Remote Sensing of Environment 331 (2025) 115034

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Strategic sampling for training a semantic segmentation model in operational mapping: Case studies on cropland parcel extraction

Rui Lu [a,b], Ronghua Liao [a,b], Ran Meng [c,d], Yingchu Hu [a,b], Yi Zhao [b], Yan Guo [e], Yingfan Zhang [a,b], Zhou Shi [a,b], Su Ye [a,b,*]

[a] State Key Laboratory of Soil Pollution Control and Safety, Zhejiang University, Hangzhou 310058, China
[b] Institute of Agricultural Remote Sensing and Information Technology Application, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] Artificial Intelligence Research Institute, Faculty of Computing, Harbin Institute of Technology, Harbin 150008, China
[d] National Key Laboratory of Smart Farm Technologies and Systems, Harbin 150008, China
[e] Institute of Agricultural Information Technology, Henan Academy of Agricultural Sciences, Zhengzhou 450002, China

ARTICLE INFO

Editor: Marie Weiss

Keywords:
Semantic segmentation
Deep learning
Training samples
Balanced sampling
Transfer strategy

ABSTRACT

Semantic segmentation of remotely sensed images has become increasingly popular for a wide range of natural resource and urban application, yielding promising results. To an operational semantic segmentation mapping project, having more samples generally enables the model to better extract target features, achieving higher accuracies. However, annotating remote sensing image samples for model training is a time-consuming and labor-intensive process. Strategic sampling aims to minimize the efforts in collecting new training samples for a mapping project, which has been not well studied yet for semantic segmentation. To approach this topic, we employed a hybrid way for combining meta-analysis and case studies to investigate the best practices for strategic sampling. Three factors relating to strategic sampling will be investigated: sample size, distribution and transferring methods. We first reviewed 334 recently published papers that adopted semantic segmentation for operational mapping projects to summarize the current status of training sample design from various mapping scenarios. Subsequently, we constructed a large dataset of over 12,000 high-quality annotated image patches for cropland parcel mapping across five study sites, and evaluated various sampling strategies using a baseline segmentation model. We also proposed a novel balanced sampling method, which leveraged patch-based entropy and edge complexity to classify sample diversity. Our findings revealed that (1) both meta-analysis and the case studies suggested that ~4 % of the total mapping patches were the optimal training sample size under random sampling, i.e., the minimum size to reach accuracy saturation; (2) compared to random sampling, the newly proposed balanced sampling was superior due to its decreasing the required sample size from ~4 % to 2.5 % of the total patches in mapped areas; (3) sample transfer and model transfer present identical performance for relaxing the average local sample demand from 2.5 % to 0.5 % of total patches, with sample transfer being slightly more accurate than model transfer (Global Total-Classification errors: 0.298 vs 0.308). This study offers a heuristic framework for applying strategic sampling in semantic segmentation, providing valuable practical guidance for implementing deep learning in an operational scenario.

1. Introduction

Semantic segmentation of remotely sensed images, i.e., assigning a category to each pixel in an image based upon state-of-art deep learning models, have become the leading approach for a wide range of surveying tasks including agricultural field extraction (Persello et al., 2019; Waldner and Diakogiannis, 2020), urban sprawl detection (Kim et al., 2024; Zhang et al., 2023), and land cover mapping (Li et al., 2020; Zhang et al., 2019). Benefited from the capability of learning intricate spatial and temporal patterns, semantic segmentation often surpasses traditional machine learning tasks particularly for mapping diverse and nuanced geographic features (Ma et al., 2019; Reichstein et al., 2019; Yuan et al., 2020; Persello et al., 2022). It is widely acknowledged that a larger training dataset could provide better exposure for various

* Corresponding author at: State Key Laboratory of Soil Pollution Control and Safety, Zhejiang University, Hangzhou 310058, China.
E-mail address: su.ye@zju.edu.cn (S. Ye).

https://doi.org/10.1016/j.rse.2025.115034
Received 19 May 2025; Received in revised form 10 September 2025; Accepted 14 September 2025

geographic patterns of different scenarios to reduce the chances of overfitting from outliers as well as to enhance the generalization to unseen data (Yu et al., 2018; Bergen et al., 2019; Yuan et al., 2021; Grift et al., 2024). However, unlike popular image label databases in the computer vision domain (e.g., ImageNet, COCO) (Deng et al., 2009; Lin et al., 2014), the training datasets are often not universally available for semantic segmentation of remotely sensed images, not only because local geographic patterns are commonly too complicated to be represented by a single dataset, but also owing to the rarity of a universal land category system (e.g., agricultural fields are defined differently among different regions) (Reichstein et al., 2019). As a result, remote sensing professionals are frequently required to manually generate as many new image patches as possible for mapping projects (Kattenborn et al., 2021), calling for considerable annotation efforts and budgets. It is essential for prioritizing a smaller, more representative training sample set, particularly when the mapping region is large and acquiring training data is expensive.

Strategic sampling is a group of techniques for minimizing the efforts in collecting new training samples while keeping the accuracy uncompromised (Brown, 2006). Traditionally, in pixel-based classification, previous discussions on strategic sampling surround sample size (Foody et al., 2006; Zhu et al., 2016; Ramezan et al., 2021; Rajput et al., 2023), and the distribution of samples across categories (Foody and Mathur, 2006; Jin et al., 2014; Mellor et al., 2015). Inadequate sample size would suffer from an underfitting of training data and fail to cover full geographic variability. Though increasing sample size generally could enhance the performance, the accuracy curve may cease to increase after a certain sample size (Heydari and Mountrakis, 2018). The ideal sample size is the minimum number of samples required to achieve a plateau in accuracy curves, balancing annotation costs with model performance. While some literature suggests using a sample size of 10–30 times the number of features for a classifier (Piper, 1992; Van Niel et al., 2005), it is more reasonable to determine the sample number by considering the total area to be mapped, as it reflects landscape variability and the complexity of the classification scene (Stehman and Wickham, 2011; Olofsson et al., 2014). As such, Zhu et al. (2016) tested the training pixel number from 2500 to 25,000 for land cover classification of a single Landsat scene based on a random-forest classifier, and reported that the accuracy stopped increasing when the training set exceeded 20,000 pixels, i.e., ~0.1 % of a Landsat tile. However, existing researches on optimal sample size have primarily focused on pixel-based samples. Different from pixel-based sampling scheme that traditional machine learning techniques mostly fall within, training samples for deep learning are mostly based upon small sub-images, i.e., image patches. For each image patch, the pixels are labeled as the classes of interest, and the number and shapes of land-cover classes are inherently variable. The different nature between image patches and pixel samples limits a direct knowledge transfer for sample selection from traditional machine learning to deep learning. In contrast to the relatively simpler model architecture of traditional classifiers (e.g., random forest, support vector machine), semantic segmentation enables a better generalization by stacking multiple convolutional and pooling layers with more model hyperparameters, thereby often requiring a larger sample size. Chen et al. (2024) suggested that in geographically similar regions, approximately one-tenth of the data might be sufficient to train a CNN model effectively. However, their study did not assess the impact of varying sample sizes on model accuracy, and whether one-tenth represented an optimal sample size remained unknown. To our knowledge, most previous semantic segmentation studies determined the training patch number based on subjective judgment or the annotation budget. It still lacks a heuristic to quickly estimate an economic sample size for a targeted mapping region without sensitivity tests.

The choice of sample distribution answers the question for how to assign the sample numbers to different classes. The literature consistently reports that balanced sampling (i.e., an equal number of samples per class) helps prevent model bias toward the majority class, compared to random or proportional sampling (i.e., assigning training samples in proportion to the area of each class), whereas it possibly suffers a lower overall accuracy (Colditz, 2015; Nguyen et al., 2020; Zhou et al., 2020). For example, Du et al. (2015) developed a voting-distribution ranked rule, which averaged multiple reliable voting distributions for each class to determine weights and increased the representation of minority classes. However, existing studies on balanced sampling methods are primarily based on pixel-level samples. In the context of semantic segmentation, there is no straightforward method for assigning sample numbers when using an image patch as the sampling unit, as it consists of multiple categories. This makes it challenging to directly apply balanced sampling to a semantic-segmentation task. As a result, professionals often resort to random sampling of image patches (Boschetti et al., 2016; Stehman et al., 2022), which can lead to poor performance for minority classes in imbalanced datasets and generate redundant samples for easily identifiable categories. Hence, there is a need for proposing a new balanced sampling strategy tailored for patch-based samples to guide sample selection for the semantic segmentation tasks.

Besides sampling design, another approach for strategic sampling is leveraging global samples (i.e., samples collected from the entire geographic region) to reduce the need for newly collected samples from a specific mapping region, i.e., local samples. Model transfer is such an approach that takes advantage of a pre-trained model from a large-scale dataset and fine-tunes it for a new targeted mapping region (Ma et al., 2024; Pan and Yang, 2010; Zhao, 2017; Zhu et al., 2021). Model transfer unfreezes the last layers of the base model and only adjusts the higher-order feature representation using the new local samples, thereby reducing the need for large amounts of training dataset. For example, Wieland et al. (2023) applied a model transfer approach, enabling a pre-trained model to identify water bodies in remote sensing data from different sources. Alternatively, it is also common for transferring local samples to a global training dataset and then building a completely new deep learning model from the scratch, i.e., sample transfer (Brown et al., 2020; Zhang et al., 2024). Sample transfer is more frequently applied in large-scale remote sensing mapping. For example, Brown et al. (2020) employed training data selected from a 3 by 3 tile window where the central tile was the tile to be classified. Zhang et al. (2024) also combined the local samples from the center tile and the global samples from its neighborhood tiles to build a classifier for the center tile. Compared to model transfer, sample transfer allows for adjusting the weights of all layers, offering greater flexibility to adapt to a new mapping region. However, it often comes with higher computational costs due to the need to train a new model from scratch. To our knowledge, no studies ever comparatively reported their performances and new local samples required for the two transfer methods.

Aiming to minimize the effort required to collect new training patches, this study will explore different strategic sampling configurations and identify the best training sample selection practices for semantic segmentation, using a hybrid approach of meta-analysis and case studies. In this study, a *sample* refers to an image patch used for training the semantic segmentation model (specifically, a 256 × 256-pixel image). *Sampling* denotes the process of selecting patches to construct the training dataset. In this study, the *sample size* is defined as the relative training patch number, which is measured as the patches selected for training divided by the total patches to be mapped. This measurement helps eliminating the influence of varying image sizes and differences in the spatial resolution of remote sensing data. Three key factors in the training sample selection process were primarily investigated: sample size, sample distribution, and transfer methods. To address various semantic segmentation cases, we first performed a meta-analysis on 334 journal articles between 2015 and 2024, covering a broad range of mapping themes. This analysis summarized the general status on sample size, sample distribution, and transferring methods for semantic segmentation. Second, through five case studies of cropland parcel mapping, we created a large open-access training dataset consisted of over 12,000 high-quality labeled patches, and then evaluated

different sampling practices concerning sample size, distribution and transferring methods based upon a baseline semantic segmentation model. Cropland parcel mapping is often considered as a challenging task owing to their complex parcel forms, diversified spectral characteristics, and incontiguous boundaries intergraded with surrounding vegetation (Masoud et al., 2020; Estes et al., 2022; Lu et al., 2024), influenced by factors such as crop type, growth stage, soil conditions, and environmental variables. This makes cropland parcel mapping an ideal application for testing the performance of different training sample selection strategies for deep learning. In the past studies, various sample configurations have been adopted to deploy semantic segmentation for extracting parcel boundaries (Waldner and Diakogiannis, 2020; Turkoglu et al., 2021; Cai et al., 2023; Pan et al., 2023), which echoed the lack of standardized criteria for selecting training samples. To compare the effectiveness of different sampling practices, we will analyze the point at which the error curve first flattens with increasing numbers of input samples. An optimal method should minimize the number of training samples needed without compromising model performance. By combining meta-analysis and case studies, this work aims to provide a heuristic solution for strategic sampling that achieves the most cost-effective training sample selection. More specifically, we will answer the three following questions:

(1) How many training samples in minimum are required when the model accuracy saturates for semantic segmentation?
(2) What is the most effective way for distributing training samples in the form of image patches?
(3) What is comparative performance for model and sample transfer in term of mitigating local sample demand?

