote Sens*. 2024;207:74–91.
18. Li Y-F, Guo L-Z, Zhou Z-H. Towards safe weakly supervised learning. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(1):334–346.
19. Zhou Z-H. A brief introduction to weakly supervised learning. *Natl Sci Rev*. 2018;5(1):44–53.
20. Tong X-Y, Dong R, Zhu XX. Global high categorical resolution land cover mapping via weak supervision. *ISPRS J Photogramm Remote Sens*. 2025;220:535–549.
21. Chen K, Liu C, Chen H, Zhang H, Li W, Zou Z, Shi Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans Geosci Remote Sens*. 2024;62:1–17.
22. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P, Girshick R. Segment anything. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2023. p. 3992–4003.
23. Chen K, Zhang J, Liu C, Zou Z, Shi Z. RSRefSeg: Referring remote sensing image segmentation with foundation models. arXiv. 2025. https://doi.org/10.48550/arXiv.2501.06809
24. Brown CF, Brumby SP, Guzder-Williams B, Birch T, Hyde SB, Mazzariello J, Czerwinski W, Pasquarella VJ, Haertel R, Ilyushchenko S, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Sci Data*. 2022;9(1):Article 251.
25. Li Z, He W, Cheng M, Hu J, Yang G, Zhang H. 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 Syst Sci Data*. 2023;15:4749–4780.
26. Liu Y, Zhong Y, Ma A, Zhao J, Zhang L. Cross-resolution national-scale land-cover mapping based on noisy label learning: A case study of China. *Int J Appl Earth Obs Geoinf*. 2023;118:Article 103265.
27. Zhou Y, Weng Q. Building up a data engine for global urban mapping. *Remote Sens Environ*. 2024;311:Article 114242.
28. Huang X, Wang Y, Li J, Chang X, Cao Y, Xie J, Gong J. High-resolution urban land-cover mapping and landscape analysis

of the 42 major cities in China using ZY-3 satellite images. *Sci Bull*. 2020;65(12):1039–1048.

29. Zhang Y, Chen G, Myint SW, Zhou Y, Hay GJ, Vukomanovic J, Meentemeyer RK. UrbanWatch: A 1-meter resolution land cover and land use database for 22 major cities in the United States. *Remote Sens Environ*. 2022;278:Article 113106.
30. Li J, Liu T, Yang J, Jiang J, Huang X. Mapping of 30m global urban boundaries from 1972 to 2021 based on multi-source geographic information fusion. *Int J Appl Earth Observ Geoinform*. 2025.
31. Zhou D, Zhao S, Zhang L, Sun G, Liu Y. The footprint of urban heat island effect in China. *Sci Rep*. 2015;5(1):11160.
32. Conn B, Arandjelović O. Towards computer vision based ancient coin recognition in the wild—Automatic reliable image preprocessing and normalization. In: *2017 International Joint Conference on Neural Networks. (IJCNN)*. IEEE; 2017. p. 1457–1464.
33. Huang X, Zhang Z, Li J. China's first sub-meter building footprints derived by deep learning. *Remote Sens Environ*. 2024;311:Article 114274.
34. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada DE, Fernández-Luna JM, editors. *Advances in information retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. Vol. 3408, p. 345–359.
35. Li J, Wang W, Huang X, Pan X, Ma S, Chen B, Zhang L. China's new-type urbanization strategies reflected by nationwide sub-meter mapping of construction sites. *Nat Sci Rev*. 2025.
36. Huang X, Yang J, Wang W, Liu Z. Mapping 10 m global impervious surface area (GISA-10m) using multi-source geospatial data. *Earth Syst Sci Data*. 2022;14(1):3649–3672.
37. Herfort B, Lautenbach S, Porto de Albuquerque J, Anderson J, Zipf A. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. *Nat Commun*. 2023;14(1):Article 3985.
38. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. Language models are few-shot learners. In: *Advances in neural information processing systems*. Curran Associates, Inc.; 2020. Vol. 33, p. 1877–901.
39. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing; 2015. p. 234–241.
40. Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, et al. Deep high-resolution representation learning for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. 2021;43(10):3349–3364.
41. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2017. p. 2881–2890.
42. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European conference on computer vision (ECCV)*. Springer (Berlin): Germany; 2018. p. 801–818.
43. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv. 2020. https://doi.org/10.48550/arXiv.2010.11929
44. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021. p. 10012–10022.
45. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: *Advances in neural information processing systems*. Curran Associates, Inc.; 2021. Vol. 34, p. 12077–12090.
46. Karra K, Kontgis C, Statman-Weil Z, Mazzariello JC, Mathis M, Brumby SP. Global land use / land cover with sentinel 2 and deep learning. *IEEE Int Geosci Remote Sens Symp IGARSS*. 2021;2021:4704–4707.
47. Van De Kerchove R, Zanaga D, Keersmaecker W, Souverijns N, Wevers J, Brockmann C, Grosu A, Paccini A, Cartus O, et al. ESA WorldCover: Global land cover mapping at 10 m resolution for 2020 based on Sentinel-1 and 2 data. 2021; 2021:GC45I-0915.
48. Gong P, Liu H, Zhang M, Li C, Wang J, Huang H, Clinton N, Ji L, Li W, Bai Y, et al. 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*. 2019;64(6):370–373.
49. Saralioglu E, Gungor O. Crowdsourcing in remote sensing: A review of applications and future directions. *IEEE Geosci Remote Sens Mag*. 2020;8(4):89–110.