Remote Sensing of Environment 311 (2024) 114274

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

China's first sub-meter building footprints derived by deep learning

Xin Huang, Zhen Zhang, Jiayi Li *

School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China

A R T I C L E I N F O

Editor: Marie Weiss

Keywords:
Building footprint
Large-scale
Datasets
Deep learning
Semi-supervised learning
Large-scale mapping
Google

A B S T R A C T

The high spatial resolution building footprints are crucial for understanding urban development and its associated applications. However, up to now, the sub-meter-level building footprint data of China is still lacking. The challenges arise from two aspects: 1) the number of training samples is inadequate for large-scale building extraction. 2) the accuracy and efficiency of current models are insufficient to conduct large-scale building extraction. Therefore, we propose a framework for large-scale building extraction in this study, including semi-automated sample generation, building extraction model, model training, and post-processing. Specifically, the main technical contributions include: 1) BldgNet (Building Extraction Network) is proposed, including the Large Window Attention, Edge Attention, and Distribution Alignment Module with consideration of spatial contextual information, to address the challenge of the multi-scale building extraction, building boundary delineation, and class imbalance, respectively; 2) a semi-supervised training approach is proposed for large-scale building extraction, leveraging the incomplete information from OpenStreetMap (OSM) to enhance the diversity of building samples and the robustness of the model. Meanwhile, we created an open-source Global Building Dataset (GBD) comprising approximately 800,000 high-resolution (0.25 m) samples. This dataset incorporates diverse building styles worldwide, offering support for global building extraction. Based on the constructed sample set and the proposed deep net, we generated China's first sub-meter (0.5 m) building footprint dataset (CBF). Through testing on 750,000 buildings from 350 cities, the overall F1 score for CBF reached 83.71%. Finally, we validated that the proposed building extraction model can achieve satisfactory results compared to existing representative deep networks. GBD and CBF datasets can be publicly available and downloadable via https://zenodo.org/doi/10.5281/zenodo.10043351.

1. Introduction

The high spatial resolution building footprint data is essential for understanding urban development and its related applications. Building footprints and their extent are important indicators for human activities (Huang et al., 2021), sustainable urbanization(Appolloni et al., 2021), building energy modeling(Byrne et al., 2015), and urban planning (Nadal et al., 2017). Furthermore, building information is related to urban energy usage(Resch et al., 2016) and greenhouse gas emissions (Borck, 2016; Marconcini et al., 2020).

With the development of sensors as well as processing techniques, increasing remote sensing and geospatial data have become available for open-source use. As a result, various large-scale building footprints, building heights, and related products are released. We can categorize these studies into three types. The first one focuses on extracting impervious surfaces or human settlement footprints. The characteristics of such products are coarse resolution (10 m–30 m) but with a long time span. For example, GISA(Huang et al., 2021), GAIA(Gong et al., 2020), and GAUD(Liu et al., 2020) utilized Landsat imagery at 30 m resolution to extract global impervious surfaces from 1972 to 2019, 1985–2018, and 1985–2015, respectively. WSF(Marconcini et al., 2020) employed Landsat 8 and Sentinel-1 imagery to extract the 10-m global human settlement footprints for 2015. GISA-10 m(Huang et al., 2022) used Sentinel-1 and Sentinel-2 data to capture global 10-m resolution impervious surface for the year of 2016. GHSL(Corbane et al., 2021) employed Sentinel-2 data to extract the 10-m resolution global human settlement footprints for 2018. However, owing to their limitations in spatial resolution, such products cannot provide information on individual buildings.

The second type of products focuses on urban building heights (Cao and Huang, 2021; P. Chen et al., 2023a; Frolking et al., 2022; Ma et al., 2023). These products often have coarse resolutions (0.3 km–1 km) and typically cover only one year. For instance, (Xuecao Li et al., 2020a) released 500 m resolution building height data for major U.S. cities.

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, PR China.
E-mail address: zjjerica@whu.edu.cn (J. Li).

https://doi.org/10.1016/j.rse.2024.114274
Received 28 January 2024; Received in revised form 18 May 2024; Accepted 14 June 2024

Table 1
The comparison of relevant product information.

Dataset	Scale	Time span	Resolution	Type
GISA(Huang et al., 2021)	Global	1972–2019	30 m	I
GAIA(Gong et al., 2020)	Global	1985–2018	30 m	I
GAUD(Liu et al., 2020)	Global	1985–2015	30 m	I
WSF(Marconcini et al., 2020)	Global	2015	10 m	I
GISA-10 m(Huang et al., 2022)	Global	2016	10 m	I
GHSL(Corbane et al., 2021)	Global	2018	10 m	I
(Xuecao Li et al., 2020a)	USA	2015	500 m	II
(Yang and Zhao, 2022)	China	2017	1000 m	II
(Zhou et al., 2022)	Global	2015	500 m	II
(Frantz et al., 2021)	Germany	2017	10 m	II
CNBH(Wu et al., 2023)	China	2020	10 m	II
Microsoft(Microsoft, 2023)	Global (not including China)	No specific time	<1 m	III
Google(Sirko et al., 2021)	Africa	No specific time	0.5 m	III
90-cities-BRA(Z. Zhang et al., 2022b)	90 cities in China	2020	1 m	III
CBRA(Liu et al., 2023)	China	2016–2021	2.5 m	III
CBF(ours)	China	2019	0.5 m	III

(Yang and Zhao, 2022) provided urban height data for China at a 1 km resolution. (Zhou et al., 2022) published global building height data for 2015 at a 500 m resolution. More recently, (Frantz et al., 2021) and (Wu et al., 2023) offered more accurate (10 m) building height data for Germany and China, respectively. Similarly, products of this type also encounter challenges in describing building information at a fine-grained level.

The third type primarily utilizes high-resolution data for extracting building footprints. Microsoft(Microsoft, 2023) released building footprints at 1 m resolution for specific global regions(year not specified). Google(Sirko et al., 2021) provided building footprints for Africa at a 0.5 m resolution (year not specified). Note that these products do not include building footprint information of China. These high-resolution products are generated using deep learning methods, requiring substantial training data and computational resources. More importantly, their reuse in other regions (such as in China) is difficult due to the lack of open-sourcing of the training data and methods. Recently, (Z. Zhang et al., 2022b) released 1 m resolution building footprint data for 90 cities in China for 2020. However, this study did not provide information in other cities and rural areas. (Liu et al., 2023) released building footprint information for China from 2016 to 2021. They employed a super-resolution segmentation method to obtain 2.5 m results based on Sentinel-2 data (10 m). Hence, a considerable gap exists between their products and sub-meter-level results, inevitably leading to incorrect adhesion or omissions of buildings. As of now, finely detailed (e.g., sub-meter-level) building footprint data of China is still lacking.

In this study, we utilized Google imagery during 2019 to 2020 to extract China's 0.5 m resolution building footprint data (CBF), aiming to address the absence of sub-meter-level building footprint data of China. Large-scale building extraction is a challenging task. Currently available high-resolution datasets for building extraction are often limited to smaller areas. Additionally, open-source data such as OpenStreetMap (OSM) has very low completeness in China (<9%)(Herfort et al., 2023), making the collection of training samples more difficult. Moreover, there is a lack of research on extracting building footprints from a massive amount of high-resolution imagery while considering both efficiency and accuracy of algorithms. To overcome these challenges, we created a global building sample dataset (GBD) using OSM and Google imagery. This dataset is widely distributed and can fulfill global building extraction sample requirements. Based on this, we propose a deep-learning approach suitable for large-scale building extraction, which can effectively balance accuracy and efficiency. In summary, the main contributions of this study are as follows:

1) Developed China's first open-source 0.5 m resolution building footprint dataset (CBF).
2) Provided an open-source global building sample dataset (GBD, resolution 0.25 m). This dataset comprises approximately 800,000 images with diverse architectural styles worldwide, each with a size of 512 × 512 pixels. It can be served as training and test samples for building extraction in different regions globally.
3) Proposed a practicable building extraction method for large-scale high-resolution imagery. The algorithm focuses on the current technical bottlenecks for the deep learning based building extraction, e.g., collection of diverse samples, inaccurate building boundaries, tiny building omissions, and foreground-background imbalance.

The remainder of this paper is arranged as follows. Sec. 2 summarizes the building-related datasets and describes relevant building extraction methods. Sec. 3 presents the data used in the study. Sec. 4 details the proposed methodology. In Sec. 5, we quantitatively evaluate and analyze the CBF dataset. Sec. 6 involves a comparison between CBF and other products, followed by a series of discussions. Finally, Sec. 7 summarizes the findings of this research.

2. Background

2.1. Building-related products

As mentioned above, we categorize products related to buildings or containing building information into three types. The specific parameters for each product are shown in Table 1. Type I products primarily provide information on impervious surfaces or human settlement footprints. These products typically use Landsat imagery as their data sources. These products often have a long temporal span and large geographical extent but relatively coarse spatial resolution. More recently, some researchers use Sentinel imagery as a data source, and increase the spatial resolution to 10 m, albeit at the cost of losing temporal span. Type II products focus on building height information, but most only cover a specific country or region in a single year. Type III products provide footprints for individual buildings, thus requiring high-resolution remote sensing data. It should be noted that the accurate prediction of individual building footprints is relied on deep learning, requiring a substantial amount of high-resolution imagery, extensive training data, efficient building extraction methods, and considerable computational resources. Therefore, the individual building extraction from high-resolution imagery is much more difficult. So far, sub-meter resolution building footprint data for China remains unavailable.

2.2. Building extraction methods

As mentioned above, the challenges of large-scale building extraction include both detection methods and sample collection. Therefore, here we describe in detail the recent advances in the two aspects.

Deep learning has made significant breakthroughs and achievements in various fields, such as natural language processing (Doveh et al., 2023; Yin et al., 2023), computer vision (Huang et al., 2023; Wang et al., 2022), medical image analysis (Huang et al., 2023; Wang et al., 2022), autonomous driving (Huang et al., 2023; Wang et al., 2022). In the building extraction, deep learning algorithms are continuously evolving. Researchers aim to enhance the performance by focusing on various modules of the deep networks. The advances include local and global information interaction (Kang et al., 2019; Zhao et al., 2017) to improve

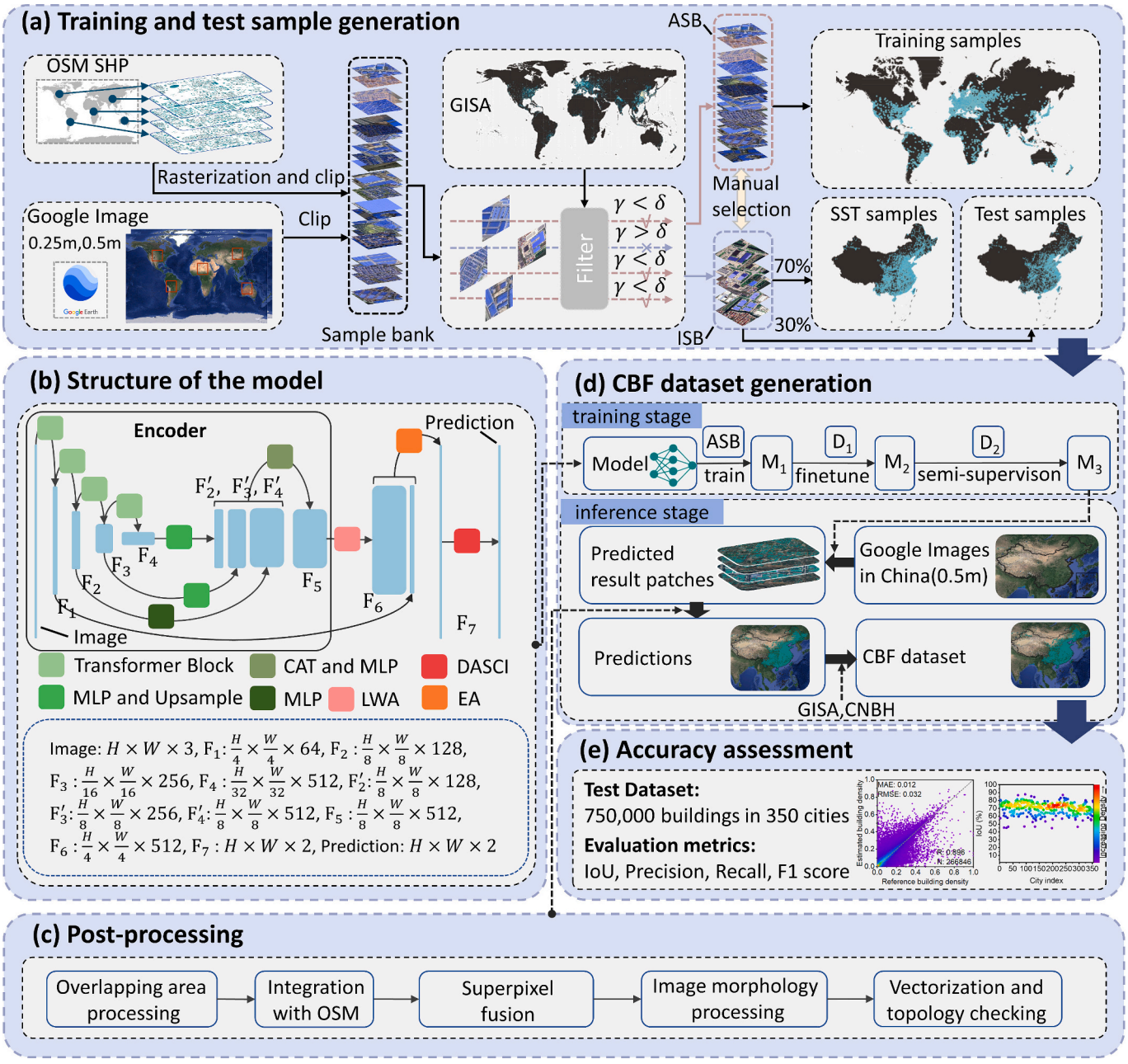

Fig. 1. The overall flowchart of the proposed framework. (a) Training and test sample generation. (b) Structure of the model. (c) Post-processing. (d) CBF dataset generation. (e) Accuracy assessment.