2. Datasets and study sites

2.1. Meta-analysis data collection

Given by a vast body of current literature on semantic segmentation, we focused on three top remote sensing journals, i.e., *Remote Sensing of Environment* (RSE), *ISPRS Journal of Photogrammetry and Remote Sensing* (ISPRS P&RS), *International Journal of Applied Earth Observation and Geoinformation* (JAG), and two data-based journals, i.e., *Scientific Data* (SD) and *Earth System Science Data* (ESSD). We did not include the technical journals such as *IEEE Transactions on Geoscience and Remote Sensing* (TGRS), as their articles primarily focused on algorithm innovation rather than land-cover mapping. Their evaluated algorithms were based on benchmark labeling datasets (e.g., Semantic Labeling Challenge datasets (Pastorino et al., 2022)), making it difficult to infer information about the total patches to be mapped. Only studies that provided the final segmentation map for their entire study area were considered for sample size demand, calculated as the proportion of the training sample patches relative to the total patches in the mapping extent, namely training patch proportion. We performed an initial search in the Science Direct (https://www.sciencedirect.com/) or Web of Science (https://webofscience.clarivate.cn/) database for each journal using the keyword "semantic segmentation" with a year range specification of "2015–2024" and an exclusion of "Review articles" (search date: Dec. 10, 2024). As a result, we obtained 334 journal articles (RSE: 27; ISPRS P&RS: 168; JAG: 117; SD: 19; ESSD: 3, see Fig. 1). The complete list of the selected articles was given in *Section S1* of the *Supplementary Material*. The year-over-year increase in publications highlighted the growing prominence of remote sensing semantic segmentation as a research focus (Fig. 1).

2.2. Case study sites and data

We selected cropland parcel mapping as a case study to address the sample selection problem in remote sensing semantic segmentation tasks. Since vector parcel boundaries could be derived from raster outputs through post-processing, our focus was on the accuracy of the raster-based cropland parcel representations generated by the model.

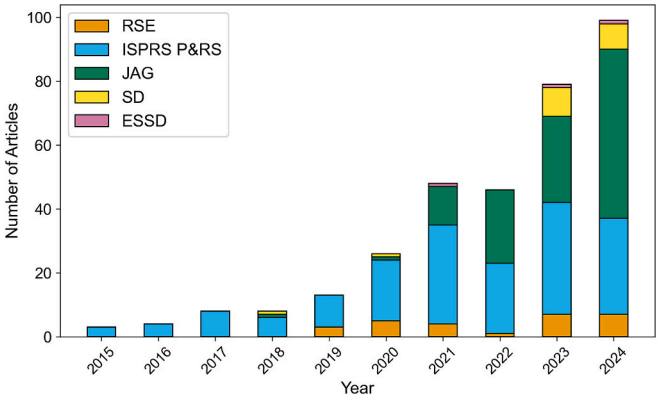

Fig. 1. Year of journals for the 334 referred articles for the meta-analysis.

2.2.1. Study sites

We selected five typical study sites in China (Fig. 2), including Xinjiang (XJ), Jilin (JL), Guangxi (GX), Hubei (HB), and Zhejiang (ZJ). Table 1 summarizes the diverse characteristics of the agricultural systems across the study sites, which represent a range of scenarios for extracting cropland parcels from high-resolution remote sensing images. These study sites were selected to capture the diversity of cropland systems, ranging from large-scale mechanized fields to smallholder-based parcels, to demonstrate the generalizability of the proposed sample selection methods.

2.2.2. Satellite images

For each study area, we selected a scene of Gaofen-2 (GF-2) high-resolution imagery. The GF-2 satellite contains a panchromatic sensor with 1-m resolution and a multispectral sensor with 4-m resolution, covering a swath of 45 km with a revisiting cycle within 5 days (Chen et al., 2022; Tong et al., 2020). The GF-2 data can be retrieved at https://data.cresda.cn. All images were acquired in 2022, during the period between crop harvest and early growth. This timing facilitated their distinction from surrounding features such as grasslands, and also allowed field roads to be clearly visible (Cheng et al., 2020). The pre-processing steps of images included radiometric calibration, atmosphere correction, and geometric correction. We applied the Gram-Schmidt Pan-Sharpening method for image fusion by resampling multispectral images with a high-resolution panchromatic image, resulting in 1-m images with four multispectral bands, i.e., blue, green, red, and near-infrared.

2.2.3. Labeled patch dataset

We generated a large number of high-quality labeled image patches for testing sampling strategies. Each GF-2 image was divided into 256 × 256 pixel patches (Li et al., 2023; Waldner and Diakogiannis, 2020). From each study site, approximately 20 % of the patches were randomly selected, resulting in around 2500 patches per site and 12,500 patches in total. A team of experienced remote sensing experts manually annotated all crop field boundaries within each patch through visual interpretation, dedicating over 1000 working hours to the annotation process. The labeled patches were then evenly split into two pools: a training sample pool (10 % of total patches), used to generate various sampling configurations, and a testing sample pool (10 %), used to independently assess model accuracy across different sampling strategies.

We have released the labeled image patches (China's Crop Parcel Training Dataset, CCPTD) as a potential benchmark dataset for future study, which is publicly available at https://doi.org/10.5281/zenodo.16595511.

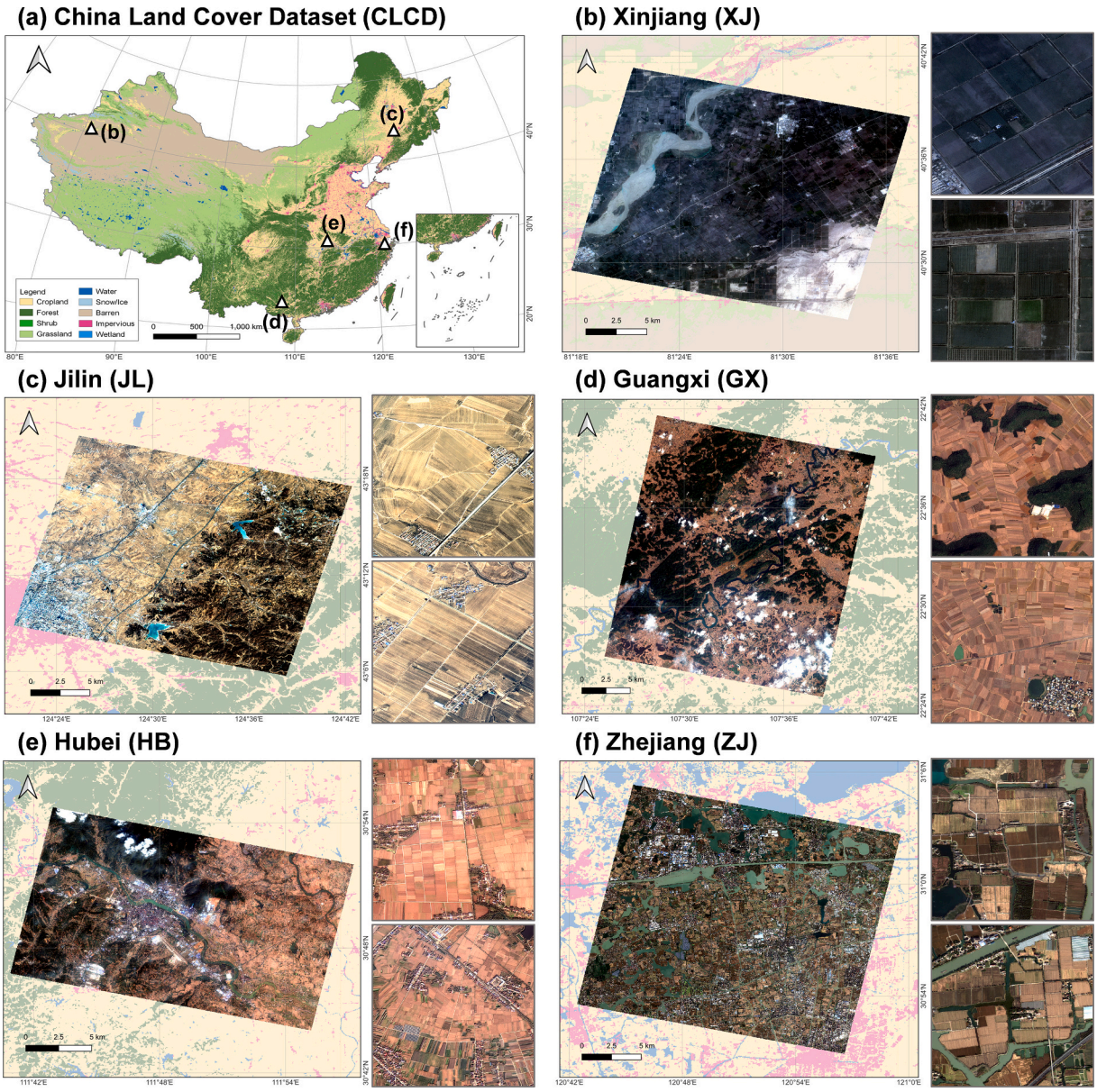

Fig. 2. The study sites with various agricultural characteristics. (a) Their locations in China. (b)-(f) The 1-m GF-2 imagery and cropland parcel examples for the five case study sites.

2.2.4. Land cover data

We used the China Land Cover Dataset (CLCD) (Yang and Huang, 2021) as a priori of cropland percentage for the newly proposed balanced sampling approach. The CLCD products were made through feeding temporal metrics derived from Landsat images into a random forest classifier, with a reported overall accuracy of 79.3 %. The CLCD dataset provides 30-m annual land-cover products from 1990 to 2022 for China, composed of nine land-cover classes including cropland. The CLCD data of the year 2022 (Fig. 1a) was selected and aligned with the GF-2 images to estimate cropland proportion for each image patch. The CLCD dataset are publicly available at https://zenodo.org/records/8176941.

3. Methods

3.1. Meta-analysis methodology

We recorded six sampling-related attributes for the 334 articles, including *mapping theme*, *number of categories*, *number of training patches*, *number of total patches to be mapped*, *sampling distribution approach*, and *transferring method*. The attribute *number of training patches* refers to the samples used for model training, whereas *number of total patches to be mapped* represents the total number of patches covering the entire mapping area. The *training patch proportion*, calculated as the ratio between these two attributes, was used as an indicator of the training sample size relative to the total mapping area. Table 2 describes ten attributes that we recorded for the 334 articles. We conducted a statistical analysis based on ten sampling-related attributes to summarize the current status of sampling strategies in practical remote sensing mapping using semantic segmentation. This meta-analysis offers valuable insights into the sampling practices reported across the literature. While our case study focuses on a typical application, the meta-analysis enables us to examine sample selection strategies across a broader spectrum of remote sensing applications. This broader perspective not only provides a more comprehensive understanding of real-world sampling approaches but also helps assess the generalizability of the findings derived from our case studies.

Of note is that many studies did not report specific information on

Table 1
Five case study sites for cropland parcel extraction to test performance of different sample configurations ("total patches to be mapped" denotes the total number of samples resulting from dividing the entire imagery into image patches).

Study sites	Region	Image size (pixels)	Total patches to be mapped	Acquisition date	Agricultural system types	Descriptions
XJ	Northwestern China	45,605 × 33,853	14,939	09/15/2022	Mechanized large-scale agriculture	The croplands are large and situated in flat terrain, characterized by highly mechanized agricultural practices. The main crops include cotton and wheat, primarily grown in a single season.
JL	Northeastern China	48,883 × 34,607	15,998	03/28/2022	Northeast Plains agricultural system	Croplands are evenly distributed across plain landscapes, with large-scale monoculture practices. Common crops include maize, soybean, and rice, predominantly cultivated in a single season.
GX	Southwestern China	37,770 × 39,527	14,981	03/12/2022	Smallholder fragmented agriculture	The region is mountainous, with small, irregular, and fragmented croplands. Typical crops include sugarcane, rice, and fruits, mostly grown in two seasons under smallholder management.
HB	Central China	37,914 × 24,832	8917	10/23/2022	Mixed plain and mountain agriculture	Croplands in this region are distributed across varied topography, including plains and mountainous areas. Main crops are rice, wheat, and rapeseed, cultivated in both single and double cropping systems.
ZJ	Eastern China	35,362 × 34,542	12,728	12/14/2022	Smallholder agriculture with abundant water systems	Croplands are fragmented and scattered, with abundant water supporting rice, wheat, oilseed rape and vegetables. The fields are managed by smallholders, with both single and double cropping systems practiced.

Table 2
Ten attributes for the reviewed articles.