the model's information acquisition; attention mechanisms (Howard et al., 2019; Hu et al., 2018; Tan and Le, 2019) for efficient self-attention; multiscale feature representation (L. Li et al., 2018b; Wu et al., 2018; Yan et al., 2022) to strengthen the model's ability to acquire global information; and edge preservation (He et al., 2021; Xiangtai Li et al., 2020; Pu et al., 2022) to enhance the model's accuracy in the boundaries. While these methods improve the model's performance to some extent, the lack of consideration in the lightweight design makes them difficult to meet the requirement for large-scale mapping, which is a complex and comprehensive task, encompassing sample acquisition, pre- and post-processing. However, a single model design alone cannot support large-scale building extraction. While Microsoft (Microsoft, 2023) and Google (Sirko et al., 2021) have released the large-scale building extraction results, their technical details have not been public yet. Furthermore, they have not provided the building extraction results of China. With respect to the samples, the existing studies typically use publicly available datasets (Ji et al., 2019; Maggiori et al., 2017) for model training and accuracy validation. Nevertheless, the current building datasets are often small in scale and homogeneous in style. It should be noted that the diversity of training data is crucial for performances of the data-driven deep learning. One approach to addressing the large-scale sample collection is to reduce the training sample requirements and enhance the model's generalization capability through the weakly supervised methods (Shen et al., 2023). For instance, (Ahn and Kwak, 2018; Ge et al., 2019; Shen et al., 2021; Y. Wang et al., 2020d; Zhou et al., 2018) employed image-level labels for supervision and achieve pixel-level segmentation. Similarly, (Hsu et al., 2019; Khoreva et al., 2017; Q. Li et al., 2018a; Oh et al., 2021) used box-level labels for network training and the subsequent pixel-level semantic segmentation. (Lin et al., 2016) supervised the network using scribble-level labels. However, it should be noted that these weakly supervised methods still rely to some extent on the quantity and quality of samples, and their

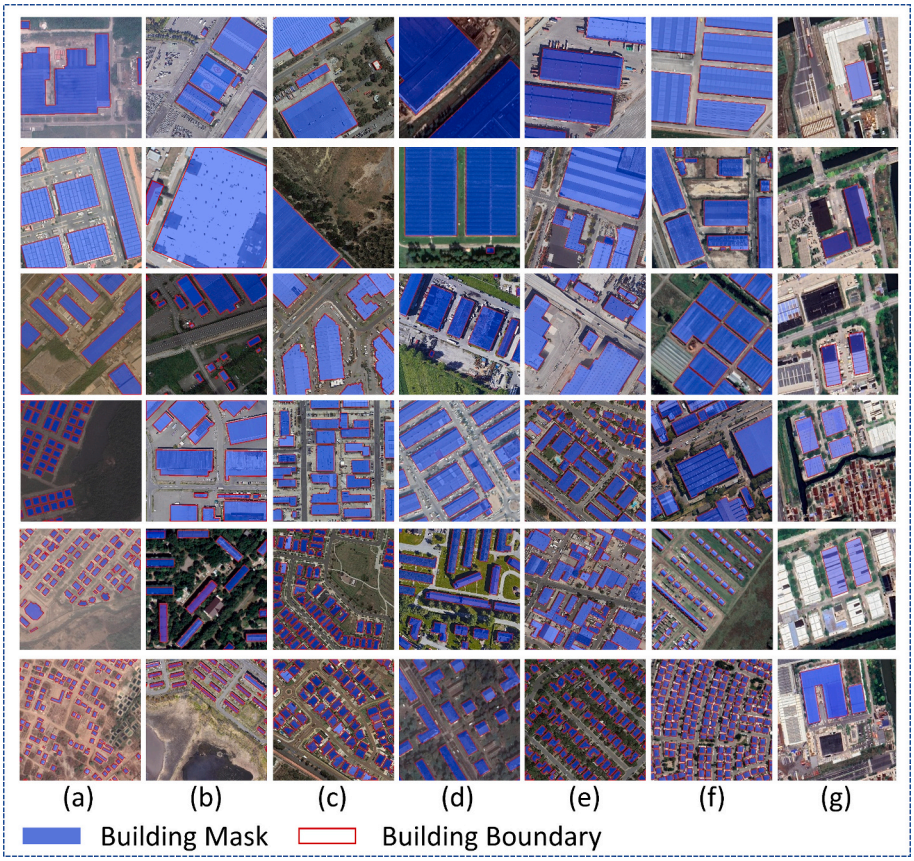

Fig. 2. Samples (a) to (f) represent instances within the ASB dataset, where (a) corresponds to Africa, (b) Asia, (c) Oceania, (d) Europe, (e) North America, and (f) South America. (a) ~ (f) are samples in the training dataset (ASB). (g) displays the ISB samples. The colour of Blue indicates building masks, while red indicates building boundaries. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

accuracy lags behind fully supervised methods. Therefore, how to generate accurate and diverse sample sets for large-scale building extraction deserves in-depth study.

3. Data

3.1. Google high resolution imagery

We collected images with resolutions of 0.25 m and 0.5 m from the Google images. To enhance the model's ability for multi-scale modeling, we employed both resolutions (0.25 and 0.5 m) as the data sources for training samples. During model predictions, only 0.5 m resolution images of China were used for mapping. The acquisition time of these images is approximately between 2019 and 2020. Due to their diverse sources, the images exhibit a variety of imaging conditions, e.g., different seasons, angles, and tones. To mitigate the impact of colour variations, we applied standardization and normalization techniques to preprocess the images (Conn and Arandjelovic, 2017).

3.2. OpenStreetMap

We utilized publicly available building information from OpenStreetMap (OSM) to generate labels for training and testing datasets. To ensure the accuracy of the training data, manual inspection and corrections were conducted. Moreover, there was no overlap between the training and testing datasets, and the evaluation was performed in a blinded manner. (See Section 4.1 for details of sample production).

3.3. GISA

GISA (Huang et al., 2021) provides a global impervious surface area map from 1972 to 2019, with an F1 score of 0.954 (in terms of a large number of randomly selected and third-party validated sample sets). GISA is generated by a novel global ISA mapping method that incorporates semi-automatic global sample collection, a locally adaptive classification strategy, and a spatio-temporal post-processing procedure. Furthermore, GISA is extracted from the entire global land area, rather than from an urban mask, thereby reducing underestimation. In this study, GISA is used for sample filtering (See Section 4.1 for details).

4. Method

The proposed framework consists of a series of key techniques for the large-scale and high-resolution building extraction (Fig. 1), including: (a) Training and test sample generation; (b) BldgNet: a deep network specialized for large-scale building extraction (denoted as BldgNet); (c) Post-processing; (d) CBF dataset generation; and (e) Accuracy assessment. The details of these modules are described in the following sections.

4.1. Training and test sample generation

The training samples are pivotal for data-driven deep-learning algorithms. Accurate sample sets are typically annotated manually, which often consumes a significant amount of manual labour and time. In this study, we proposed a semi-automated approach to generate samples, delegating only key steps to manual processing. This method ensures sample accuracy while significantly reducing the manual workload. The

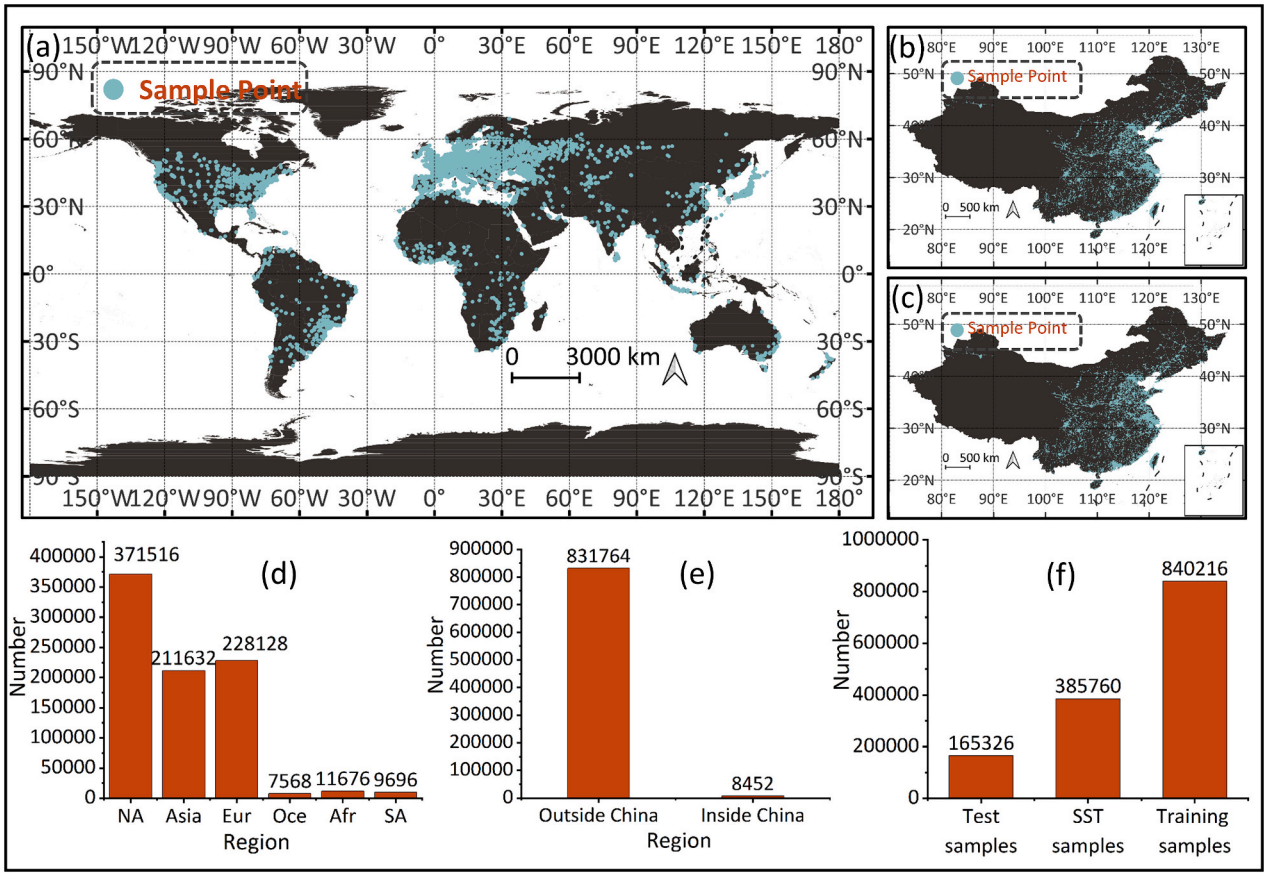

Fig. 3. (a) The sample distribution of the ASB (training dataset). (b) The sample distribution of the semi-supervised training dataset. (c) The sample distribution of the test dataset. (d) A comparison of sample quantities across continents within the ASB. (e) A comparison of sample quantities within and outside the Chinese region in the ASB. (f) A comparison of sample quantities across various datasets.

specific process is illustrated in Fig. 1(a). Initially, we rasterized the global building information from OSM (OpenStreetMap) and then cropped it along with Google images (0.25 m and 0.5 m) into image pairs of 512 × 512 pixels, forming a sample bank. Due to the varied quality of OSM annotations, primarily the issue of omissions, we chose the samples by considering the correlation between impervious surface density and building density. Specifically, we calculated the impervious surface density α and building density β as well as their difference, denoted as γ, for each image. α is calculated by dividing the impervious surface area (obtained from GISA) within each 512 × 512 patch by the total area of the patch, and β is computed by dividing the building area (obtained from OSM) within the patch by the total area of the patch. Image pairs with γ exceeding a threshold δ were identified as building omissions. In this way, the original sample bank is divided into the Accurate Sample Bank (ASB) and the Incomplete Sample Bank (ISB). As shown in Fig. 2, samples (a) to (f) belong to ASB with more complete annotations, while sample (g) belongs to ISB with higher omissions. Finally, we visually inspected and modified samples in ASB to eliminate misclassifications and incorrect annotations, further enhancing ASB's accuracy. Simultaneously, we manually filtered and revised samples in ISB to ensure that the existing building annotations were accurate. It should be noted that despite omissions in ISB's building samples, in this study, we adopted a semi-supervised training approach (detailed in Section 4.2.5), leveraging the building information from ISB to enhance the training of the BldgNet.

Figure 3(a) shows the distribution of ASB samples. ASB contains approximately 840,000 samples from various regions globally. It encompasses diverse architectural styles that can provide potential samples for large-scale or even global building extraction. Fig. 2(a)~(f) showcases building samples from different continents within ASB. We utilized ASB as the training dataset for our model. Fig. 3(d) presents the statistics of ASB samples across continents: 211,632 in Asia, 228,128 in Europe, 7568 in Oceania, 11,676 in Africa, 9696 in South America, and 371,516 in North America. As depicted in Fig. 3(e), we partitioned this training dataset (ASB) into two parts: the first part consists of samples outside the study area (China), totaling 831,764 samples. The second one is the samples within China, approximately 8400 samples, denoted as D_1. The small number of samples in China can be attributed to the lack of building datasets in China (Sun et al., 2023), which underscores the necessity and significance of our study. To fully leverage OSM's buildings samples and enhance the model's performance within the study area, we designed a semi-supervised training approach capable of utilizing the samples with incomplete annotations, such as those in ISB, to improve the model's generalization in the study area (elaborated in Section 4.2.5). Notice that the existing building annotations in ISB have been manually corrected, to maintain a high level of accuracy despite omissions in ISB. Subsequently, 70% of the ISB samples were randomly chosen and used to construct the semi-supervised training dataset (Fig. 3 (b)), designated as D_2. The remaining 30% was used as the test dataset (Fig. 3(c)) to evaluate the mapping results, containing 750,000 buildings from 350 cities in China. Fig. 3(f) displays the sample counts for each dataset, with no overlap between the test, the training and semi-supervised training datasets.

4.2. The structure of BldgNet

4.2.1. Overview

CNN (convolutional neural network) and Transformer are the two

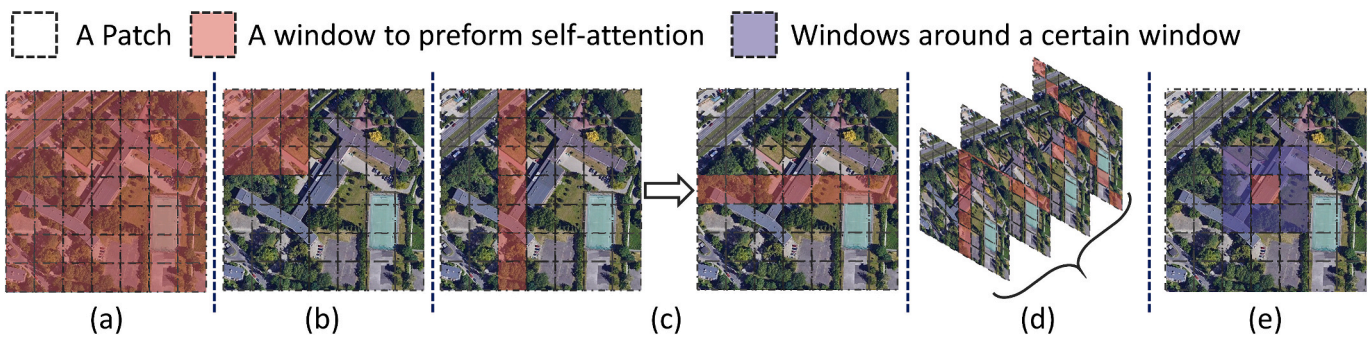

Fig. 4. Demonstration of feature map partitioning for different attention methods. (a) Full self-attention. (b) Shifted local self-attention. (c) Sequential axial self-attention. (d) Directional window self-attention. (e) Cross-attention calculation between the current window and the surrounding windows.

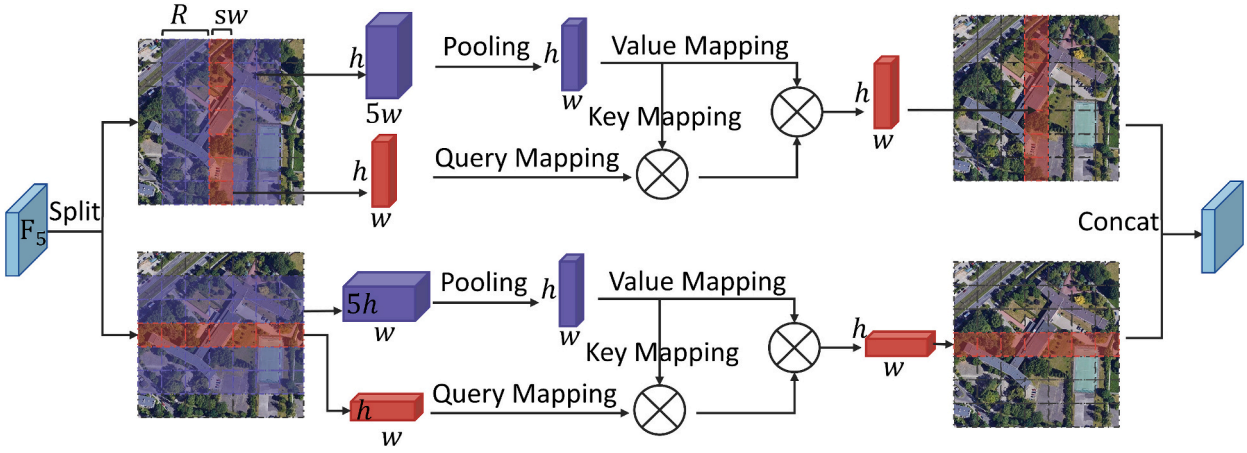

Fig. 5. The structure of LWA.

predominant architectures in deep learning models. Some researchers have also proposed novel structures like multilayer perceptron (Tolstikhin et al., 2021), deep learning clustering (Xu Ma et al., 2023), and deep learning support vector machines (Tarzanagh et al., 2023). However, with the increasing volume of data, models based on the Transformer architecture have exhibited superior capabilities in fitting large datasets compared to other ones. For tasks such as building extraction from large-scale high-resolution remote sensing images, the efficiency of models is crucial due to the substantial amount of data (labels and images). Simultaneously, large-scale building extraction poses unique challenges, including diverse building sizes, difficulties in precise delineation of building boundaries, and foreground-background imbalances. To deal with these challenges, we propose the BldgNet (the Buildings Extraction Network), a deep learning network based on the Transformer architecture specifically tailored for large-scale building extraction. The overall structure of the model is shown in Fig. 1(b). The image is first passed through an encoder (see Section 4.2.2) to extract and fuse the multi-level features. The fused feature (F_5) is fed into the Large Window Attention (LWA) module (see Section 4.2.3), for improving the contextual modeling capabilities and capturing global information to enhance the performance of extracting buildings with different sizes. Its output is recorded as F_6. Subsequently, the Edge Attention (EA) Module is performed on F_1 and F_6, to strengthen the model's delineation to building boundaries (see Section 4.2.4). Finally, the Distribution Alignment Module with Consideration of Spatial Contextual Information (DASCI) is used to ameliorate the model accuracy degradation caused by imbalanced foreground-background ratios (see Section 4.2.5). Additionally, our proposed semi-supervised training method is detailed in Section 4.3.

4.2.2. Encoder

The encoder structure is shown in Fig. 1(b). The image data ($H \times W \times 3$) first undergoes token embedding through a Transformer Block (Dosovitskiy et al., 2021), resulting in a feature map with dimensions $\frac{H}{4} \times \frac{W}{4} \times 64$, denoted as F_1. Subsequently, three additional Transformer Blocks are adopted for feature extraction, with each Transformer Block reducing the size of the feature map by half and doubling the number of channels. Therefore, in sequence, the output feature map sizes of the Transformer Blocks are $\frac{H}{8} \times \frac{W}{8} \times 128$, $\frac{H}{16} \times \frac{W}{16} \times 256$, $\frac{H}{32} \times \frac{W}{32} \times 512$, denoted as F_2, F_3, and F_4, respectively. Subsequently, these multi-scale features (F_2, F_3, F_4) are processed by a multilayer perceptron (MLP) (Tolstikhin et al., 2021) and are upsampled to the same size. They are then stacked along the channel dimension and fused using an MLP, resulting in the fused feature denoted as F_5.

4.2.3. The LWA module

The attention mechanism is the core of the Transformer architecture. The computation of full self-attention (Fig. 4(a)) increases proportionally with the square of image dimensions. Considering the affordability of computational resources, many researchers divide the feature map into specific windows (as illustrated in Fig. 4(b) and Fig. 4(c)), with the attention computation performed simultaneously and independently within each window (Dong et al., 2021; Ho et al., 2019; Liu et al., 2021; Zhang et al., 2023). To enhance information exchange between different windows, shifted local self-attention (Fig. 4(b)) performs convolution or the self-attention calculations again by utilizing the windows that are partially overlapped with the preceding windows. Besides, sequential axial self-attention (Fig. 4(c)) conducts self-attention calculations sequentially using horizontal and vertical windows. These approaches improve the computational efficiency of the attention mechanism and

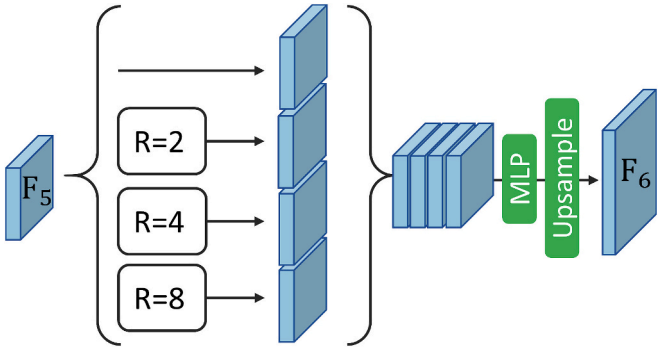

Fig. 6. The overall structure of the LWA module.

enhance the model's accuracy. (Zhang et al., 2023) investigated the impact of different window shapes and window overlapping methods of the attention mechanism on model accuracy and efficiency. The proposed directional window self-attention (DWSA) (Fig. 4(d)) outperformed other attention computation methods in both speed and accuracy. DWSA divides the feature map into n segments along the channel dimension, with each segment performing self-attention calculations in different directions. However, this indirect information exchange among windows can limit the contextual modeling capabilities and the capture of global information. In this study, this limitation hinders the model from effectively capturing super-sized buildings or densely connected residential areas. In contrast to indirect information exchange, (Yan et al., 2022) employed direct cross-attention calculations between the current and surrounding windows (Fig. 4(e)) to enhance information exchange among different windows. However, relying on a single mode of exchange (indirect or direct) still falls short of effectively representing contextual information.

Therefore, in this study, we propose the Larger Window Attention (LWA) Module, which integrates the two modes (both indirect and direct) of information exchange between windows. This approach ensures that the self-attention scope along different directional paths overlaps spatially (indirect information exchange) and enables each window to obtain information from surrounding windows through cross-attention calculations (direct information exchange). This enhancement facilitates the model's capability of contextual modeling and global information acquisition. The structure of the LWA is illustrated in Fig. 5, where the feature F_5 ($\frac{H}{8}\times\frac{W}{8}\times 512$) is divided into n segments along the channel dimension ($\frac{H}{8}\times\frac{W}{8}\times\frac{512}{n}$). Each segment concurrently performs strip-wise attention in different directions. Considering the computational complexity, n is set as 2 in this study, representing the vertical and horizontal directions. This partitioning method improves computational speed while enlarging the attention scope along different directions. Taking the vertical direction as an example, by dividing the feature map ($\frac{H}{8}\times\frac{W}{8}\times 256$) into $\frac{W}{8\times sw}$ patches, the dimension of each patch is $\frac{H}{8}\times sw\times 256$, where sw is the strip-wise window width. The value of sw can be adjusted to balance accuracy and efficiency. To further expand the attention scope and enhance contextual information retrieval for each window in both directions, each patch (taking the red window in the figure as an example) performs attention calculations with patches within a radius of R (depicted by the purple window in the figure). Subsequently, the feature maps from both directions are stacked along the channel dimension to restore the original dimensions ($\frac{H}{8}\times\frac{W}{8}\times 512$).

Figure 5 illustrates the case of $R = 2$. Employing window attention with various R sizes, i.e., spatial pyramid pooling, allows for further extraction of global information (Yan et al., 2022). Therefore, in this study, we consider multiple R values (depicted in Fig. 6). Following the application of window attention at different radii, these feature maps are stacked with the original feature map, which are then fused with a Multilayer Perceptron (MLP). Subsequently, upsampling is performed to further restore the spatial information.