ID	Attributes	Definition	Values
1	Journal	Published journal of the paper	
2	Year	Year of publication	
3	Mapping theme	Object of remote sensing semantic segmentation mapping	1) Land cover ; 2) Agricultural; 3) Forest; 4) Impervious surface; 5) Water body; 6) Disaster; 7) Others
4	Number of categories	Count of distinct classes to be segmented in the semantic segmentation task	
5	Number of training patches	Number of patches selected for model training	
6	Number of total patches to be mapped	Number of patches to be mapped for a semantic-segmentation task	
7	Training patch proportion	The proportion of training patches over the total patches to be mapped	
8	Sampling distribution approach	The sample distribution strategy of the training samples	1) Random: samples are selected randomly; 2) Balanced: samples are selected to ensure equal representation across classes; 3) Systematic: samples are selected based on a predefined selection rule; 4) No information
9	Transferring method	Transferring strategy used when validating model transfer capabilities or mapping in other regions	1) Direct use of an existing model; 2) Sample transfer: model is trained using a mix of local and global samples; 3) Model transfer: model is pretrained with global samples and then fine-tuned with local samples; 4) No transfer
10	Comments	Other situations encountered in information extraction	

the number of training patches or total mapping patches. If the number of training patches was unavailable, the study was excluded from the sample size statistics (i.e., attributes 5–7 in Table 2). When the number of training patches was reported but the total number of patches was missing, we estimated the latter by dividing the total number of pixels in the mapping area by the size of a training patch. If the total number of pixels was not provided, we calculated it using the reported mapping area and the spatial resolution of the remote sensing imagery. Studies lacking all these information were excluded from the sample size statistics.

3.2. Case study design

3.2.1. Deep learning model

Our study employed ResUNet-a model (Diakogiannis et al., 2020), an established deep learning model for cropland parcel extraction, which has shown superior performance in past studies (Waldner and Diakogiannis, 2020; Jong et al., 2022; Li et al., 2023). Its structure was presented in Fig. 3. ResUNet-a is a deep learning model built upon the UNet architecture (Ronneberger et al., 2015), featured by a symmetric encoder-decoder structure with skip connections between corresponding stages. This design enhances the model ability to capture contextual information while maintaining fine-scale details. The core design of ResUNet-a is the ResBlock-a module (Fig. 3b) that integrates multiple parallel atrous convolutions within residual block. The atrous convolutional branches with different dilation rates perform feature extraction at various receptive fields, improving identification of targets at different scales and locations. To enhance the model performance, ResUNet-a employs a pyramid scene parsing pooling (PSP pooling) layer between the encoder and decoder, as well as before the output layer, and then applies a multi-task learning to achieve more accurate cropland parcel extraction by leveraging the constraints between related tasks.

We employed the Tanimoto distance as the loss function during the training stage to aid in model convergence while maintaining a balance across multiple tasks (Waldner and Diakogiannis, 2020). Tanimoto loss, a variant of Dice loss, focuses on maximizing the overlap between predicted and true labels in segmentation tasks. It provides smoother gradient changes during optimization, leading to more stable convergence (Diakogiannis et al., 2020). This results in better boundary accuracy and overall segmentation quality, particularly in improving the shape of parcels. The Tanimoto distance loss function is defined as:

$$\widetilde{T}(p,l) = \frac{T(p,l) + T(1-p,1-l)}{2} \tag{1}$$

with

$$T(p,l) = \frac{\sum_i (p_i l_i)}{\sum_i \left(p_i{}^2 + l_i{}^2\right) - \sum_i (p_i l_i)} \tag{2}$$

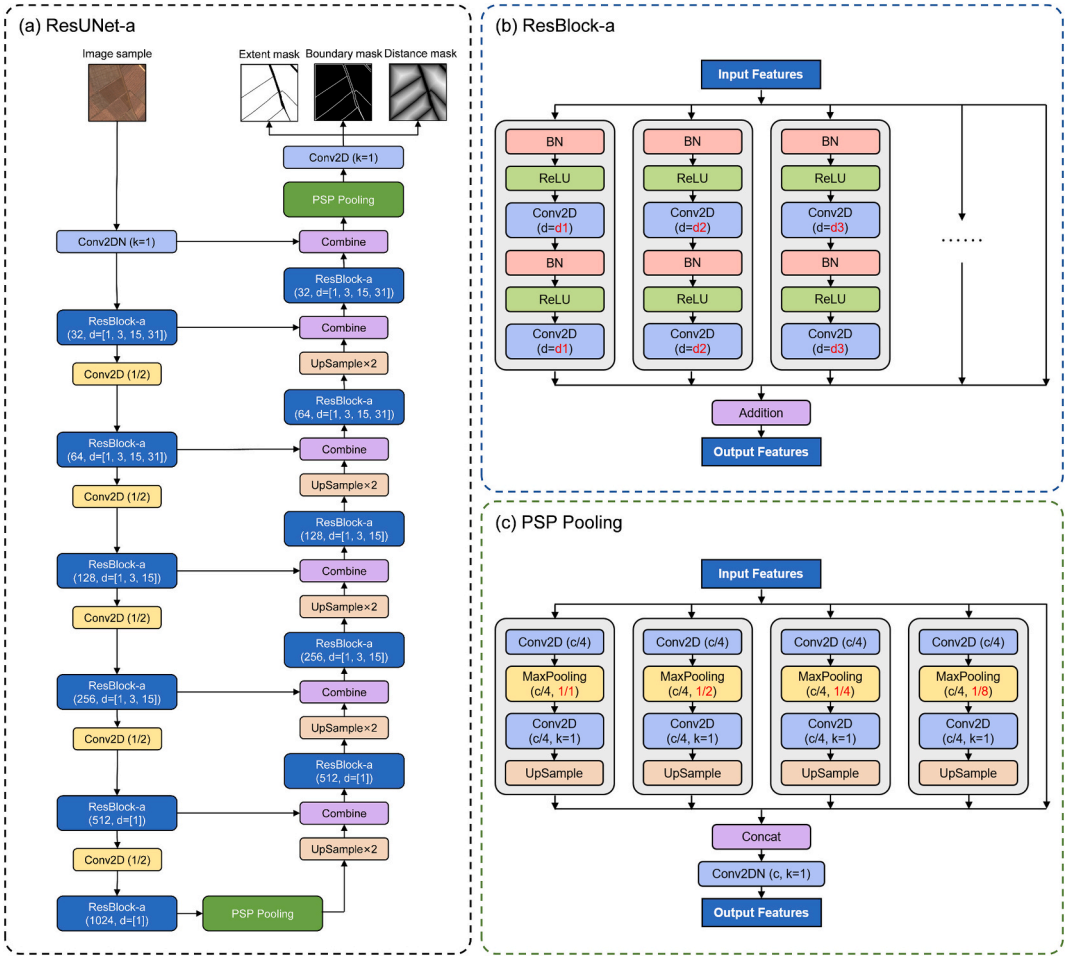

Fig. 3. Graphic explanation for ResUNet-a model (Diakogiannis et al., 2020). (a) The overview of the ResUNet-a architecture. (b) The ResBlock-a module that integrates residual connection and parallel atrous convolutions with different dilation rates. (c) PSP pooling layer.

where p presents the probability map output by the model, and l represents the corresponding sample label. For multi-task learning, we used the average of the loss functions for all tasks:

$$\widetilde{T}_{MTL}(p,l) = \frac{\widetilde{T}_{extent}(p,l) + \widetilde{T}_{boundary}(p,l) + \widetilde{T}_{distance}(p,l)}{3} \tag{3}$$

We used the AdamW optimizer (Loshchilov and Hutter, 2019) with a weight decay of 1×10^{-5}. The maximum number of training epochs was set to 200, with an initial learning rate of 1×10^{-4}. After 100 epochs, the learning rate decayed to 10 % of the initial value, resulting in a learning rate of 1×10^{-5} for the remaining 100 epochs. The batch size was set to 16 during training.

3.2.2. Performance evaluation

In this study, we chose the GTC (Global Total-Classification) to evaluate the accuracy of deep learning models with different sampling configuration. GTC is an object-based evaluation metric that comprehensively assesses over-segmentation and under-segmentation errors in model predictions, making it well-suited for assessing both the geometric and thematic accuracy of parcel segmentation tasks (Li et al., 2023; Zhao et al., 2025). GTC is formulated as follows:

$$\mathrm{GTC} = \sum_{i=1}^{n}\left(S_T(i) \times \frac{area(P_i)}{\sum_{i=1}^{n} area(P_i)}\right) \tag{4}$$

with

$$S_T(i) = \sqrt{\frac{S_O(i)^2 + S_U(i)^2}{2}} \tag{5}$$

$$S_O(i) = 1 - \frac{area(R_i \cap P_i)}{area(R_i)} \tag{6}$$

$$S_U(i) = 1 - \frac{area(R_i \cap P_i)}{area(P_i)} \tag{7}$$

where R_i denotes the reference parcels in ground truth, P_i denotes the predicted parcels in model results, n represents the parcel number, S_O is the over-segmentation error, and S_U is the under-segmentation error.

We generated the GTC curve along with increasing sample sizes. The best sample size was identified when the GTC curve just reached the level-off point, which was defined as the first sample size reaching 95 % of the maximum accuracy, i.e., $1 - 0.95 \times (1 - GTC_{min})$, following the practice for the semi-variogram in geostatistics (Chen and Gong, 2004; Garrigues et al., 2008). Given that the GTC typically decreases with an increasing number of training samples and eventually will be stabilized, its trend is analogous to an inverted semi-variogram curve, where the point of stabilization at which GTC first levels off corresponds to *sill*. Unlike significance testing methods (Collins et al., 2020; Foody et al., 2006), which identify pre-defined sample sizes or accuracy thresholds, the variogram-based method enables us to pinpoint the exact point when the error curve first flattens out, even when the number of samples at that point is not predetermined. The Y value of the level-off point reflects the minimum GTC (i.e., the best model performance), while its X value is

the minimum required sample size to reach the optimal sample size for the best balance between the accuracy and the annotation costs. The smaller X value of the level-off point indicates the less sample patches required to reach the best model performance.

To validate our findings, we conducted the same curve-based analysis using the F_1-score, another widely used performance metric for semantic segmentation in the literature (the results were shown in the *Supplementary Material*). In addition, we will perform a visual comparative evaluation of parcel maps generated under different sampling configurations to confirm the accuracy of the model achieved by the selected best sampling strategy.

3.2.3. Experiment setup

To answer three research questions for this study, we designed three sub-experiments respectively for training size, sample distribution and transferring method (Fig. 4):

3.2.3.1. Training size. For each study site, we established a series of training sample sets using the proportions of training patches out of the total patches in mapping areas, with a step size of 0.2 % for the range from 0 % to 1 %, and a step size of 1 % for the range from 1 % to 10 %. We assessed the GTC curves under two sample distribution approaches, i.e., the random sampling and the newly proposed balanced sampling (see "*sample distribution*"). Their optimal training size was determined as the X value of the level-off point of their GTC curve.

3.2.3.2. Sample distribution. Sampling design in machine learning can generally be divided into three categories: random sampling, stratified sampling (i.e. balanced sampling), and systematic sampling (Olofsson et al., 2014). Systematic sampling is inherently rule-based and context-dependent, making it difficult to evaluate comparatively. Therefore, we will test the other two sample distribution approaches, random sampling and stratified sampling for semantic segmentation. Existing stratified sampling methods for semantic segmentation studies often used arbitrary definition for patch-based stratum. For example, Qurratulain et al. (2023) interpreted the majority category for each training patch and then assigned higher weights for those patches dominated by minority category. Also, focusing solely on categorical attributes overlooked the morphological complexity of the geographic features. For this study, we proposed a new stratified sampling to balance the training patch selection, hereafter named as *balanced sampling*. We defined training strata based upon two patch-based features: thematic entropy (H) and edge complexity (E). Thematic entropy captures the amount of information related to the categorical composition within each patch, while edge complexity quantifies the structural intricacy of object boundaries. Together, these two features characterize training patches from complementary perspectives, i.e., thematic and morphological measurement. Entropy calculation requires external data that provide prior knowledge about the class distribution, whereas edge complexity can be derived directly using edge detection algorithms, without the need for additional input data.