4.2.4. The EA module

The commonly adopted Hierarchical structure design in neural networks leads to the gradual loss of spatial details, especially for edge-related information. This can also lead to large uncertainty in predicting building edges. Many researchers (e.g., (He et al., 2021; Xiangtai Li et al., 2020)) address this issue by introducing Edge branches or Edge Processing Modules (EPM) to achieve more precise building edge predictions. Using Edge branches (as shown in Fig. 7(a1)) can increase the model's parameter and computation complexity, particularly in the structures with multiple edge branches (He et al., 2021). This approach hampers the model's speed, making it unsuitable for large-scale building extraction. Similarly, edge pixels constitute only a small portion of the image, and Edge Processing Modules (as illustrated in Fig. 7(a2)) utilize the entire feature map for computation, thereby increasing the computational effort of the model. In our research, to guide the model's focus on building edges without introducing additional computation, we propose the Edge Attention (EA) module (Fig. 7(a3)). This method performs feature extraction only on edge pixels, resulting in improved edge prediction with lower computational complexity. The edge pixels

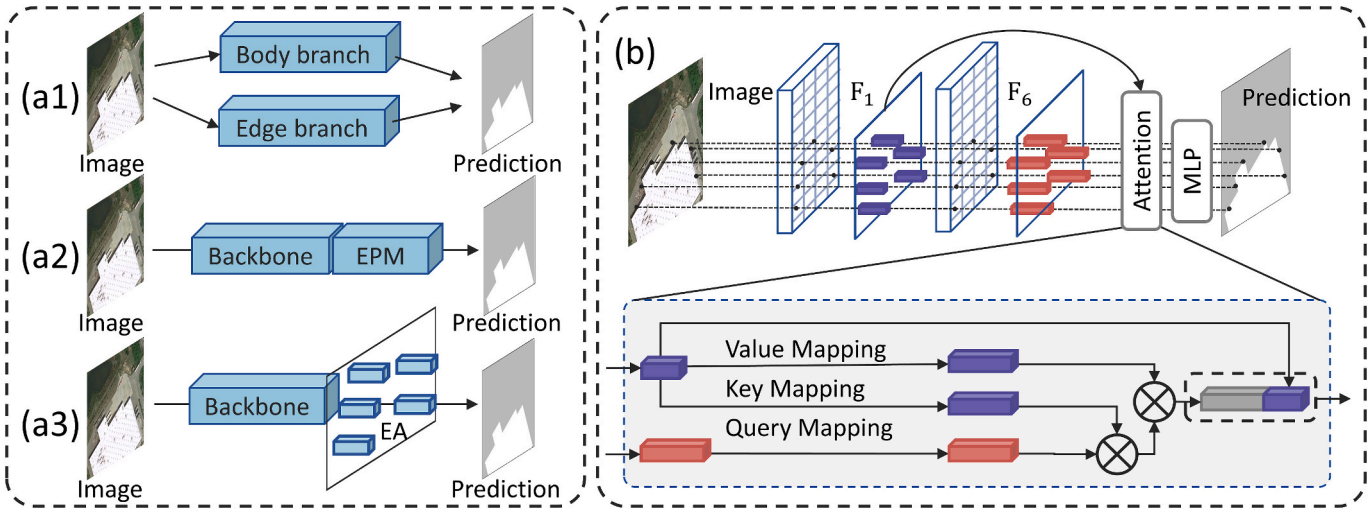

Fig. 7. (a1) Structure of the model with the edge branch. (a2) Structure of the model with the edge processing module. (a3) Structure of the model incorporating the proposed EA module. (b) EA module structure.

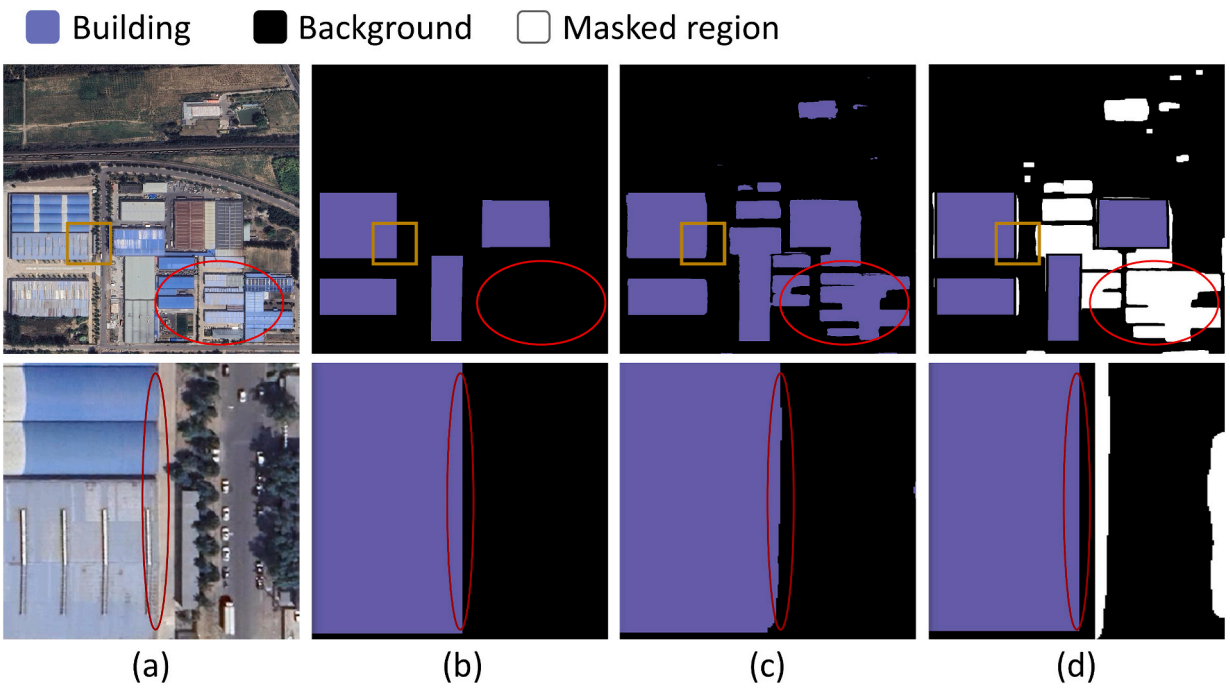

Fig. 8. (a) is the Google imagery, (b) is the sample from D_2 (70% of ISB), (c) is the prediction of the network (BldgNet) trained on ASB data, and (d) illustrates the label after masking by the semi-supervised training. The second row represents the enlarged view of the yellow-boxed region in the first row. The red circles in the second row demonstrate the labels at the edges.

are identified based on the absolute difference in prediction probabilities between the two categories (building and non-building). Pixels with small differences are typically located at or near edges (Kirillov et al., 2020). Therefore, we consider n points with the smallest differences in the EA module. Specifically, as depicted in Fig. 7(b), the EA module utilizes an attention mechanism (Dosovitskiy et al., 2021) to fuse information from the edge pixels in the initial feature map F_1 (containing more spatial details) with the corresponding pixels in the feature map F_6 (containing more semantic information) obtained after LWA. Subsequently, the output is stacked with the edge pixels from F_1, which are then predicted using an MLP for each point. During training, a cross-entropy loss is used to supervise the prediction of these points.

4.2.5. The DASCI module

The imbalance between the number of building and non-building classes can lead to biased decision boundaries of the network. Therefore, it is essential to implement class-balancing strategies to improve the model's performance. The efforts to mitigate the adverse effects of long-tailed class distributions (i.e., class imbalance) can be categorized into two types: One-stage Imbalance Learning and Two-stage Imbalance Learning. One-stage Imbalance Learning includes resample-based methods (Buda et al., 2018; Chawla et al., 2002; Han et al., 2005; Mahajan et al., 2018; Shen et al., 2016; Wang et al., 2019), loss function reweighting (Cao et al., 2019; Cui et al., 2019; Huang et al., 2016; Khan et al., 2018; Ren et al., 2018), and transfer learning (Wang et al., 2018, 2017; J. Wu et al., 2020a, 2020b). Resampling-based methods aim to either downsample the classes with more instances or upsample the classes with fewer instances. However, downsampling tends to reduce the amount of training data, while upsampling increases training time and leads to overfitting for minority classes. Similarly, reweighting the loss function during training may lead to overfitting of large-weighted classes and insufficient training for small-weighted classes.

Two-stage Imbalance Learning decouples the learning of representation and the classifier head (Kang et al., 2020; Menon et al., 2020; Tang et al., 2020; T. Wang et al., 2020a, 2020b). This type of methods implements the feature extraction in the first stage and correct the biased decision boundaries of the model in the second stage by properly re-balancing the classifier head or directly adjusting the prediction scores. However, this approach often requires intricate hyperparameter tuning in practice. Distribution Alignment (Zhang et al., 2021) unifies class-balancing paradigms in semantic segmentation and introduces trainable parameters that can be adjusted adaptively. However, the trainable parameters neglect spatial information, which is crucial for semantic segmentation. In this regard, (Zhang et al., 2023) proposed the Distribution Alignment Module with Consideration of Spatial Contextual Information (DASCI) to cope with this problem. Specifically, the network parameters are frozen after the network has achieved feature extraction capabilities through training. Subsequently, additional parameters are trained to adjust biased decision boundaries while considering the spatial information. Readers can refer to (Zhang et al., 2023) for the details of the DASCI module.

4.3. The semi-supervised training method based on data masking

As mentioned earlier, the lack of data in China results in a relatively limited number of training samples, which hinders the model's ability to fully capture the characteristics of the buildings in the study area. Label-efficient deep image segmentation methods (Shen et al., 2023), including self-supervised, weakly supervised, and semi-supervised approaches, are potential for generating more training samples. For instance, (Zou et al., 2021) refined and purified pseudo-labels to create trainable samples. (Ahn and Kwak, 2018; Y. Wang et al., 2020d) utilized class activation maps from networks trained with image-level labels to generate pixel-level labels. (Liu et al., 2022) leveraged the model's tolerance to correct erroneous labels. In this study, during the dataset generation process (Section 4.1), a significant number of incompletely annotated building samples were collected in the ISB. Transforming the existing incomplete annotations (samples in the ISB) into trainable samples is a natural idea. Therefore, we propose a semi-supervised learning method based on data masking. This method masks the erroneous parts of the ISB samples to make them invisible to the model and then feeds the correctly annotated portions of the samples to the model

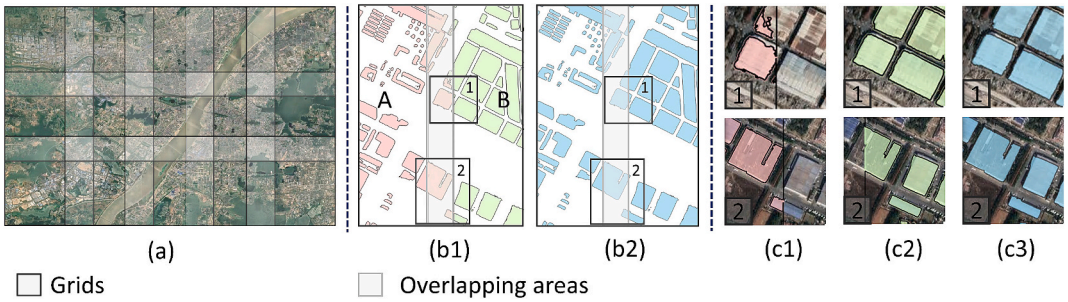

Fig. 9. (a) Demonstration of the grid division. (b1) Prediction results of two adjacent grids, where the red represents the prediction of grid A, the green represents the prediction of grid B, and the gray striped area indicates their overlapping region. (b2) Result after merging grid A and B. (c1) Enlarged view of Box 1 and Box 2 areas in grid A. (c2) Enlarged view of Box 1 and Box 2 areas in grid B. (c3) Enlarged view of Box 1 and Box 2 areas in the merged result. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

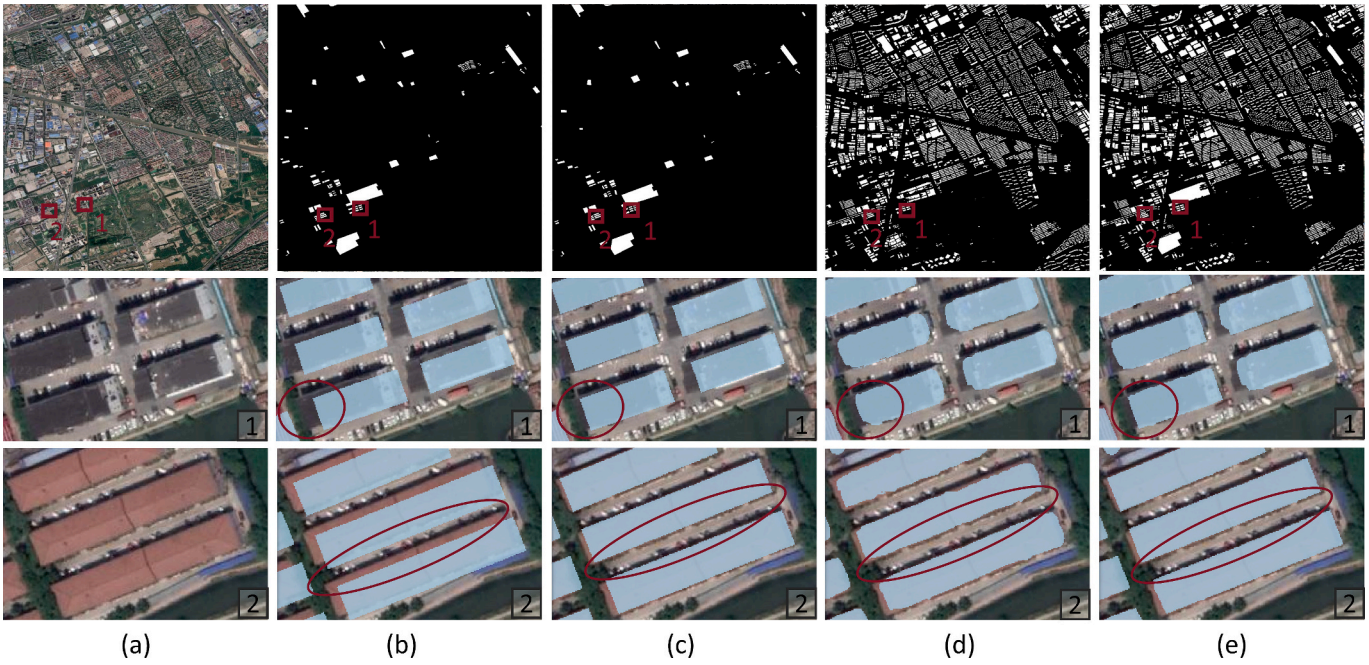

Fig. 10. (a) The Google imagery. (b) The OSM building footprints. (c) The adjusted OSM building footprints. (d) The prediction of the model. (e) The fused results. The second and third rows display an enlarged view of the Box 1 and the Box 2 regions in the first row, respectively. The red circles in the second and third rows illustrate the changes before and after processing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

for training. Note that our proposed approach differs in motivation from existing label-efficient deep image segmentation methods (Lai et al., 2021; Lin et al., 2016; Zou et al., 2021). The latter primarily aims to maintain model accuracy (compared to full supervision) while reducing the requirements for the precision, quantity, and annotation of existing samples. In contrast, our data masking algorithm aims to enhance data usability in a specific region (China) and make the model more specialized for that area.

As shown in Fig. 8, (a) is the Google imagery, (b) depicts the samples from D_2 (70% of ISB), (c) displays the predictions of the network (BldgNet) trained on ASB data, and (d) illustrates the labels after masking by the present method, i.e., the semi-supervised training labels. It can be observed that annotations in D_2 exhibit omissions for buildings, as indicated by the red circles in the first row. Directly utilizing labels from D_2 for network training can introduce numerous annotation errors, leading to a decrease in model accuracy. Therefore, it is crucial to identify the correct portions of the labels and mask out the mislabeled ones. Given that the ISB has been manually corrected, we consider the building annotations within it (designated as $region_1$) to be reliable. The background regions ($region_2$), consisting of x pixels around the buildings ($region_1$), can be deemed correct negative annotations. The $region_2$ is obtained by dilating the D_2 samples with x pixels and then taking the difference from the original samples. Note that the priority of the building annotations ($region_1$) is higher than the negative annotations ($region_2$), ensuring that the dilation operation does not lead to mislabeling the building areas as non-building. Consequently, the dilated region ($region_2$) does not introduce background errors. The predictions of the network (BldgNet) trained on the ASB data exhibit fewer building omissions, as shown in Fig. 8(c). Therefore, BldgNet is used to indicate the incorrect areas within the background class of ISB samples, which should be masked out subsequently. However, its predictions to the edge region are relatively inaccurate. Therefore, the areas containing building annotations in the prediction and the surrounding y pixels are designated as uncertain regions, labeled as $region_3$. The $region_3$ is obtained by dilating the prediction results with y pixels. The priority of the uncertain regions ($region_3$) is the lowest, ensuring that it does not overwrite the accurate regions ($region_1$ and $region_2$). In this study, we leverage these three regions ($region_1$, $region_2$, and $region_3$) to generate

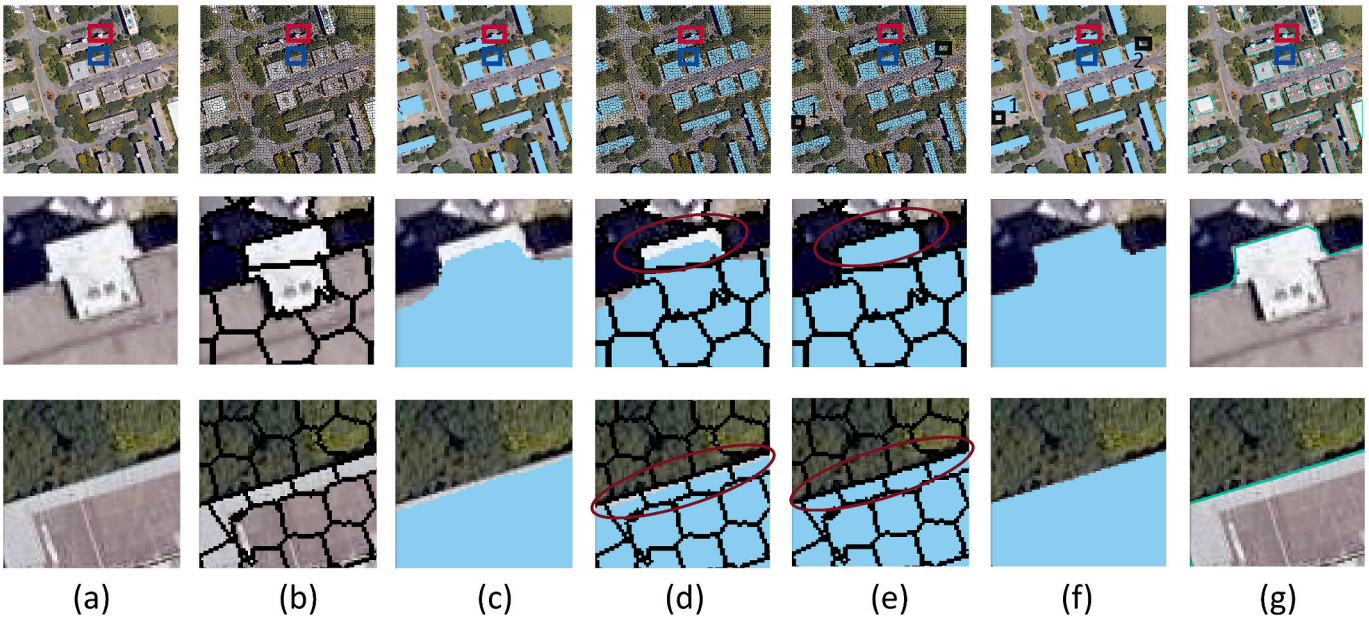

Fig. 11. (a) The Google imagery. (b) The superpixel segmentation result. (c) The prediction of the network. (d) The superpixel segmentation result overlaid with the network prediction. (e) The result after superpixel-based fusion. (f) The result after image morphology processing. (g) The result after vectorization and topology checking. The second and third row provide an enlarged view for the red-boxed and blue-boxed region in the first row, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

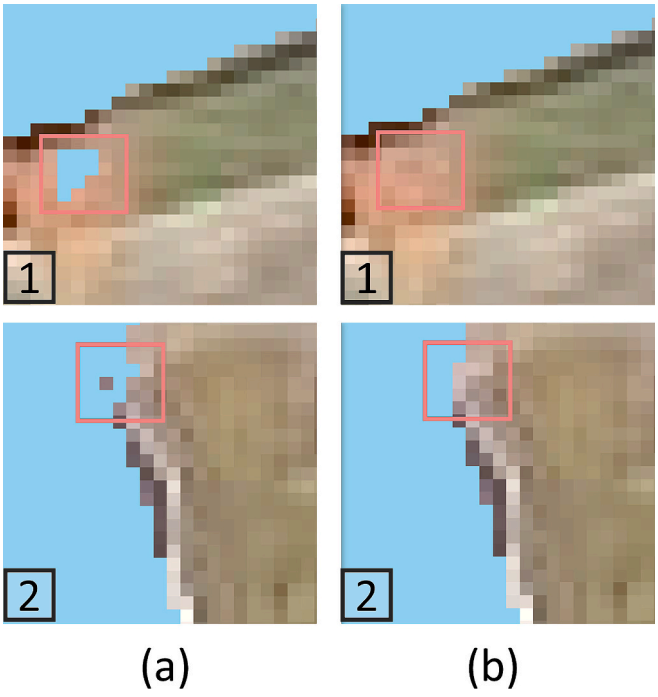

Fig. 12. The result before (a) and after (b) the morphological processing.

Table 2
Comparison of different products in the study area.

Dataset	IoU (%)	Precision (%)	Recall (%)	F1 score (%)
90-cities-BRA(Z. Zhang et al., 2022b)	65.98	75.12	80.23	77.59
CBRA(Liu et al., 2023)	55.88	61.72	74.78	67.63
CBF(ours)	**73.98**	**81.15**	**86.45**	**83.71**

semi-supervised training labels based on the following rules:

$$\text{New label} = \begin{cases} Building = region_1 \\ Background = \left(R - \left(region_1 \cup region_3\right)\right) \cup region_2 \\ Masked\ region = region_3 - \left(region_1 \cup region_2\right) \end{cases} \#$$

(1)

where *Building* signifies the building region, *Background* designates the non-building region, *Masked region* denotes the areas to be masked, and *R* is the entire labeled region. In Fig. 8, the second row represents an enlarged view of the yellow-boxed region in the first row. It is shown that the new labels incorporate accurate portions from D_2 samples, particularly near the edges. Meanwhile, the white regions (*Masked region*) represent the omissions in D_2 samples identified by the prediction of the BldgNet (see (c)). The mask of the white regions eliminates misidentifications in the background of (b). The white regions are not involved in network training and loss function calculations, thus preventing the introduction of erroneously labeled data. This process can effectively utilize the incomplete samples (ISB) for training, which is crucial in the case of scarcity of training and test samples in the study area (China).

4.4. Post-processing

Post-processing aims to refine and modify the predictive results of the model. It does not require additional model training and can compensate for deficiencies in the model's predictions. Current research in large-scale building mapping (Liu et al., 2023; Z. Zhang et al., 2022b) has given relatively less attention to the post-processing techniques. Therefore, in this study, we propose a series of post-processing algorithms to enhance the predictive outcomes of building extraction (depicted in Fig. 1(c)):

1) Overlapping Area Processing. In the large-scale mapping, models cannot be implemented on all the data simultaneously. Therefore, the regular practice is to divide the study area into grids and process the data grid by grid (Fig. 9(a)). However, this approach often leads to errors in the edge regions owing to the lack of surrounding

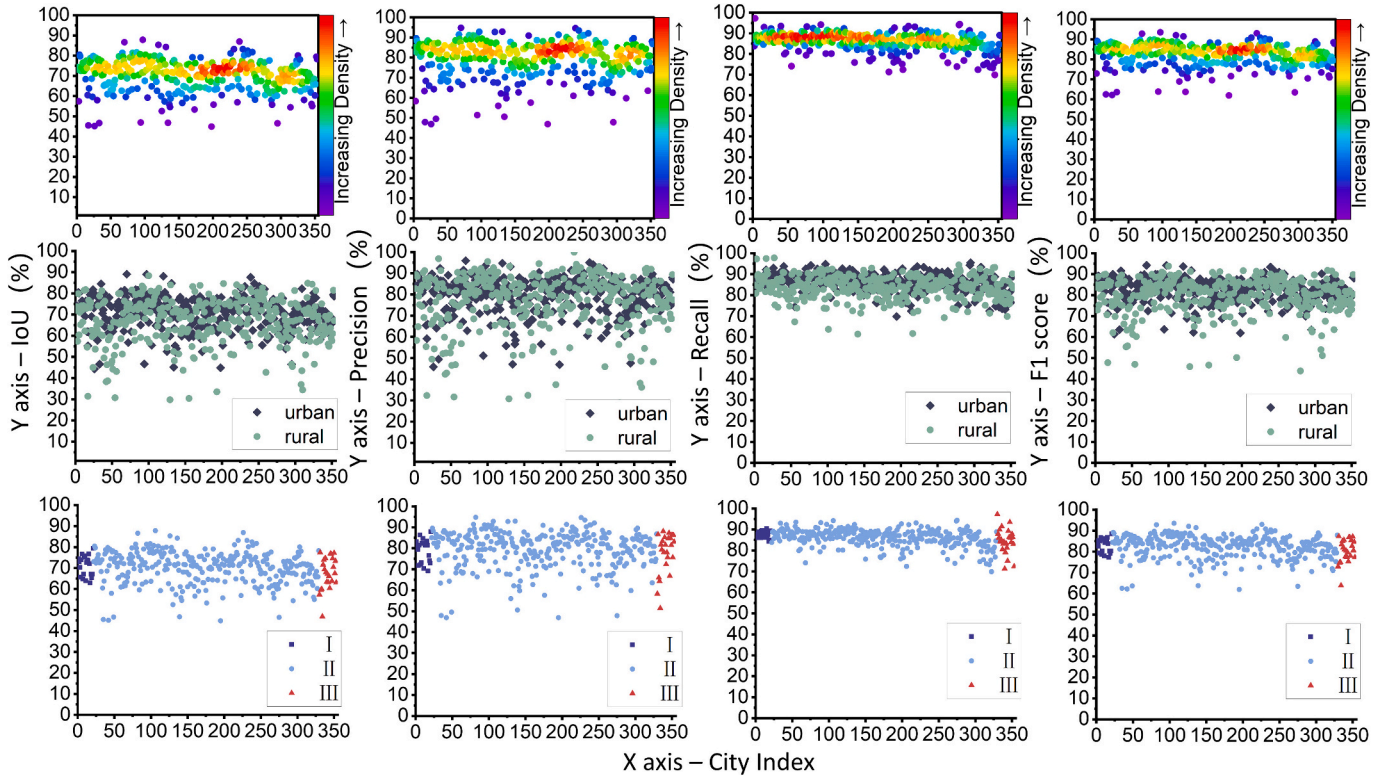

Fig. 13. Accuracy assessment of the CBF dataset. The first row presents the accuracy scores of various metrics in each city. The second row illustrates the results for urban and rural areas of each city. The third row shows the evaluation for cities with different grades. The vertical axis represents the relevant evaluation metrics, while the horizontal axis corresponds to the city index labels.