Fig. 5 illustrates the process of the proposed balanced sampling. The thematic entropy reflects the variation of map categories, and the patch with low entropy has more uniform pixel labels. We derived a rough estimation of the cropland percentage from a prior crop product (i.e., the CLCD product), which often exists for most remote sensing applications, and then calculated the thematic entropy (H) using the following formula:

$$H = -\sum_{i=1}^{n} P_i \log P_i \tag{8}$$

where n presents the number of categories to be classified, and P denotes the proportion of each category. Based on the geographic coordinates of each training sample, the corresponding area in the land-cover product was identified to obtain land-cover data for that sample. Even though the spatial resolutions of the two datasets differ, aligning them by sample extent ensures that the land-cover data can still capture the proportion of cropland within the sample region. Edge complexity refers to the level of detail or intricacy in the boundaries (edges) of objects within an image. We computed edge complexity by calculating edge length after applying the Canny edge detection algorithm (Turker and Kok, 2013). We divided each feature into three groups using quantile-based thresholds, resulting in nine strata for the image patches with two features. These strata represented different cropland patterns. For

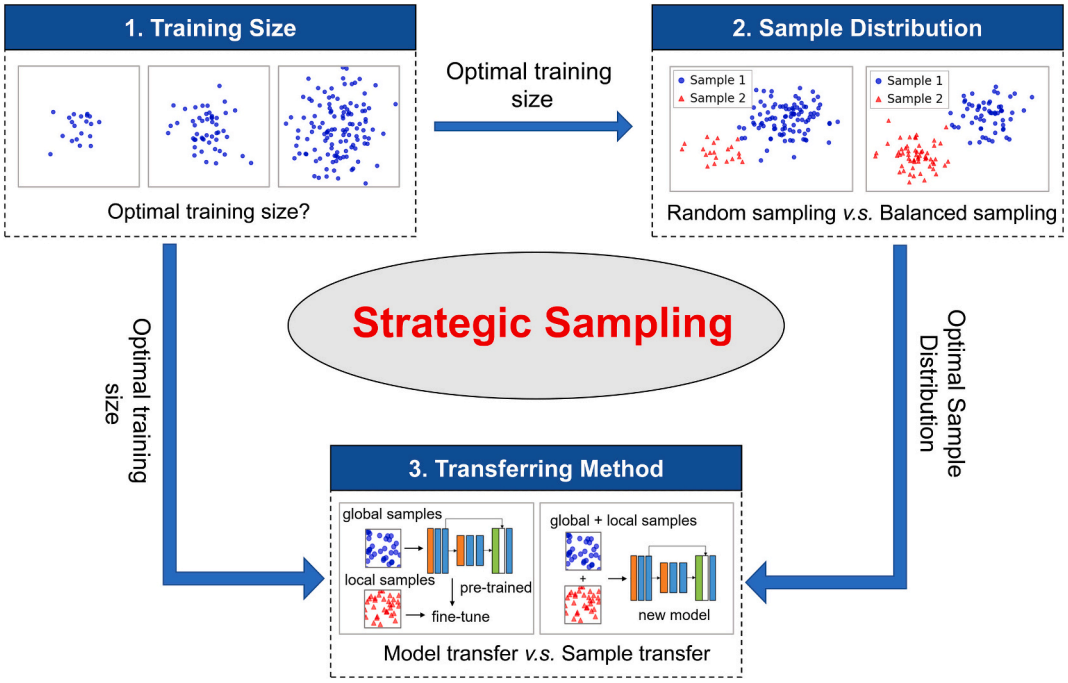

Fig. 4. The experiment design for exploring best practices for the three key factors of strategic sampling for semantic segmentation: 1) training size, 2) sample distribution, and 3) transferring method.

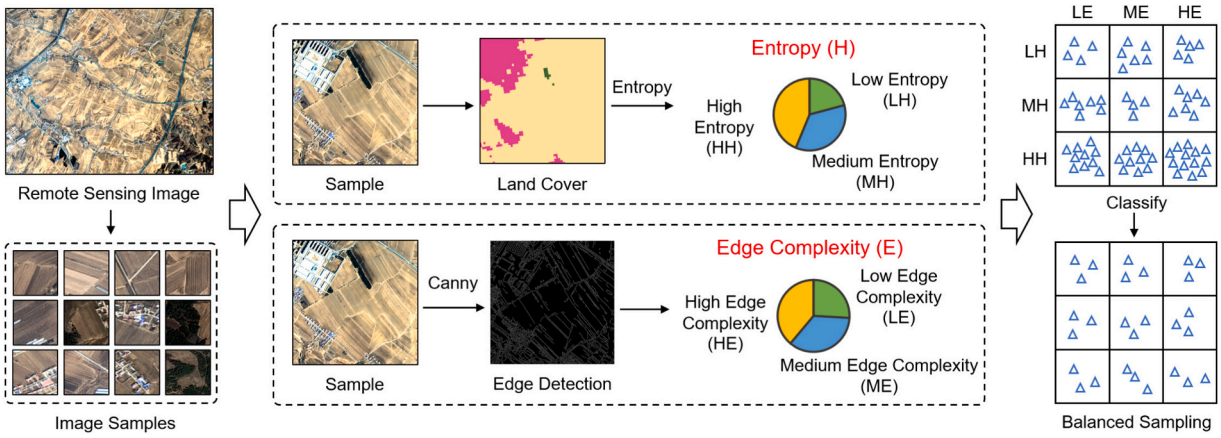

Fig. 5. The process of the proposed balanced sampling. The balanced sampling defines nine strata by two-dimensional measurements, i.e., thematic entropy and edge complexity, and then assigns training patches to each stratum using quantile-based thresholds. HH: High thematic entropy. MH: Medium thematic entropy. LH: Low thematic entropy. HE: High edge complexity. ME: Medium edge complexity. LE: Low edge complexity.

example, "low entropy – low edge complexity" corresponded to large-scale industrial cropland or non-agricultural areas with clean edges, while "high entropy – high edge complexity" referred to complex regions where scattered, small-scale croplands were interwoven with non-cropland backgrounds. The newly proposed balanced sampling will assign an equivalent number of training patches to each stratum, ensuring a high diversity of training patches, based upon the hypothesis that the more diversified and representative sample patches could reduce the demand of the total training samples needed. We will test the random and proposed balanced sampling, and compare the differences of level-off point location for the two approaches.

The code script for our proposed balanced sampling method is publicly available at https://github.com/Remote-Sensing-of-Land-Resource-Lab/Training-Sample-Selection.

3.2.3.3. Transferring methods. Based on the optimal training size and distribution strategy, we created the sample set for each site. To evaluate data and model transfer performance, we designated and rotated one of the five study sites as the local sample pool, with all samples from the remaining four sites serving as the global pool. This allowed for sample transfer and model transfer to be applied and assessed five times, each time using a different site as the testing region. For each site, we incrementally increased the training patch number from the local sample pool from 0 to 3.0 %, generated a series of semantic segmentation models from applying different transferring methods for the mixture of the subsets of the local pool and the global pool. We compared three approaches: no transfer, sample transfer, and model transfer. *No transfer* refers to training a model from scratch at the target site without any transfer strategy, using the previously described balanced sampling method to select training samples. *Sample transfer* involves combining local and global samples to form a new training set for training a deep learning model. *Model transfer*, by contrast, follows a transfer learning paradigm, where a pre-trained model is adapted to the target domain. This approach entails training a pre-trained model using global samples, after which the encoder parameters are frozen and only the decoder parameters are fine-tuned using local samples from the specific site to be tested. Specifically, the pre-trained model was first trained using global samples for 1000 rounds, and then fine-tuned on local samples for 100 rounds. During fine-tuning, the encoder parameters were kept frozen, and the decoder parameters were updated using a learning rate of 10^{-5}. We applied the same GTC curve as the evaluation tool for the comparative performance of the two transferring methods based upon an out-of-bag testing sample set. The better transferring method should require fewer local samples to reach its GTC level-off point.

4. Results

4.1. Meta-analysis results

Fig. 6 summarizes the meta-analysis on sematic segmentation for remote sensing applications. For sample size, 102 out of the 334 articles provided explicit information on the proportion of training patches over the total mapping area (Fig. 6a). The training patch proportion for semantic segmentation concentrated on either very small ("0–1 %") or very large sample sizes (">20 %"), with a median of 4.2 % total patches. We identified that the distribution strategies for semantic segmentation were dominated by random sampling which was adopted by more than half of the studies (51.1 %) (Fig. 6b). Some articles employed systematic sampling methods by formulating specific rules to mitigate this imbalance caused by random sampling (Chen et al., 2020). A very small proportion of studies (5.2 %) adopted a balanced sampling approach (Descals et al., 2021; Liu et al., 2023; He et al., 2024; Zhao et al., 2024). For example, Zhao et al. (2024) used stratified sampling, categorizing samples by time and changes in areas, selecting 300 training samples for each category to avoid imbalanced sampling. For transferring methods (Fig. 6c), only 7.9 % of the articles adopted either model or sample transfer to reduce sample annotation efforts, while 27.0 % of the articles chose the direct use of the model trained from somewhere else. This highlighted that the transferring method was still an emerging topic that required in-depth investigation. In general, model transfer was more commonly used (5.9 %) than sample transfer (2.0 %). Sample transfer was employed mainly for a large-scale mapping (e.g., Zhou and Weng, 2024). Moreover, a variety of mapping themes were identified (Fig. 6d), including "impervious surfaces" (26.7 %), "land cover" (25.1 %), and "agriculture" (10.0 %), which reflected a wide range of mapping themes in our survey to identify current states of strategic sampling.

Lastly, among the articles reporting training patch proportion, binary classification tasks were slightly more common than multi-class classification tasks (Fig. 6e). The binary classification tasks in remote sensing semantic segmentation included the extraction of ground objects such as buildings (Zhou and Weng, 2024), cropland parcels (Waldner and Diakogiannis, 2020), and water bodies (Hertel et al., 2023), while multi-class classification tasks mainly focused on the land cover classification (Xiong et al., 2024) and crop type classification (Cai et al., 2024). Surprisingly, we found no obvious correlations between the training patch proportion and the category number to be classified (Fig. 6f). The median training patch proportion in binary classification studies was even higher than that in multi-classification studies, indicating that the increase of category number did not necessarily lead to the increase of training samples.

Fig. 6. Results of the meta-analysis after extracting information from the selected papers. (a) Distribution and statistical analysis of training patch proportion, (b) sample distribution, (c) transferring method, (d) mapping theme, (e) class types, and (f) training patch proportions across different categories.

4.2. Case study results

4.2.1. Training size and sample distribution

Fig. 7 presents the GTC curves with increasing training size using the newly proposed balanced and random sampling methods for five case studies of cropland parcel mapping. As the number of training sample patches increased, the GTC value gradually decreased until they levelled off, with the dashed line indicating where the level-off point, i.e., $1 - 0.95 \times (1 - GTC_{min})$, was first reached. For most study sites, the GTC for balanced sampling (blue curve) reached the level-off point earlier than for random sampling (orange curve), as indicated by the blue dashed line on the x-axis being to the left of the orange dashed line. The distance between the two dashed lines on the x-axis reflected the difference in the minimum required training patch proportion between the two sampling schemes. In the XJ and GX study areas, the differences in required sample size between the two sampling methods were relatively small. This was attributed to the geographically consistent morphological patterns of cropland parcels in the remote sensing images, which resulted in a limited level of sample imbalance even under random sampling. The XJ study area required the fewest training samples due to its relatively simple parcel morphology, characterized by large, regularly shaped cropland parcels. Overall, the results suggested that the proposed balanced sampling method generally required fewer sample patches to train a satisfactory model compared to random sampling.

Table 3 summarizes the minimum training patch proportion for the level-off point and its corresponding GTC value, highlighting the superiority of the proposed balanced sampling method. The optimal training patch proportion for random sampling (3.3 %) was lower than the average training size summarized from the meta-analysis (4.2 %, see Fig. 6a). Compared to the random sampling method, the proposed balanced sampling strategy reduced the required training patch proportion from 3.3 % to 2.5 %, resulting in a reduction of approximately 25 % in sample annotation workload. Despite fewer samples used for model training, the balanced sampling achieved the level-off GTC almost identical to random sampling on average (0.257 vs. 0.267).