Table 3
Comparison of IoU between urban and rural areas among different products.

Dataset	Urban	Rural
90-cities-BRA	66.48%	55.15%
CBRA	59.23%	45.12%
CBF(ours)	**74.27%**	**70.12%**

Table 4
IoU comparison across cities of different grades for various products.

Dataset	I	II	III
90-cities-BRA	65.97%	66.07%	–
CBRA	55.46%	56.56%	48.87%
CBF(ours)	**73.05%**	**74.44%**	**68.45%**

information. A possible solution is to consider the overlapping regions between adjacent grids, in order to enhance the certainty of the prediction in the edge areas. Specifically, in Fig. 9(b1), the prediction results of a grid are represented in Red (referred to as A), the results of its neighboring grid are indicated in Green (referred to as B), and their overlapping region is marked by the gray stripes. We adopt a union approach to merge the results of the overlapping region, and the fused result is shown in Fig. 9(b2). Fig. 9(c1) ~ Fig. 9(c3) presents enlarged views of two boxed regions (i.e., Box 1 and 2). Fig. 9 (c1) and (Fig. 9(c2) depict the prediction results of grid A and B, respectively. The third column (Fig. 9(c3)) shows their fused result. It can be observed that the fusion of overlapping areas can mitigate the omissions of the mapping results.

2) Integration with OSM. As illustrated in Fig. 10, the OSM data accurately delineate the footprints for large high-rises and factories, in spite of conspicuous omissions in other regions. Moreover, there exists a slight positional offset between OSM data and the imagery, as shown by the red circles in Fig. 10(b). However, on the other hand, the position of the model's predictions is accurate (Fig. 10(d)). Therefore, during the post-processing, the results of the model (BldgNet) and OSM are merged in a decision fusion, to leverage their respective strengths (Fig. 10).

Specifically, let $B = [b_1, b_2, \ldots, b_m]$ and $H = [h_1, h_2, \ldots, h_n]$ represent the set of buildings in OSM and model prediction, respectively. For each building b_i in OSM, we calculate its Intersection over Union (IOU) with each building h_j in the prediction set H resulting in $[IOU_1, IOU_2, \ldots, IOU_n]$. The building h_x with the highest IOU is regarded responsible for b_i. Afterwards, b_i is shifted within 5 pixels to generate a candidate set of buildings $[b_{i,1}, b_{i,2}, \ldots, b_{i,25}]$, and the IOU between h_x and each building in the candidate set is calculated, yielding $[iou_1, iou_1, \ldots, iou_{25}]$, where the building with the highest IOU is considered as the corrected building b_i'. The corresponding value of IOU is regarded as its confidence level of the OSM building, denoted as p. If $0.8 < p < 0.95$, the union of h_x and b_i' is used as the final prediction result. If $0.95 \leq p$, b_i' is employed as the ultimate prediction result. As illustrated by the red circles in Fig. 10(e), the fused result ensures both accurate positioning and edge precision.

3) Superpixel Fusion. The superpixel segmentation (S. Chen et al., 2023b; Jampani et al., 2018; Liu et al., 2018; Stutz et al., 2018) results often adhere well to object boundaries. Therefore, in this study, a post-processing algorithm based on superpixel segmentation is designed to reduce building omissions and enhance edge accuracy. Specifically, we first use the SLIC (Liu et al., 2018) segmentation method to divide the Google imagery (Fig. 11(a)) into a series of superpixels (Fig. 11(b)). Then, the network predictions (Fig. 11(c)) are overlaid onto the superpixel results (Fig. 11(d)), and the ratio of

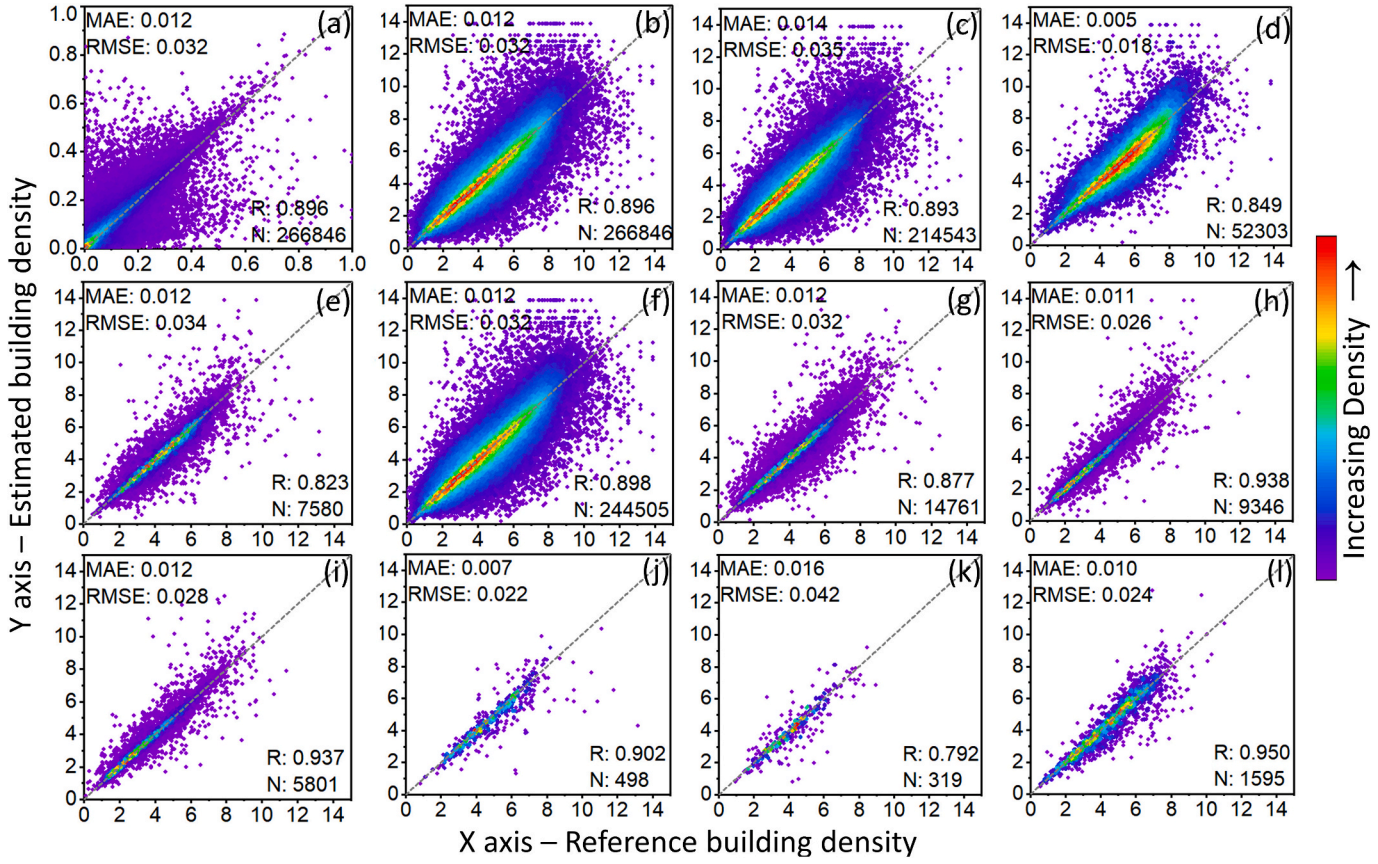

Fig. 14. Accuracy assessment of Building Area. (a) Overall results of the area evaluation, with the x-axis representing the reference building density and the y-axis the predicted building density. (b) Overall results after applying logarithmic scaling to both the x and y axes. (c) and (d) depict the results for urban and rural areas, respectively. (e), (f), and (g) present the overall results for cities of different levels: I (e), II (f), and III (g). (h) ~ (l) display the results for Beijing, Shanghai, Bijie, Haikou, and Tainan City, respectively.

building pixels to the total pixels within each superpixel is calculated. The superpixels with this ratio exceeding 0.5 are identified as buildings. From Fig. 11, it can be observed that the superpixel fusion algorithm can effectively improve edge accuracy (as indicated by the red circles in Fig. 11(d) and Fig. 11(e)).

4) Image Morphology Processing. The image morphology processing techniques are further utilized to remove isolated pixels and small patches and fill holes. The outcome of this processing is shown in Fig. 11(f). To provide a clearer demonstration of the changes before and after processing, Fig. 12 zooms in on two black-boxed regions from Fig. 11(e) and Fig. 11(f).

5) Vectorization and Topology Checking. Finally, the results are vectorized and the topological errors are corrected manually (Fig. 11 (g)).

4.5. CBF dataset generation

4.5.1. Data generation procedures

The production process is shown in Fig. 1(d). Initially, all the samples from ASB were used for training, to ensure that the model (M_1) can learn various architectural styles and achieve better generalization. Subsequently, fine-tuning was performed using data within China (D_1) to optimize the model (M_2) and achieve superior performance within the study area. Finally, the network was trained using the semi-supervised training dataset (D_2) to fully explore the characteristics of buildings, resulting in model M_3. The CBF building footprints were generated using model M_3 with the post-processing steps.

4.5.2. Parameter settings

The initial learning rate of the fully supervised training was set to 0.001. The cosine annealing decay strategy was employed. The AdamW optimizer was utilized with a weight decay of 1e−4 to optimize the model. The batch size was configured to 36, and the maximum number of epochs was set to 120. In the fine-tuning and semi-supervised training stages, a fixed learning rate of 1e-6 was used, maintaining consistency with other settings of the fully supervised stage. Data augmentation techniques, such as random rotation, scale transformation, and colour jitter, were applied across all training stages. The proposed method was implemented using the PyTorch framework and executed on a computer equipped with four NVIDIA GeForce RTX 3090 GPUs.

4.6. Accuracy assessment

The model performance and accuracy are validated using intersection-over-union (IoU), precision, recall, and the F1 score, which are commonly employed metrics for semantic segmentation. The construction of the test samples has been elaborately described in Section 4.1, encompassing 750,000 buildings across 350 cities in China. In addition, the accuracy of building footprints is assessed using reference building density (detailed in Section 5.1). In terms of efficiency, the model speed is evaluated by measuring the number of images processed per second, denoted as frames per second (FPS).

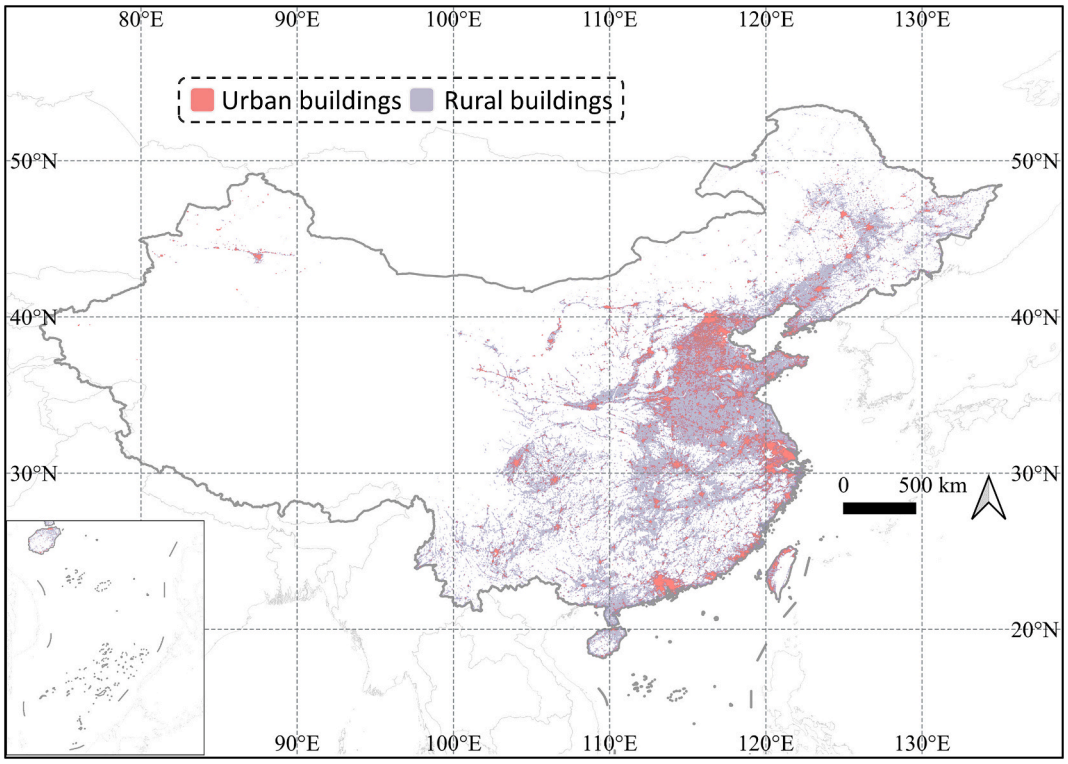

Fig. 15. Spatial distribution of CBF. Red denotes urban building footprints, while blue indicates rural building footprints. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Results

5.1. Quantitative analysis

350 cities in China, including 750,000 buildings, were used for the accuracy assessment. There is no overlap among the test, training, and semi-supervised samples. The overall IoU, Precision, Recall, and F1 score is 73.98%, 81.15%, 86.45%, and 83.71%, respectively, representing an overall satisfactory result. As compared in Table 2, the IoU values for 90-cities-BRA (Z. Zhang et al., 2022b) and CBRA (Liu et al., 2023) are 65.98% and 55.88%, respectively. The first row of Fig. 13 presents the evaluation results for each city of the generated CBF dataset. It can be observed that the IoU for the majority of cities (75%) exceeds 65%.

Compared to urban areas, rural samples are relatively scarce, with smaller building sizes and more complex backgrounds. Therefore, extracting buildings from rural areas is more challenging. Existing large-scale mapping results are unsatisfactory in rural areas (Liu et al., 2023; Marconcini et al., 2020; Sirko et al., 2021). In this study, we also used the urban boundary provided by (Xuecao Li et al., 2020b) to define urban and rural areas and conduct separate accuracy assessments. As shown in Table 3, the IoU of 90-cities-BRA and CBRA in rural areas is significantly lower than that in urban areas. In contrast, our model, trained on diverse samples from global regions, exhibits stronger generalization. Additionally, fine-tuning and semi-supervised training in the study area enable our method to better capture the characteristics of China's buildings. This, to some extent, improves the effectiveness of rural mapping and reduces the gap of mapping accuracy between urban and rural areas. For CBF, the IoU, Precision, Recall, and F1 scores in urban areas are 74.27%, 81.42%, 86.54%, and 83.90%, respectively. The corresponding results in rural areas are 70.12%, 77.48%, 83.40%, and 80.33%. Meanwhile, the second row of Fig. 13 shows the results for urban and rural areas within each city of the CBF dataset.

Cities of different levels exhibit variations in sample quantity and building morphology, leading to differences in mapping performance. To thoroughly assess our building extraction results across cities of varying levels, we categorize the cities into three grades:

I: 21 M-cities or super-large cities according to the latest urban ratings from the National Bureau of Statistics (National Bureau of Statistics, 2021).

II: Cities with a population exceeding one million (except for grade I).

III: Cities with a population equal to or less than one million.

Table 4 presents the assessment results for different products across cities of various levels. Our CBF demonstrates higher accuracy in all the levels of cities. Building extraction from Level I cities exhibit more challenges compared to level II, due to a greater number of ultra-high-rise buildings. The level III cities are akin to rural areas, with relatively fewer training samples and smaller building sizes, resulting in lower prediction accuracy than other ones. Fig. 13, in the third row, shows the results for each city of our CBF product across different city levels.

To further evaluate the building extraction accuracy, we divided the test dataset into 260,000 grids of 100 m × 100 m. The accuracy of the building area is assessed based on the predicted building density and the reference building density within each grid. Building density is the ratio of building area to the grid area. The overall results, as shown in Fig. 14 (a), exhibit consistency between the predicted and reference building density: Mean Absolute Error (MAE) is 0.012, Root Mean Squared Error (RMSE) is 0.032, and the correlation coefficient (R) is 0.896. An MAE of 0.012 implies a difference of $1.2m^2$ in the predicted building area compared to the reference building area within a $100m^2$ plot. In Fig. 14 (b), the logarithmic scaling was applied to both the x and y axes for better distribution visualization. Fig. 14(c) and Fig. 14(d) present the results for urban and rural areas, respectively. The correlation coefficient in rural areas is slightly lower than in urban areas. Figs. 14(e), (f), and (g) show the results of cities at levels I, II, and III, respectively. Additionally, Figs. 14(h)~(l) depict examples of cities from different

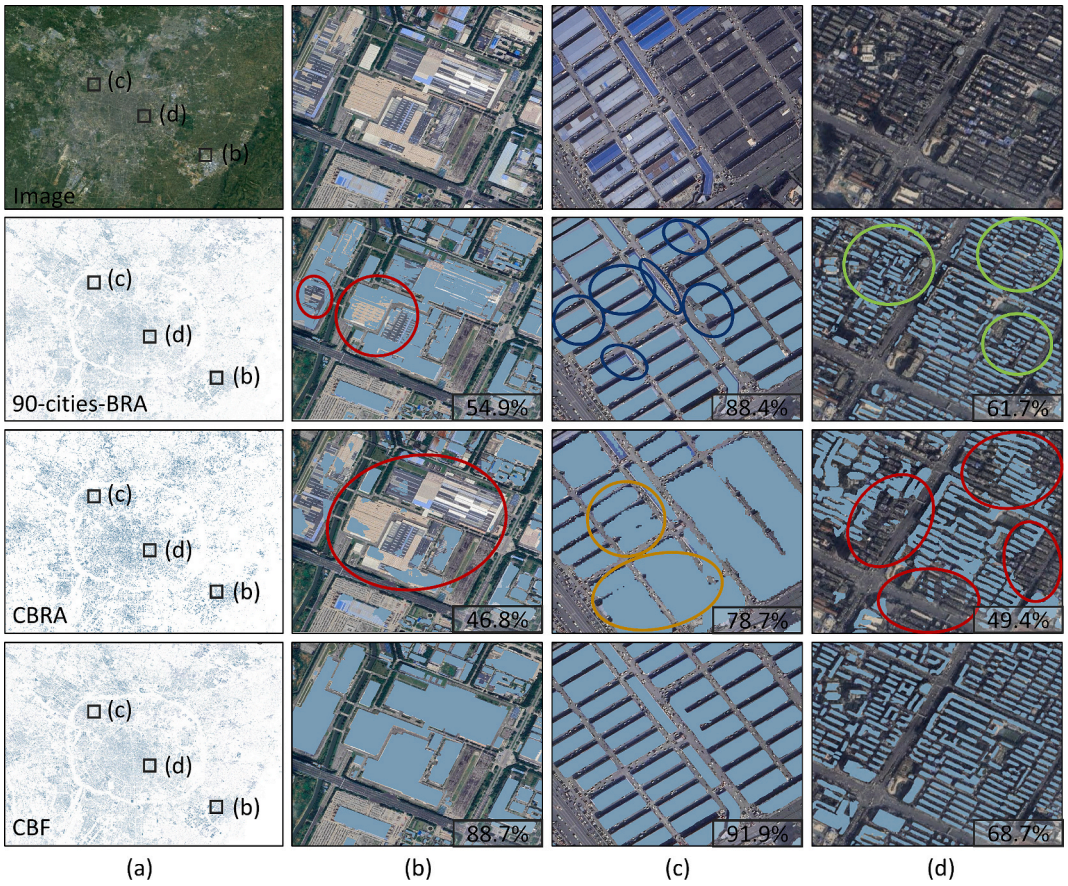

Fig. 16. Comparison of different products in Chengdu. The first row is the Google imagery, the second, third, and the fourth row show the 90-cities-BRA, the CBRA, and our CBF, respectively. Column (a) provides an overview of Chengdu, with (b), (c), and (d) displaying enlarged views of the black-boxed regions. The numerical values in the bottom right corner indicate the IoU of the current predicted result.

levels, including Beijing (I), Shanghai (I), Bijie (III), Haikou (II), and Tainan (II). It can be observed that the building areas in various scenarios are effectively reflected in the CBF dataset.

It deserves noting that we should be cautious for the accuracy comparison between CBF and other building products, as they are not on the same benchmark.

5.2. Overview of the CBF dataset

Figure 15 illustrates the nationwide spatial distribution of CBF. CBF contains approximately 185 million buildings across China, with a total floor area of 598.80 billion square meters. The average spacing between buildings is about 8.6 m. We conducted a separate statistical analysis by dividing the regions into urban and rural areas based on the 2020 urban boundary (Xuecao Li et al., 2020b). The total building area in urban regions is 357.52 billion square meters, and 241.27 billion square meters in rural areas. The average inter-building distance in urban areas is around 6.9 m, and in rural areas, it is approximately 9.5 m. The total number of buildings is 66.7 million in urban areas and 119 million in rural areas. Overall, urban buildings tend to be denser and larger than rural ones. Note that the statistical analysis to the CBF dataset is beyond scope of this research, and will be investigated in future.

6. Discussions

6.1. Comparison with other products

We visualized and compared different building products in various regions of China. Fig. 16 illustrates the three products in Chengdu City. The first row displays the Google imagery, the second, third, and fourth row exhibit the 90-cities-BRA, the CBRA, and our CBF datasets, respectively. Column (a) represents an overview of results in Chengdu city, while columns (b), (c), and (d) provide enlarged views of the black-boxed regions. The numerical values in the bottom right corner of each image denote its IoU of the predicted result. It can be observed, particularly in column (b) within the red circle, that 90-cities-BRA and CBRA exhibit holes and significant omissions in the large buildings. This possibly can be attributed to the limited receptive field of these models, failing to adequately capture the contextual information of large buildings. In addition, the textures on the roofs of large buildings as well as the presence of other structures (such as solar panels, air conditioning units) further complicate the extraction. In contrast, our proposed LWA module can effectively capture global information, and enhance the model's ability of contextual modeling. This module can significantly improve the extraction results for large buildings.

In the (c) column, within the blue circles, the boundary of the extraction results from 90-cities-BRA is not precise enough. Our proposed EA module can improve this issue by extracting additional features from the edge regions. Furthermore, the post-processing method based on superpixels further raises the accuracy of the boundaries, as demonstrated in the third row of the (c) column, where the building boundaries align well with the reference. Simultaneously, in the yellow circles of the (c) column, CBRA, portrays rough edges and tends to treat adjacent buildings as a single entity, which might be attributed to its coarser spatial resolution. Due to the interference from building shadows or other factors, the extraction results of the (d) column (especially in green circles) have more fragmented spots, which can be mitigated by the morphological post-processing adopted in our method.

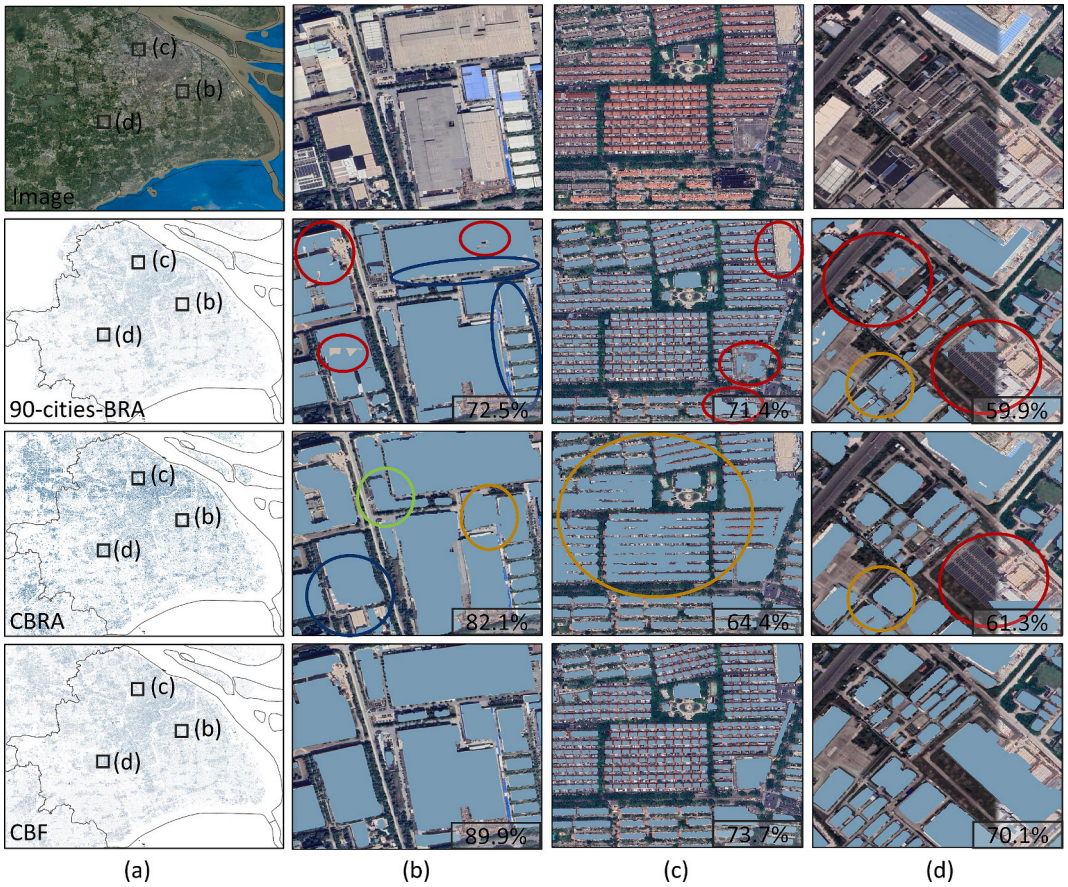

Fig. 17. Comparison of different products in Shanghai. The first row is the Google imagery, the second, third, and the fourth row show the 90-cities-BRA, the CBRA, and our CBF, respectively. Column (a) depicts an overview of predictions in Shanghai, with (b), (c), and (d) showing enlarged views of the black-boxed regions. The numerical values in the bottom right corner indicate the IoU of the current predicted result.

The (d) column in red circles exhibit omission of densely packed buildings in other products. The omission of dense buildings might stem from CBRA's lower resolution and the smaller receptive field of its extraction method.