Fig. 8 confirms our findings from the predicted parcel maps using two sampling strategies with their optimal sample sizes. The random sampling method used 3.3 % of the total patches for training, while the

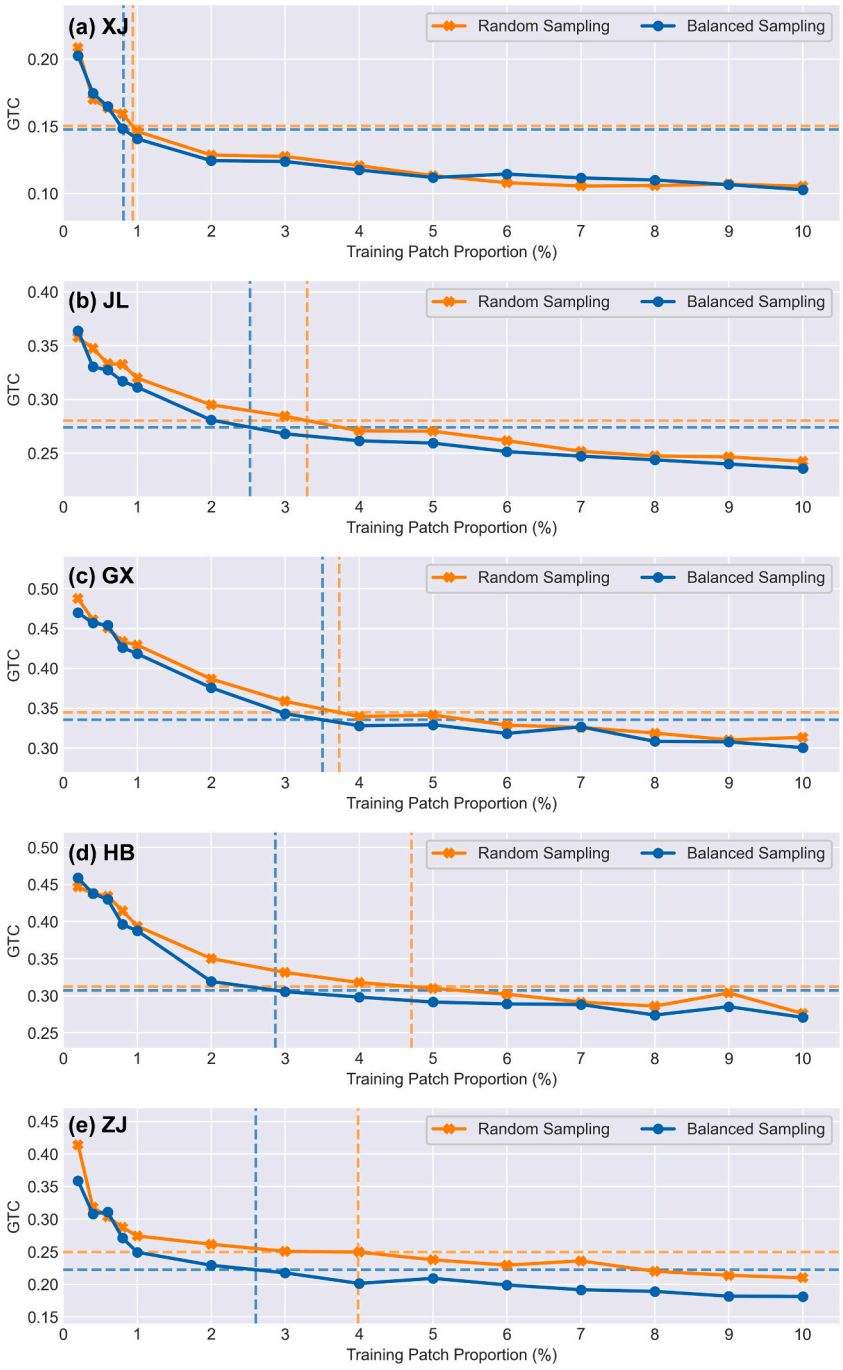

Fig. 7. Impacts of training sample size and distribution on model accuracy across five study areas. The blue line represents the proposed balanced sampling and the orange line represents the random sampling. The dashed line at the y-axis indicates the first point when GTC levels off. The dashed line at the x-axis represents the proportion of total patches as training sizes for the level-off point. The closer to the origin means the less samples needed to reach the level-off point. XJ: Xinjiang; JL: Jilin; GX: Guangxi; HB: Hubei; ZJ: Zhejiang. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The minimum training patch proportion when GTC reaches the level-off point and its corresponding GTC value for the five study sites. XJ: Xinjiang; JL: Jilin; GX: Guangxi; HB: Hubei; ZJ: Zhejiang.

	Sample distribution	XJ	JL	GX	HB	ZJ	Average
Minimum training patch proportion (%) for the level-off GTC	Random sampling	0.9	3.3.	3.7	4.7	4.0	**3.3**
	Balanced sampling	0.8	2.5	3.5	2.9	2.6	**2.5**
Level-off GTC	Random sampling	0.150	0.280	0.345	0.312	0.250	**0.267**
	Balanced sampling	0.148	0.274	0.335	0.307	0.222	**0.257**

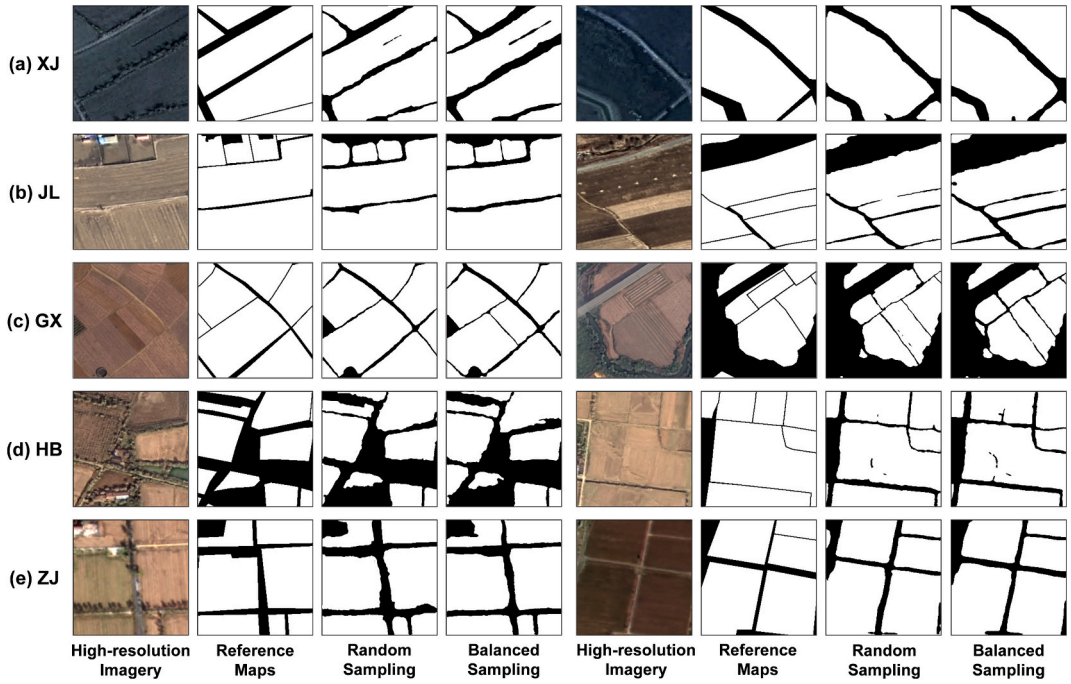

Fig. 8. Cropland parcel extraction results using 3.3 % training samples selected by random sampling and 2.5 % by balanced sampling. XJ: Xinjiang; JL: Jilin; GX: Guangxi; HB: Hubei; ZJ: Zhejiang.

balanced sampling method used only 2.5 %. Both methods achieved comparable performance in delineating cropland parcels, as visually observed from predictions. Despite using approximately 25 percentage of fewer training samples, the balanced sampling strategy produced result maps that were nearly identical to those generated by the random sampling method. We chose the balanced sampling with its optimal training sample size (2.5 % of total patches in mapped areas) as the best configuration for the next stage of the transferring test.

4.2.2. *Transferring methods*

Fig. 9 and Table 4 compare the GTC error curves and level-off points for three strategies: sample transfer, model transfer, and training solely with local samples (i.e., no transfer). When the number of local training samples was limited, both transfer-based methods yielded lower GTC errors than the no-transfer approach, benefiting from the inclusion of global samples. However, as the volume of local training data increased, the performance of the no-transfer model gradually surpassed that of both transfer strategies. This suggested that in data-rich scenarios, transfer methods might introduce conflicting information between global and local samples, ultimately reducing segmentation accuracy.

The average local training patch proportion required to reach the level-off point was 0.5 % for both sample and model transfer, only one-fifth of the requirement under the no transfer strategy, which reached the level-off at 2.5 %. However, this efficiency came with a trade-off: both transfer methods resulted in higher average level-off GTC values (0.298 for sample transfer and 0.308 for model transfer) compared to no transfer (0.257). This implied that while transfer strategies were advantageous in low-data settings, they might slightly compromise accuracy due to the inclusion of out-of-region global samples.

Between the two transfer approaches, sample transfer achieved slightly better model performance than model transfer (0.298 vs. 0.308), despite their same optimal sample sizes. This might be attributed to the fact that sample transfer trained the model from scratch, allowing better integration of local and global data, whereas model transfer fine-tuned only the decoder parameters of a pre-trained model, potentially limiting its adaptability.

4.2.3. *Implications of training sample size for parcel products*

Fig. 10 illustrates the cropland parcel extraction results based on full-scene imagery from the JL and ZJ study sites. We conducted experiments using the balanced sampling strategy under three training patch proportion scenarios: sparse (0.5 %), optimal (2.5 %), and excessive (10.0 %) of the total patches to be mapped. The model trained with 0.5 % of the data exhibited noticeable omission errors and poor boundary extraction. In contrast, the outputs from the 2.5 % and 10.0 % training sizes were visually similar, indicating that increasing the sample size beyond the optimal level provided no substantial improvement. We further evaluated key parcel-level statistics for each scenario, i.e., the number of cropland parcels, average parcel area, and total cropland area. The results for the 2.5 % and 10.0 % cases displayed highly consistent patterns, reinforcing the conclusion that their mapping quality was comparable.

4.2.4. *Summary*

Fig. 11 summarizes the results where optimal components were systematically added for strategic sampling to analyze how their addition impacted the required sample sizes and GTC values. It started from a random sampling of image patches which averagely required 3.3 % of total patches in mapped areas as training sample size to reach the GTC level-off point. The preferred distribution strategy, the newly proposed balanced sampling, could reduce the number of samples required for semantic segmentation model training from 3.3 % to 2.5 %, and also slightly reduced GTC errors (0.267 vs 0.258). Sample transfer would greatly reduce the local training sample demand, resulting in 0.5 % of total patches in mapped areas as the final local sample size, although it caused the GTC to increase from 0.258 to 0.298. We recommend using balanced sampling generally. However, when the cost of sample annotation is prohibitively high, applying sample transfer could significantly reduce annotation workload, albeit with a slight decrease in accuracy.

Besides, we applied the same analysis procedure based upon F_1-score curve. The results were almost the same as using the GTC curve (balanced vs. random sampling: 2.5 % vs. 3.5% % for minimum required training area; sample vs. model transfer: 0.6 % vs. 0.7 % for optimal training patch proportion). For the details, please refer to *Section S2* of

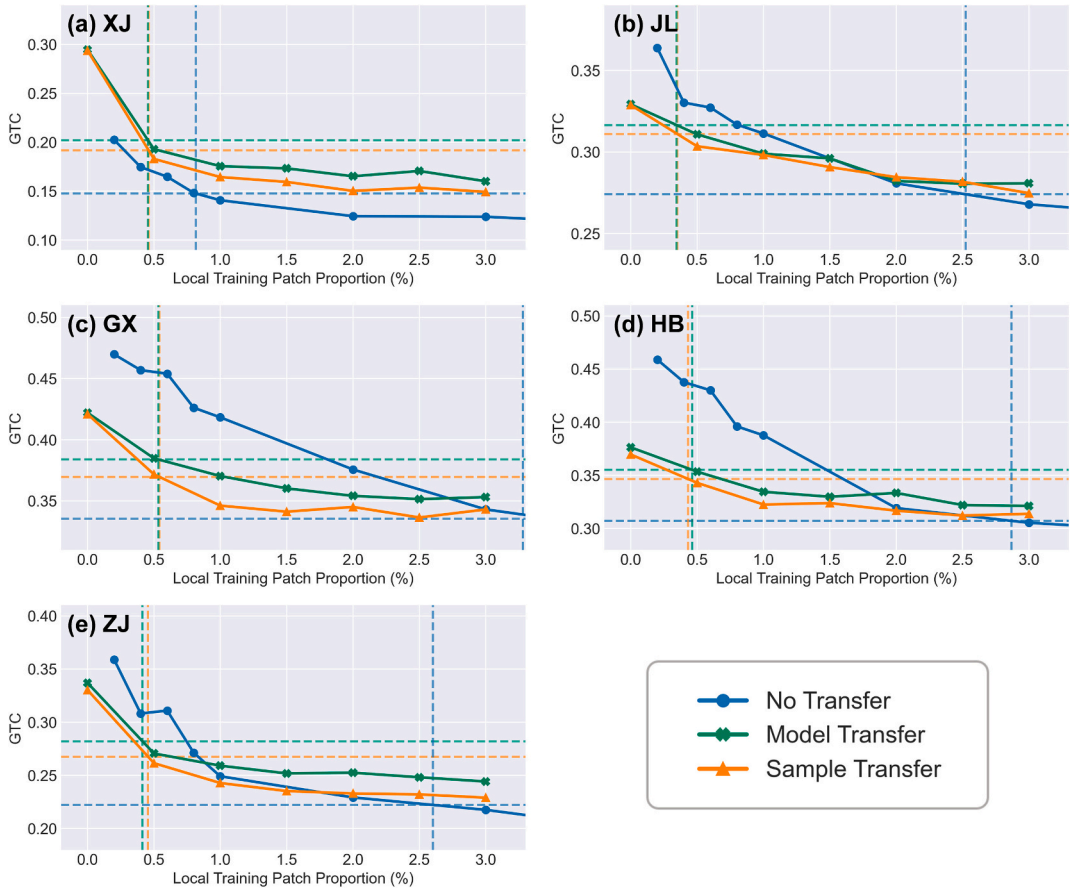

Fig. 9. Impacts of local training size and transferring method on model accuracy for five study sites. The local training patch proportion represents the proportion of patches selected for training from the site to be mapped over the total patches of the same site. The dashed line on the y-axis indicates the level-off point that first reaches 95 % of the maximum accuracy. The dashed line on the x-axis represents the minimum number of local samples required for the model to reach the level-off point. No transfer requires much higher sample size to reach its level-off point than sample and model transfer. XJ: Xinjiang; JL: Jilin; GX: Guangxi; HB: Hubei; ZJ: Zhejiang.

Table 4
The minimum local training sample size and its level-off GTC value for different transferring methods. XJ: Xinjiang; JL: Jilin; GX: Guangxi; HB: Hubei; ZJ: Zhejiang.

	Transfer strategy	XJ	JL	GX	HB	ZJ	Average
Minimum training patch proportion (%) for level-off GTC	No transfer	0.8	2.5	3.5	2.9	2.6	**2.5**
	Sample transfer	0.5	0.4	0.5	0.4	0.5	**0.5**
	Model transfer	0.5	0.4	0.5	0.5	0.4	**0.5**
Level-off GTC	No transfer	0.148	0.274	0.335	0.307	0.222	**0.257**
	Sample transfer	0.192	0.311	0.370	0.347	0.268	**0.298**
	Model transfer	0.202	0.317	0.384	0.355	0.282	**0.308**

the *Supplementary Material*.

5. Discussion

5.1. Training size

The meta-analysis of various semantic segmentation applications and case studies on cropland parcel mapping consistently demonstrated that approximately 4 % of the total patches were the optimal training size when the most commonly used patch-based sampling scheme, random sampling, was applied. The training size directly impacts the costs in terms of sample annotation, as larger datasets require more labeled examples, often leading to higher manual effort or the need for automated labeling tools. Additionally, redundant sample preparation—where duplicate or overly similar data is included—can unnecessarily inflate the budget, as it fails to provide aids in improving model performance while still requiring resources for annotation and storage. Admittedly, the "best" sample size is not a one-size-fits-all scenario. A variety of factors, beyond the total patches to be mapped, influence the optimal size of the training dataset, including geographic variability of the target, image resolution, and seasonal fluctuations. Our study provides an important heuristic to quickly determine the optimal sample size that strikes the balance between sample size and model performance, serving as a powerful tool for initiating operational mapping projects with minimal upfront data. It would allow for more informed decision-making, especially in time-sensitive or resource-constrained projects where preliminary tests might not be feasible.

The meta-analysis revealed an unusual pattern in which most studies utilized either less than 1 % or more than 20 % of available data as training samples (Fig. 6a). This bimodal distribution may be explained by two main factors. First, in many existing studies, the size of the training dataset is primarily constrained by data acquisition and

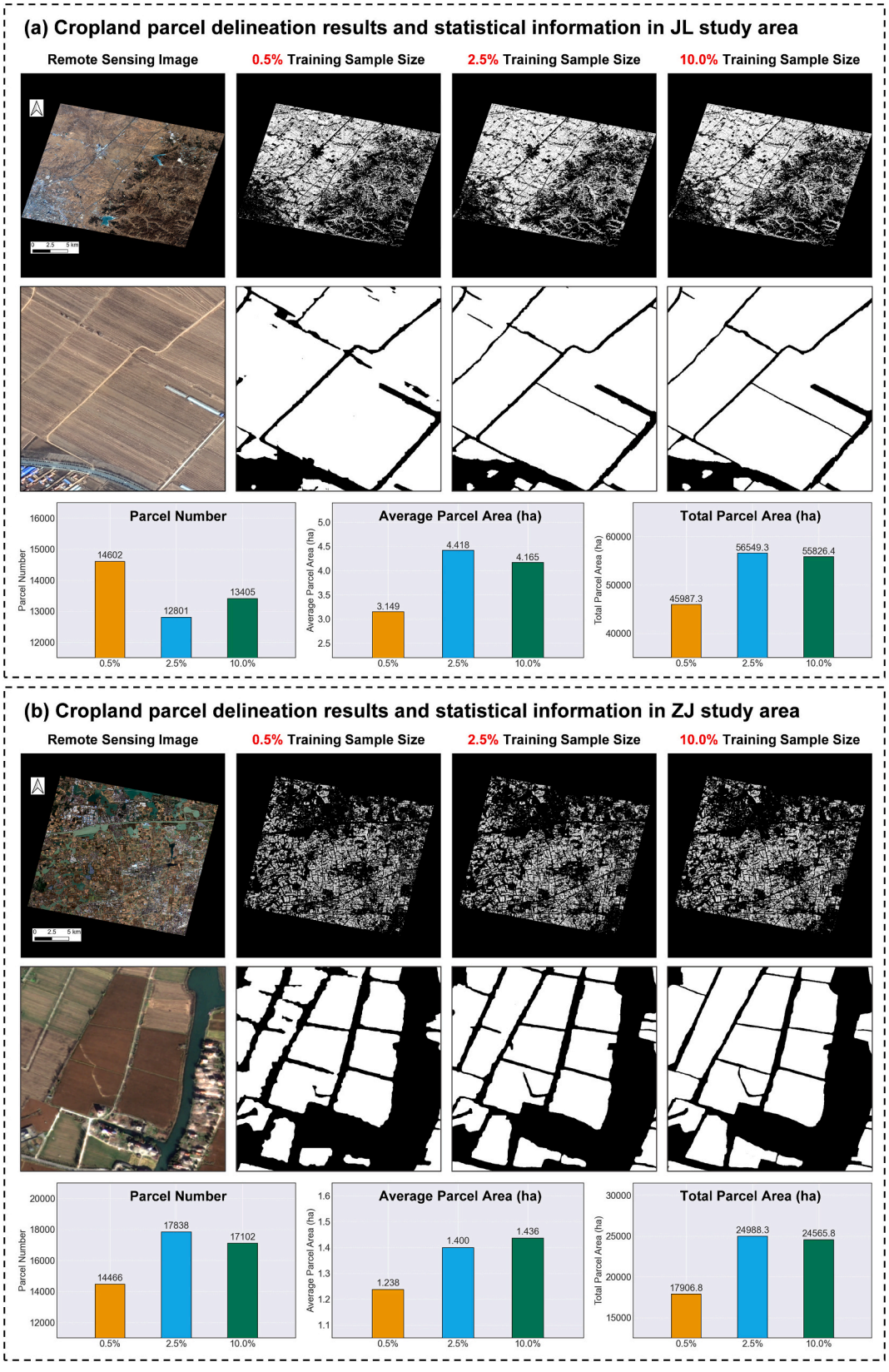

Fig. 10. Effects of different training sample sizes on regional mapping. Taking (a) JL and (b) ZJ study sites as examples, cropland parcel extraction results based on remote sensing imagery were presented using sparse (0.5 %), optimal (2.5 %), and excessive (10.0 %) training sample proportions. The number of cropland parcels, average parcel area, and total cropland area within the image extent were summarized.

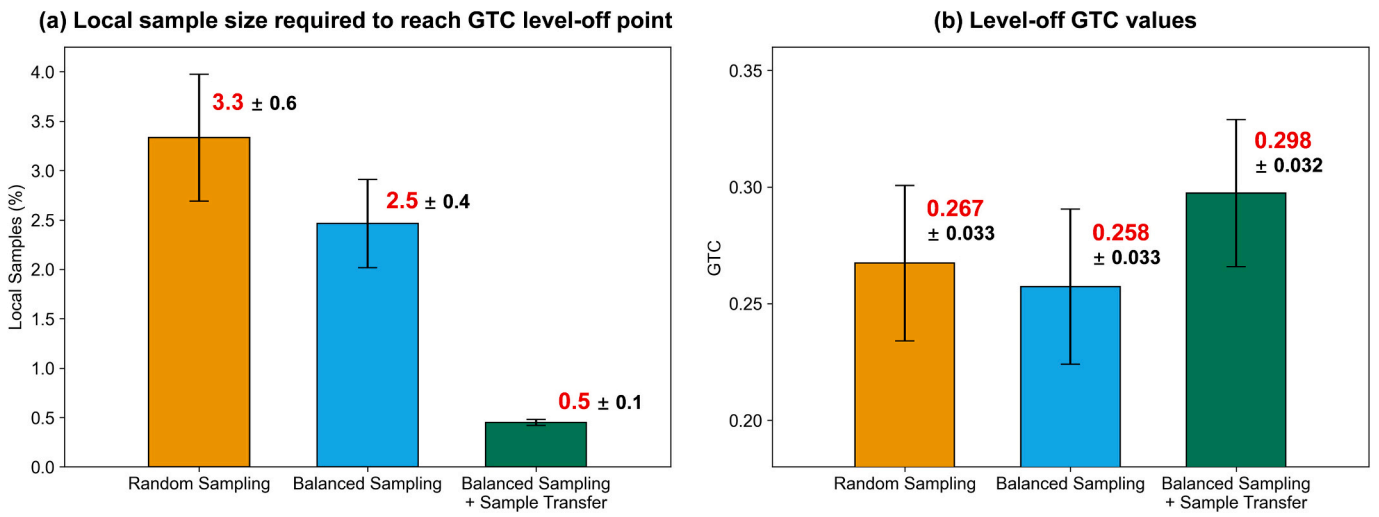

Fig. 11. Comparison of the results for random sampling, balanced sampling, and sample transfer. (a) The local sample size required to reach GTC level-off point as the proportion of the total patches to be mapped. (b) The level-off GTC values for the minimum local sample size.

annotation costs, rather than determined by the number of target classes. Consequently, there is no consistent correlation between the number of training samples and the number of classification categories reported in the literature. Second, current semantic segmentation researches tend to follow two distinct directions: few-shot learning, which focuses on model performance with minimal labeled data, and large-scale training, which relies on extensive annotated datasets. This divergence has contributed to the clustering of studies at both extremes of the sample size spectrum. The emergence of this pattern underscores the absence of standardized guidelines for determining training sample sizes in remote sensing applications, highlighting the relevance and necessity of our investigation.

In multi-class classification scenarios, it has been suggested that the training sample size should increase proportionally with the number of categories to ensure adequate representation for each class (Foody et al., 2006). However, our meta-analysis did not reveal clear evidence supporting this assumption in the context of semantic segmentation. In contrast, we observed that binary classification tasks tended to utilize a larger proportion of training samples than multi-class tasks (Fig. 6f). This trend likely reflects the fact that the most influential factor determining sample size is the available annotation resources, rather than the number of target categories. Given that binary classification tasks typically involve simpler and less time-consuming labeling processes, they can accommodate more training samples within the same annotation budget. A rigorous assessment of the relationship between the number of categories and the required training sample size with an assumption on unlimited annotation resources would be excessively complex and falls beyond the scope of this study. Future research is needed to determine whether optimal training sample sizes scale proportionally with the number of categories in semantic segmentation tasks. For example, if 2.5 % of the total image patches are sufficient for a balanced binary classification task, a seven-class task might require approximately 8.8 % of the patches for training (2.5 % / 2 × 7 = 8.8 %). This hypothesis, however, remains to be empirically validated.

5.2. Sample distribution

Random sampling is the dominant approach for generating training image patches for semantic segmentation in past studies. More than half of the studies used random sampling from our meta-analysis, owing to its simplicity and ease of implementation. However, random sampling can lead to repetitive or highly similar patches, which do not add much new information to the model and could cause certain classes to be underrepresented. We proposed the new diversity-based balanced sampling specifically for image patches, which distributed the even sample number into several patch strata defined by object edges and categorical proportion from a priori map. From five case studies (Fig. 7), we observed that the new balanced-sampling approach generally reached the accuracy level-off point earlier than random sampling, and reduced the sample size demand from 3.3 % to 2.5 % of the total mapped region. Also, the balanced sampling reached a slightly lower level-off GTC error (0.257 vs. 0.267), as such diversity-based sampling captures the variance in the data and prevents overfitting to a specific region of the feature space. It is simple to expand balanced sampling for multi-class scenarios in terms of various patch-based deep-learning applications. Therefore, the new balanced sampling is recommended for future practice of strategic sampling.

The proposed balanced sampling method relies on existing land cover products to obtain categorical proportion priors. Our meta-analysis revealed that most mapping themes, such as impervious surfaces, land cover, agriculture, forest, and water bodies, were well supported by existing remote sensing products. Cases where no prior information is available are relatively rare. For specific applications lacking such products (e.g., disaster mapping or greenhouse extraction), users can either rely on the single indicator of edge complexity or adopt a fallback strategy by randomly selecting approximately 4 % of the total samples. While this may increase data preparation costs or introduce some sampling bias, the negative impacts could be minimal, thereby preserving the overall practicality of the method even in the absence of prior data. The accuracy of the existing products is another possible factor for compromising the effectiveness of the balanced sampling through affecting the estimate of categorial proportion for patch-based entropy. However, it is important to note that the entropy calculation relies solely on the pixel quantity of each class, rather than the precise spatial distribution. Therefore, the use of an existing land cover product, even one with moderate spatial accuracy or a coarser resolution, remains an effective strategy for guiding sample selection, provided that the quantity errors are not excessively high.

5.3. Transferring method

Transferring methods have demonstrated their superiority for reducing sample preparation, but only were used in 7.9 % of the past semantic segmentation studies from our meta-analysis. Our case studies showed that the model transfer and the sample transfer presented an equal performance for decreasing the demand of the local samples from 2.5 % to 0.5 %, only a quarter of the optimal sample number when no transfer was applied. Sample transfer exhibited a slightly lower GTC

error (0.298 vs 0.308), benefiting from its process of rebuilding the model from scratch with more flexibility in tuning the weights of all hidden layers, making it the preferred approach for this study. Sample transfer is ideal when a large amount of labeled data is available in the source domain, but needs to be transferred to a target domain where data is sparse—a common scenario in remote sensing applications. However, if the source and target domains are too dissimilar in terms of data distributions, feature types, or underlying structures, the transferred samples may not be as effective, potentially resulting in poor generalization. On the other hand, model transfer leverages the generalization capabilities of a model that has already learned to detect basic features from a large dataset, considerably reducing training time and computational resources. However, its performance heavily depends on the quality of the pre-trained model. If a pre-trained model has been built with a large, representative dataset and computational resources are limited, model transfer is recommended because it significantly reduces computation costs. In other situations, particularly when global samples from similar domains are available, sample transfer is preferred due to the higher accuracy achieved by developing a deep learning model from the scratch.

5.4. Generalizability

We used cropland parcel extraction as a case study to investigate strategic sampling approaches for remote sensing semantic segmentation tasks. As a representative semantic segmentation challenge, cropland parcel extraction involves substantial variation in parcel size, shape, and spectral characteristics across regions, making it broadly reflective of the complexities faced in many segmentation applications. The five study areas selected for this analysis are distributed across diverse geographic regions of China, each exhibiting distinct landscapes and agricultural systems. For example, the XJ region is characterized by mechanized agriculture, resulting in large, regularly shaped parcels, whereas the GX region, located in the hilly terrain of southwestern China, is dominated by smallholder farming, producing fragmented and irregular parcels (Fig. 2). Our experiments demonstrated that the proposed balanced sampling approach consistently reduced annotation effort compared to conventional random sampling across all five regions, highlighting its broad applicability.

To ensure comparability across different spatial resolutions, training sample size was measured as the proportion of training patches rather than absolute pixel counts. Also, the strategy was specifically designed for regional mapping tasks, where geographic and cropland characteristics tended to be relatively homogeneous, minimizing generalization issues. For large-scale mapping efforts, we recommend first performing geographic stratification and then applying the strategic sampling method within each sub-region to construct effective training datasets.

It is important to note that evaluating strategic sampling across all possible segmentation tasks within a single study is nearly impossible. To address this limitation, we conducted a meta-analysis encompassing a broader range of remote sensing applications. This meta-analysis reviewed sample selection practices across various classification tasks, including impervious surface mapping, land cover classification, and agricultural monitoring. Over half of the studies employed random sampling, with a median training sample proportion of approximately 4 % (Fig. 6). In our case study, the optimal training sample proportion under random sampling was 3.3 %, which was slightly lower but closely aligned with the meta-analysis result. This strong consistency between the meta-analysis and our case study suggested that the conclusions drawn from our study could be generalizable to other semantic segmentation tasks. While the meta-analysis offered limited insight into the comparative performance of different sampling strategies and transfer learning methods, our local experiments addressed these two critical issues using one of the most representative segmentation tasks, i.e., cropland parcel extraction.

Admittedly, the optimal sample size may vary depending on the specific segmentation objectives, landscape characteristics, and data sources. A more rigorous way would be to visualize the learning curve where increasing the dataset size yields diminishing improvements in model performance and then to pinpoint the optimal dataset size that just makes the performance plateau. Nevertheless, to find this optimal size, it would require the professionals to annotate the unnecessary sample patches to evaluate the sample size much larger than the level-off point. Our recommendations, i.e., balanced sampling using 2.5 % of the total mapping region or sample/model transfer using 0.5 %, offer an important baseline for users to test the optimal sample size for their own applications. These findings are especially valuable for time-sensitive or resource-constrained projects, where conducting a full parameter sensitivity analysis may not be feasible.

6. Conclusion

This study combined meta-analysis on 334 papers from various remote sensing applications and case studies of cropland parcel extraction to explore the best practices for strategic sampling for semantic segmentation, mainly focusing on three key factors: training size, sample distribution and transferring methods. The meta-analysis and case studies both identified ~4 % of the total mapping patches is the optimal training size. The patch-based balanced sampling newly proposed in this study, which takes account of category entropy and edge complexity, could further decrease the sample demand to 2.5 %, compared to random sampling which was adopted in over half of the papers in our survey. While transferring methods were only used in 7.9 % of the papers, our case studies showed that sample and model transfer could considerably reduce the required local sample size from 2.5 % (i.e., the sample demand for balanced sampling) to 0.5 % of the total mapping patches, with sample transfer being slightly more accurate than model transfer (GTC errors: 0.298 vs 0.308). This finding suggests transferring methods have great potential for decreasing the amount of new training dataset, despite the decreased model performance. By leveraging both the meta-analysis and the representative case study of cropland parcel extraction, our strategic sampling recommendations have the potential to be generalized to other remote sensing mapping tasks. Targeted on an operational mapping project based upon semantic segmentation, this study will provide important guidance to minimize the cost of training data collection while maintaining or even improving model's accuracy.

CRediT authorship contribution statement

Rui Lu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Ronghua Liao:** Validation, Methodology, Investigation. **Ran Meng:** Writing – review & editing, Validation, Supervision. **Yingchu Hu:** Methodology, Investigation. **Yi Zhao:** Methodology, Investigation. **Yan Guo:** Validation, Conceptualization. **Yingfan Zhang:** Writing – review & editing, Visualization. **Zhou Shi:** Supervision, Funding acquisition, Conceptualization. **Su Ye:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgement

The authors are very grateful for the financial support provided by National Key Research and Development Program of China (2023YFD1900100), National Natural Science Foundation of China (U24A20575), and National Key Research and Development Program of

China (2022YFB3903503). The authors would also like to thank the editors and anonymous reviewers of the journal for their thorough review and insightful comments that have helped improving the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2025.115034.

Data availability

I have shared the link to my data and code in the manuscript.

References

Bergen, K.J., Johnson, P.A., de Hoop, M.V., Beroza, G.C., 2019. Machine learning for data-driven discovery in solid earth geoscience. Science 363, eaau0323. https://doi.org/10.1126/science.aau0323.

Boschetti, L., Stehman, S.V., Roy, D.P., 2016. A stratified random sampling design in space and time for regional to global scale burned area product validation. Remote Sens. Environ. 186, 465–478. https://doi.org/10.1016/j.rse.2016.09.016.

Brown, A., 2006. Strategic sampling. In: Hurford, C., Schneider, M. (Eds.), Monitoring Nature Conservation in Cultural Habitats: A Practical Guide and Case Studies. Springer, Netherlands, Dordrecht, pp. 43–54. https://doi.org/10.1007/1-4020-3757-0_5.

Brown, J.F., Tollerud, H.J., Barber, C.P., Zhou, Q., Dwyer, J.L., Vogelmann, J.E., Loveland, T.R., Woodcock, C.E., Stehman, S.V., Zhu, Z., Pengra, B.W., Smith, K., Horton, J.A., Xian, G., Auch, R.F., Sohl, T.L., Sayler, K.L., Gallant, A.L., Zelenak, D., Reker, R.R., Rover, J., 2020. Lessons learned implementing an operational continuous United States national land change monitoring capability: the land change monitoring, assessment, and projection (LCMAP) approach. Remote Sens. Environ. Time Ser. Anal. High Spat. Resol. Imag. 238, 111356. https://doi.org/10.1016/j.rse.2019.111356.

Cai, Z., Hu, Q., Zhang, X., Yang, J., Wei, H., Wang, J., Zeng, Y., Yin, G., Li, W., You, L., Xu, B., Shi, Z., 2023. Improving agricultural field parcel delineation with a dual branch spatiotemporal fusion network by integrating multimodal satellite data. ISPRS J. Photogramm. Remote Sens. 205, 34–49. https://doi.org/10.1016/j.isprsjprs.2023.09.021.

Cai, Z., Xu, B., Yu, Q., Zhang, X., Yang, J., Wei, H., Li, S., Song, Q., Xiong, H., Wu, H., Wu, W., Shi, Z., Hu, Q., 2024. A cost-effective and robust mapping method for diverse crop types using weakly supervised semantic segmentation with sparse point samples. ISPRS J. Photogramm. Remote Sens. 218, 260–276. https://doi.org/10.1016/j.isprsjprs.2024.09.017.

Chen, Q., Gong, P., 2004. Automatic variogram parameter extraction for textural classification of the panchromatic IKONOS imagery. IEEE Trans. Geosci. Remote Sens. 42, 1106–1115. https://doi.org/10.1109/TGRS.2004.825591.

Chen, T.-H.K., Qiu, C., Schmitt, M., Zhu, X.X., Sabel, C.E., Prishchepov, A.V., 2020. Mapping horizontal and vertical urban densification in Denmark with Landsat time-series from 1985 to 2018: a semantic segmentation solution. Remote Sens. Environ. 251, 112096. https://doi.org/10.1016/j.rse.2020.112096.

Chen, L., Letu, H., Fan, M., Shang, H., Tao, J., Wu, L., Zhang, Y., Yu, C., Gu, J., Zhang, N., Hong, J., Wang, Z., Zhang, T., 2022. An introduction to the Chinese high-resolution earth observation system: Gaofen-1~7 civilian satellites. J. Remote Sens. 2022. https://doi.org/10.34133/2022/9769536.

Chen, T.-H.K., Kincey, M.E., Rosser, N.J., Seto, K.C., 2024. Identifying recurrent and persistent landslides using satellite imagery and deep learning: a 30-year analysis of the Himalaya. Sci. Total Environ. 922, 171161. https://doi.org/10.1016/j.scitotenv.2024.171161.

Cheng, T., Ji, X., Yang, G., Zheng, H., Ma, J., Yao, X., Zhu, Y., Cao, W., 2020. DESTIN: a new method for delineating the boundaries of crop fields by fusing spatial and temporal information from WorldView and planet satellite imagery. Comput. Electron. Agric. 178, 105787. https://doi.org/10.1016/j.compag.2020.105787.

Colditz, R.R., 2015. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. Remote Sens 7, 9655–9681. https://doi.org/10.3390/rs70809655.

Collins, L., McCarthy, G., Mellor, A., Newell, G., Smith, L., 2020. Training data requirements for fire severity mapping using Landsat imagery and random forest. Remote Sens. Environ. 245, 111839. https://doi.org/10.1016/j.rse.2020.111839.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Descals, A., Wich, S., Meijaard, E., Gaveau, D.L.A., Peedell, S., Szantoi, Z., 2021. High-resolution global map of smallholder and industrial closed-canopy oil palm plantations. Earth Syst. Sci. Data 13, 1211–1231. https://doi.org/10.5194/essd-13-1211-2021.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.

Du, S., Zhang, F., Zhang, X., 2015. Semantic classification of urban buildings combining VHR image and GIS data: an improved random forest approach. ISPRS J. Photogramm. Remote Sens. 105, 107–119. https://doi.org/10.1016/j.isprsjprs.2015.03.011.

Estes, L.D., Ye, S., Song, L., Luo, B., Eastman, J.R., Meng, Z., Zhang, Q., McRitchie, D., Debats, S.R., Muhando, J., Amukoa, A.H., Kaloo, B.W., Makuru, J., Mbatia, B.K., Muasa, I.M., Mucha, J., Mugami, A.M., Mugami, J.M., Muinde, F.W., Mwawaza, F. M., Ochieng, J., Oduol, C.J., Oduor, P., Wanjiku, T., Wanyoike, J.G., Avery, R.B., Caylor, K.K., 2022. High resolution, annual maps of field boundaries for smallholder-dominated croplands at National Scales. Front. Artif. Intell. 4.

Foody, G.M., Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sens. Environ. 103, 179–189. https://doi.org/10.1016/j.rse.2006.04.001.

Foody, G.M., Mathur, A., Sanchez-Hernandez, C., Boyd, D.S., 2006. Training set size requirements for the classification of a specific class. Remote Sens. Environ. 104, 1–14. https://doi.org/10.1016/j.rse.2006.03.004.

Garrigues, S., Allard, D., Baret, F., Morisette, J., 2008. Multivariate quantification of landscape spatial heterogeneity using variogram models. Remote Sens. Environ. 112, 216–230. https://doi.org/10.1016/j.rse.2007.04.017.

Grift, J., Persello, C., Koeva, M., 2024. CadastreVision: a benchmark dataset for cadastral boundary delineation from multi-resolution earth observation images. ISPRS J. Photogramm. Remote Sens. 217, 91–100. https://doi.org/10.1016/j.isprsjprs.2024.08.005.

He, H., Yan, J., Liang, D., Sun, Z., Li, J., Wang, L., 2024. Time-series land cover change detection using deep learning-based temporal semantic segmentation. Remote Sens. Environ. 305, 114101. https://doi.org/10.1016/j.rse.2024.114101.

Hertel, V., Chow, C., Wani, O., Wieland, M., Martinis, S., 2023. Probabilistic SAR-based water segmentation with adapted Bayesian convolutional neural network. Remote Sens. Environ. 285, 113388. https://doi.org/10.1016/j.rse.2022.113388.

Heydari, S.S., Mountrakis, G., 2018. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. Remote Sens. Environ. 204, 648–658. https://doi.org/10.1016/j.rse.2017.09.035.

Jin, H., Stehman, S.V., Mountrakis, G., 2014. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. Int. J. Remote Sens. 35, 2067–2081. https://doi.org/10.1080/01431161.2014.885152.

Jong, M., Guan, K., Wang, S., Huang, Y., Peng, B., 2022. Improving field boundary delineation in ResUNets via adversarial deep learning. Int. J. Appl. Earth Obs. Geoinf. 112, 102877. https://doi.org/10.1016/j.jag.2022.102877.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010.

Kim, J., Kim, D., Jun, H.-J., Heo, J.-P., 2024. The detection of residential developments in urban areas: exploring the potentials of deep-learning algorithms. Comput. Environ. Urban. Syst. 107, 102053. https://doi.org/10.1016/j.compenvurbsys.2023.102053.

Li, W., Dong, R., Fu, H., Wang, J., Yu, L., Gong, P., 2020. Integrating Google earth imagery with Landsat data to improve 30-m resolution land cover mapping. Remote Sens. Environ. 237, 111563. https://doi.org/10.1016/j.rse.2019.111563.

Li, M., Long, J., Stein, A., Wang, X., 2023. Using a semantic edge-aware multi-task neural network to delineate agricultural parcels from remote sensing images. ISPRS J. Photogramm. Remote Sens. 200, 24–40. https://doi.org/10.1016/j.isprsjprs.2023.04.019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.

Liu, S., Chen, L., Zhang, L., Hu, J., Fu, Y., 2023. A large-scale climate-aware satellite image dataset for domain adaptive land-cover semantic segmentation. ISPRS J. Photogramm. Remote Sens. 205, 98–114. https://doi.org/10.1016/j.isprsjprs.2023.09.007.

Loshchilov, I., Hutter, F., 2019. Decoupled Weight Decay Regularization. https://doi.org/10.48550/arXiv.1711.05101.

Lu, R., Zhang, Y., Huang, Q., Zeng, P., Shi, Z., Ye, S., 2024. A refined edge-aware convolutional neural networks for agricultural parcel delineation. Int. J. Appl. Earth Obs. Geoinf. 133, 104084. https://doi.org/10.1016/j.jag.2024.104084.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., Johnson, B.A., 2019. Deep learning in remote sensing applications: a meta-analysis and review. ISPRS J. Photogramm. Remote Sens. 152, 166–177. https://doi.org/10.1016/j.isprsjprs.2019.04.015.

Ma, Y., Chen, S., Ermon, S., Lobell, D.B., 2024. Transfer learning in environmental remote sensing. Remote Sens. Environ. 301, 113924. https://doi.org/10.1016/j.rse.2023.113924.

Masoud, K.M., Persello, C., Tolpekin, V.A., 2020. Delineation of agricultural field boundaries from Sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks. Remote Sens 12, 59. https://doi.org/10.3390/rs12010059.

Mellor, A., Boukir, S., Haywood, A., Jones, S., 2015. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. ISPRS J. Photogramm. Remote Sens. 105, 155–168. https://doi.org/10.1016/j.isprsjprs.2015.03.014.

Nguyen, L.H., Joshi, D.R., Clay, D.E., Henebry, G.M., 2020. Characterizing land cover/land use from multiple years of Landsat and MODIS time series: a novel approach using land surface phenology modeling and random forest classifier. Remote Sens.

Environ. Time Ser. Anal. High Spat. Resol. Imag. 238, 111017. https://doi.org/10.1016/j.rse.2018.12.016.
Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
Pan, Y., Wang, X., Zhang, L., Zhong, Y., 2023. E2EVAP: end-to-end vectorization of smallholder agricultural parcel boundaries from high-resolution remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 203, 246–264. https://doi.org/10.1016/j.isprsjprs.2023.08.001.
Pastorino, M., Moser, G., Serpico, S.B., Zerubia, J., 2022. Semantic segmentation of remote-sensing images through fully convolutional neural networks and hierarchical probabilistic graphical models. IEEE Trans. Geosci. Remote Sens. 60, 1–16. https://doi.org/10.1109/TGRS.2022.3141996.
Persello, C., Tolpekin, V.A., Bergado, J.R., de By, R.A., 2019. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sens. Environ. 231, 111253. https://doi.org/10.1016/j.rse.2019.111253.
Persello, C., Wegner, J.D., Hänsch, R., Tuia, D., Ghamisi, P., Koeva, M., Camps-Valls, G., 2022. Deep learning and earth observation to support the sustainable development goals: current approaches, open challenges, and future opportunities. IEEE Geosci. Remote Sens. Mag. 10, 172–200. https://doi.org/10.1109/MGRS.2021.3136100.
Piper, J., 1992. Variability and bias in experimentally measured classifier error rates. Pattern Recogn. Lett. 13, 685–692. https://doi.org/10.1016/0167-8655(92)90097-J.
Qurratulain, S., Zheng, Z., Xia, J., Ma, Y., Zhou, F., 2023. Deep learning instance segmentation framework for burnt area instances characterization. Int. J. Appl. Earth Obs. Geoinf. 116, 103146. https://doi.org/10.1016/j.jag.2022.103146.
Rajput, D., Wang, W.-J., Chen, C.-C., 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24, 48. https://doi.org/10.1186/s12859-023-05156-9.
Ramezan, C.A., Warner, T.A., Maxwell, A.E., Price, B.S., 2021. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. Remote Sens 13, 368. https://doi.org/10.3390/rs13030368.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
Stehman, S.V., Wickham, J.D., 2011. Pixels, blocks of pixels, and polygons: choosing a spatial unit for thematic accuracy assessment. Remote Sens. Environ. 115, 3044–3055. https://doi.org/10.1016/j.rse.2011.06.007.
Stehman, S.V., Mousoupetros, J., McRoberts, R.E., Næsset, E., Pengra, B.W., Xing, D., Horton, J.A., 2022. Incorporating interpreter variability into estimation of the total variance of land cover area estimates under simple random sampling. Remote Sens. Environ. 269, 112806. https://doi.org/10.1016/j.rse.2021.112806.
Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322. https://doi.org/10.1016/j.rse.2019.111322.
Turker, M., Kok, E.H., 2013. Field-based sub-boundary extraction from remote sensing imagery using perceptual grouping. ISPRS J. Photogramm. Remote Sens. 79, 106–121. https://doi.org/10.1016/j.isprsjprs.2013.02.009.
Turkoglu, M.O., D'Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J. D., 2021. Crop mapping from image time series: deep learning with multi-scale label hierarchies. Remote Sens. Environ. 264, 112603. https://doi.org/10.1016/j.rse.2021.112603.
Van Niel, T.G., McVicar, T.R., Datt, B., 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. Remote Sens. Environ. 98, 468–480. https://doi.org/10.1016/j.rse.2005.08.011.
Waldner, F., Diakogiannis, F.I., 2020. Deep learning on edge: extracting field boundaries from satellite images with a convolutional neural network. Remote Sens. Environ. 245, 111741. https://doi.org/10.1016/j.rse.2020.111741.
Wieland, M., Martinis, S., Kiefl, R., Gstaiger, V., 2023. Semantic segmentation of water bodies in very high-resolution satellite and aerial images. Remote Sens. Environ. 287, 113452. https://doi.org/10.1016/j.rse.2023.113452.
Xiong, S., Zhang, X., Lei, Y., Tan, G., Wang, H., Du, S., 2024. Time-series China urban land use mapping (2016–2022): an approach for achieving spatial-consistency and semantic-transition rationality in temporal domain. Remote Sens. Environ. 312, 114344. https://doi.org/10.1016/j.rse.2024.114344.
Yang, J., Huang, X., 2021. The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019. Earth Syst. Sci. Data 13, 3907–3925. https://doi.org/10.5194/essd-13-3907-2021.
Yu, H., Yang, Z., Tan, L., Wang, Y., Sun, W., Sun, M., Tang, Y., 2018. Methods and datasets on semantic segmentation: a review. Neurocomputing 304, 82–103. https://doi.org/10.1016/j.neucom.2018.03.037.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens. Environ. 241, 111716. https://doi.org/10.1016/j.rse.2020.111716.
Yuan, X., Shi, J., Gu, L., 2021. A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Syst. Appl. 169, 114417. https://doi.org/10.1016/j.eswa.2020.114417.
Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019. Joint deep learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187. https://doi.org/10.1016/j.rse.2018.11.014.
Zhang, Q., Zhang, Z., Xu, N., Li, Y., 2023. Fully automatic training sample collection for detecting multi-decadal inland/seaward urban sprawl. Remote Sens. Environ. 298, 113801. https://doi.org/10.1016/j.rse.2023.113801.
Zhang, X., Zhao, T., Xu, H., Liu, W., Wang, J., Chen, X., Liu, L., 2024. GLC_FCS30D: the first global 30 m land-cover dynamics monitoring product with a fine classification system for the period from 1985 to 2022 generated using dense-time-series Landsat imagery and the continuous change-detection method. Earth Syst. Sci. Data 16, 1353–1381. https://doi.org/10.5194/essd-16-1353-2024.
Zhao, W., 2017. Research on the deep learning of the small sample data based on transfer learning. AIP Conf. Proc. 1864, 020018. https://doi.org/10.1063/1.4992835.
Zhao, W., Lyu, R., Zhang, Jinming, Pang, J., Zhang, Jianming, 2024. A fast hybrid approach for continuous land cover change monitoring and semantic segmentation using satellite time series. Int. J. Appl. Earth Obs. Geoinf. 134, 104222. https://doi.org/10.1016/j.jag.2024.104222.
Zhao, H., Wu, B., Zhang, M., Long, J., Tian, F., Xie, Y., Zeng, H., Zheng, Z., Ma, Z., Wang, M., Li, J., 2025. A large-scale VHR parcel dataset and a novel hierarchical semantic boundary-guided network for agricultural parcel delineation. ISPRS J. Photogramm. Remote Sens. 221, 1–19. https://doi.org/10.1016/j.isprsjprs.2025.01.034.
Zhou, Y., Weng, Q., 2024. Building up a data engine for global urban mapping. Remote Sens. Environ. 311, 114242. https://doi.org/10.1016/j.rse.2024.114242.
Zhou, Q., Tollerud, H., Barber, C., Smith, K., Zelenak, D., 2020. Training data selection for annual land cover classification for the land change monitoring, assessment, and projection (LCMAP) initiative. Remote Sens 12, 699. https://doi.org/10.3390/rs12040699.
Zhu, Z., Gallant, A.L., Woodcock, C.E., Pengra, B., Olofsson, P., Loveland, T.R., Jin, S., Dahal, D., Yang, L., Auch, R.F., 2016. Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative. ISPRS J. Photogramm. Remote Sens. 122, 206–221. https://doi.org/10.1016/j.isprsjprs.2016.11.004.
Zhu, W., Braun, B., Chiang, L.H., Romagnoli, J.A., 2021. Investigation of transfer learning for image classification and impact on training sample size. Chemom. Intell. Lab. Syst. 211, 104269. https://doi.org/10.1016/j.chemolab.2021.104269.