Figure 17 provides another comparative example, illustrating the results of Shanghai from different datasets. The red circle in Fig. 17 highlights the omission of large buildings. Compared to 90-cities-BRA and CBRA, CBF (ours) exhibits more accurate extraction without holes or omissions. The yellow circle depicts the results of closely spaced buildings, showing adhesion of buildings in 90-cities-BRA and CBRA. The blue circle showcases the delineation of building boundaries, where our CBF can achieve a more precise boundary compared to 90-cities-BRA and CBRA. The green circle represents a false alarm in CBRA, which is caused by the similar textures and shapes between buildings and other objects in the background (e.g., parking lots). This problem could be effectively addressed by introducing a substantial number of diverse negative samples. However, the inclusion of a large number of negative samples can lead to imbalance between buildings and backgrounds in the training samples, making the model overly focus on non-building categories and thus lowering building extraction performance. In this regard, our approach deals with this issue by incorporating the DASCI module to assist the model in adjusting biased decision boundaries (between buildings and non-buildings).

In smaller cities or rural areas, buildings tend to be smaller, scattered, and embedded in complex backgrounds. Consequently, mapping buildings in these regions poses more challenges. Meanwhile, the western regions in China exhibit lower population densities than the eastern ones. Therefore, mapping results in these areas can further reflect the quality of the product. We chose Linzhi and Karamay, two western cities of China with populations under one million, for illustration. They are situated in Xizang and Xinjiang, respectively. Fig. 18 shows the visual results for Linzhi, where the first, second, and third row present the results of 90-cities-BRA, CBRA, and our CBF, respectively. It can be seen that 90-cities-BRA did not produce mapping results in this region. CBRA's results are blurred in details and boundaries ((b) and (d) columns, black circles) and exhibit omissions ((c) column, black circle). Possible explanations include: 1) in this region the roofs are spectrally similar with the background, confusing the model's decisions; 2) the image resolution is coarse (10 m) and the size of buildings is small, and CBRA's super-resolution method fails to capture spatial details adequately. This situation has been significantly improved in our CBF dataset. It can be attributed to the training strategy that combines both full supervision and semi-supervision, resulting in more diverse samples and more robust models.

Karamay City (Fig. 19), similar to Linzhi City, exhibits clumped predictions and building omissions in circles (b) and (d). Furthermore, the predictions within the black circle in column (c) demonstrate clear-cut incisions, since the region's position is located at the grid's edge, and the insufficient contextual information can lead to the wrong or incomplete predictions. This issue can be properly resolved by our post-processing approach (i.e., overlapping predictions).

6.2. Comparison with other methods

For large-scale mapping, models need to meet requirements in both accuracy and speed. Therefore, we compared the accuracy and efficiency of different building extraction models. In this section, we randomly sampled 25,000 images (20,000 for training, 5000 for testing)

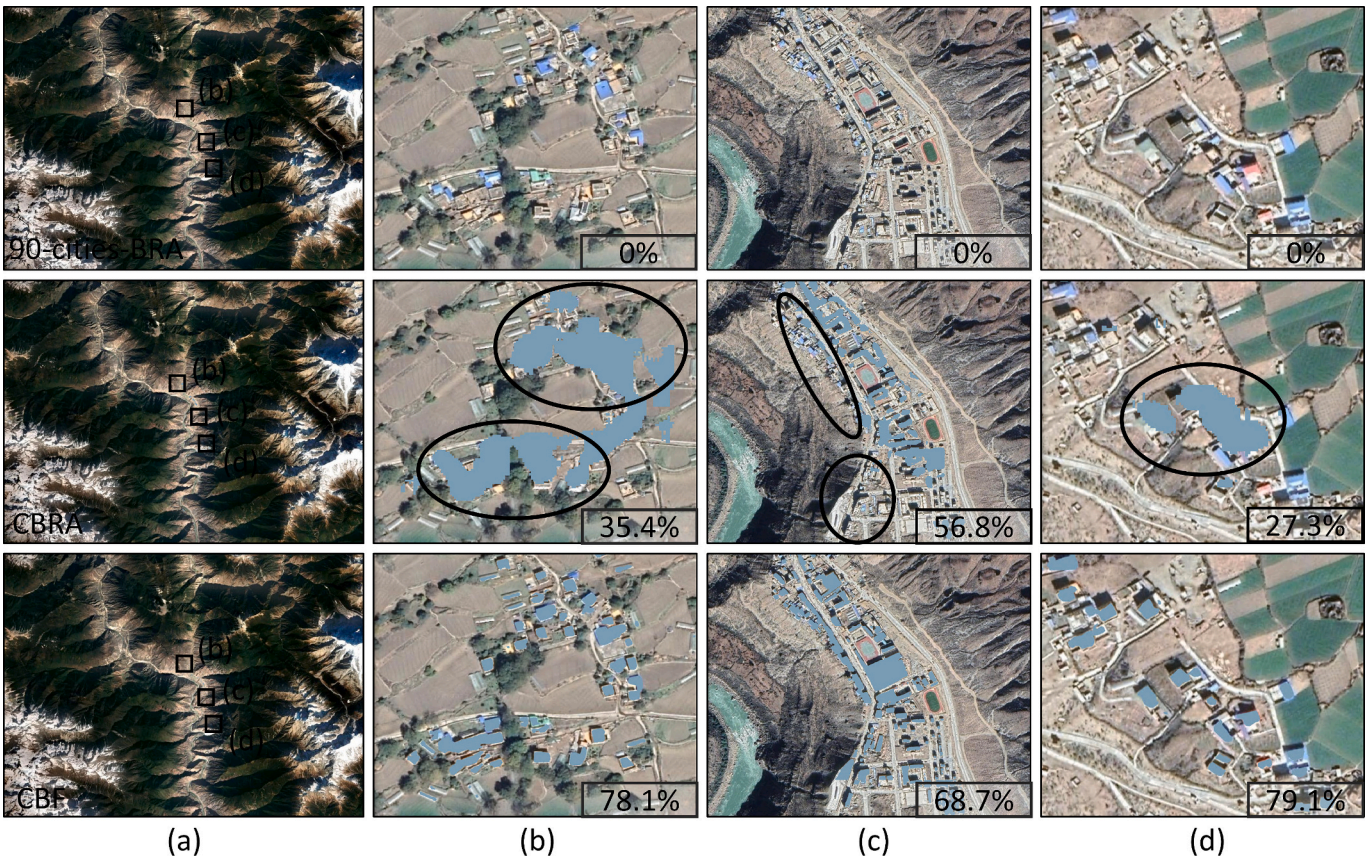

Fig. 18. Comparison of different products in Nyingchi. The first, second, and third row present the results of 90-cities-BRA, CBRA, and our CBF, respectively. Column (a) provides an overview of Linzhi, with (b), (c), and (d) showing enlarged views of the black-boxed regions. The numerical values in the bottom right corner indicate the IoU of the current predicted result.

from the ASB dataset for the comparison. As shown in Table 5, lightweight models (i.e., models with lightweight designs) such as TopFormer (W. Zhang et al., 2022a), EdgeFormer-S (Pu et al., 2022), and Segformer-B3 (Xie et al., 2021) achieve an IoU over 62% with >100 FPS, showing a good balance between accuracy and efficiency. However, these networks are not specifically designed for building extraction, resulting in subpar performances. On the other hand, conventional models like ST-UNets and HRFormer, while superior in accuracy, lack lightweight design, making their operating speed unsuited for large-scale mapping. BldgNet achieves an IoU of 72.84% at a relatively higher speed. Compared to lightweight models (e.g., MobileNet, EfficientNet), our model notably improves the metric of recall. Meanwhile, compared to models with a larger number of parameters (e.g., HRFormer, ST-UNets), our proposed BldgNet achieves a comparable level of precision but has a much faster computation speed (i.e., much smaller FPS). FPS is measured using an Intel(R) Xeon(R) Silver 4114 CPU, 3 NVIDIA RTX 3090 GPUs, and the maximum batch size it can afford. Furthermore, when the model is applied to high-resolution large-scale mapping, hundreds of epochs and billions of image patches are required for training and prediction. Consequently, the difference of efficiency can be substantially large among different models. For example, it took 7 days to complete China's building mapping using our model, whereas it would take three to four months using the HRFormer. However, it should be noted that while the efficiency of the method presented in this paper is advantageous, the degree of its efficiency advantage may become less significant with an appropriate increase in computational power in the future.

We visually compared the results of the aforementioned methods on the test dataset. As illustrated by the blue circles in Fig. 20, the predictions of the lightweight network may overlook smaller buildings or those situated at the patch's edge, and have less accuracy in delineating building edges. Moreover, these results often exhibit scattered false alarms. This suggests that the lightweight networks struggle to meet the accuracy standards for large-scale mapping. It also underscores the importance of the modules proposed in this study, which aim to enhance model performance while minimizing the computational burden.

6.3. Effects of the proposed method

The efficacy of different modules in the proposed BldgNet is analyzed using the experimental dataset aforementioned (25,000 images). The results (as shown in Table 6) indicate that LWA and EA modules increase the IoU by 2.04 and 0.94 at the cost of decreasing 19 FPS and 6 FPS, respectively. DASCI can raise the IoU by 2.09 without affecting model FPS, but it requires an additional 1–10 epochs for training. Furthermore, to evaluate the effects of model training, we assessed the accuracy of the prediction results with different training datasets using the test dataset (750,000 buildings in 350 cities), as shown in Table 7. As defined in Section 4.4, M_1 represents the model trained on ASB data, M_2 represents the model fine-tuned in the study area, and M_3 represents the model with the semi-supervised training. It is observed that the semi-supervised training contributes most to the accuracy gain, showing that the large amount of incomplete samples from the OSM are beneficial for the large-scale building extraction.

6.4. Accuracy of ASB Samples

Given the significance of the sample accuracy, this section delves into the relevant factors in the sample production process that could influence the sample accuracy.

Fig. 19. Comparison of different products in Karamay. The first, second, and third row present the results of 90-cities-BRA, CBRA, and our CBF, respectively. Column (a) provides an overview of Karamay, with (b), (c), and (d) showing enlarged views of the black-boxed regions. The numerical values in the bottom right corner indicate the IoU of the current predicted result.

Table 5
The comparison of accuracy and speed among different models.

Method	IoU (%)	Precision (%)	Recall (%)	FPS
TopFormer-B(W. Zhang et al., 2022a)	63.71	86.57	71.38	276
EdgeFormer-S(Pu et al., 2022)	62.45	84.53	72.92	182
Segformer-B3(Xie et al., 2021)	66.49	88.73	72.65	113
MobileNet(Howard et al., 2019)	62.73	84.92	70.65	172
EfficientNet(Tan and Le, 2019)	61.84	78.86	74.21	125
ST-UNets(He et al., 2022)	73.82	86.24	83.75	10
HRFormer(Yuan et al., 2021)	74.15	85.85	84.41	18
UNet(Ronneberger et al., 2015)	70.36	84.16	81.28	68
HRNet(J. Wang et al., 2020c)	71.24	87.18	79.71	65
BldgNet(ours)	72.84	85.49	83.33	88

1) Division between ASB and ISB. Impervious surfaces and buildings are not synonymous. Impervious surfaces contain buildings and other associated features (roads, squares, etc.). Consequently, a larger threshold δ may result in misclassification of ISB samples (missing part of the buildings) into ASB, whereas a smaller δ may leads to misclassification of eligible samples into ISB. However, in this study, both cases can be appropriately dealt with through the proposed sample construction process. Specifically, on the one hand, we enhance the quality of the ASB by eliminating the unqualified samples from the initial ASB through intensive visual inspection. On the other hand, the division of ASB (qualified and complete) samples into ISB (incomplete) does not trigger errors in the proposed semi-supervised sample construction process. Therefore, the uncertainties can be effectively controlled and suppressed during the sample construction process. With regard to the threshold, a large value can retain more ASB samples but entail high labor intensity to screen out incorrect samples. In contrast, a small value may yield fewer ASB samples, and weaken the sample diversity, yet reduce the workload for manual inspection.
2) How to control manual errors during sample production. In this study, we dedicated significant effort to ensure the accuracy of the sample datasets. 70 experienced interpreters (specialized in remote sensing) were involved in screening samples for the initial ASB, spending more than five months. To ensure sample accuracy, we established stringent criteria, i.e., only samples with no building footprint omissions, offsets, distortions, and false alarms were considered as candidates. In this way, approximately 800,000 accurate samples (about a 5% pass rate) were selected from a pool of 15.92 million samples (each sample is a 512 × 512 image patch) worldwide. Furthermore, to further minimize errors, each sample should be confirmed by two interpreters before being considered as an ASB candidate sample. Additionally, all candidate samples are checked by an administrator before being designated as ASB samples.
3) Validation of ASB sample accuracy. To further assess the accuracy of the samples generated through the proposed process, we conducted additional sample accuracy assessment. Specifically, 300 samples (in 512 × 512 patches) were randomly chosen and manually annotated to evaluate the accuracy of the samples. The results are presented in Table 8. As depicted in the table, the sample accuracy is remarkably high, and close to the level of manual annotation, thus satisfying the sample requirements for deep learning.

6.5. Accuracy of ISB Samples

As aforementioned, a large quantity of ISB samples are collected, which is crucial for strengthening the diversity and quality of the samples for large-scale building footprint mapping. However, manually correcting such a vast number of ISB samples is impracticable. Conse-

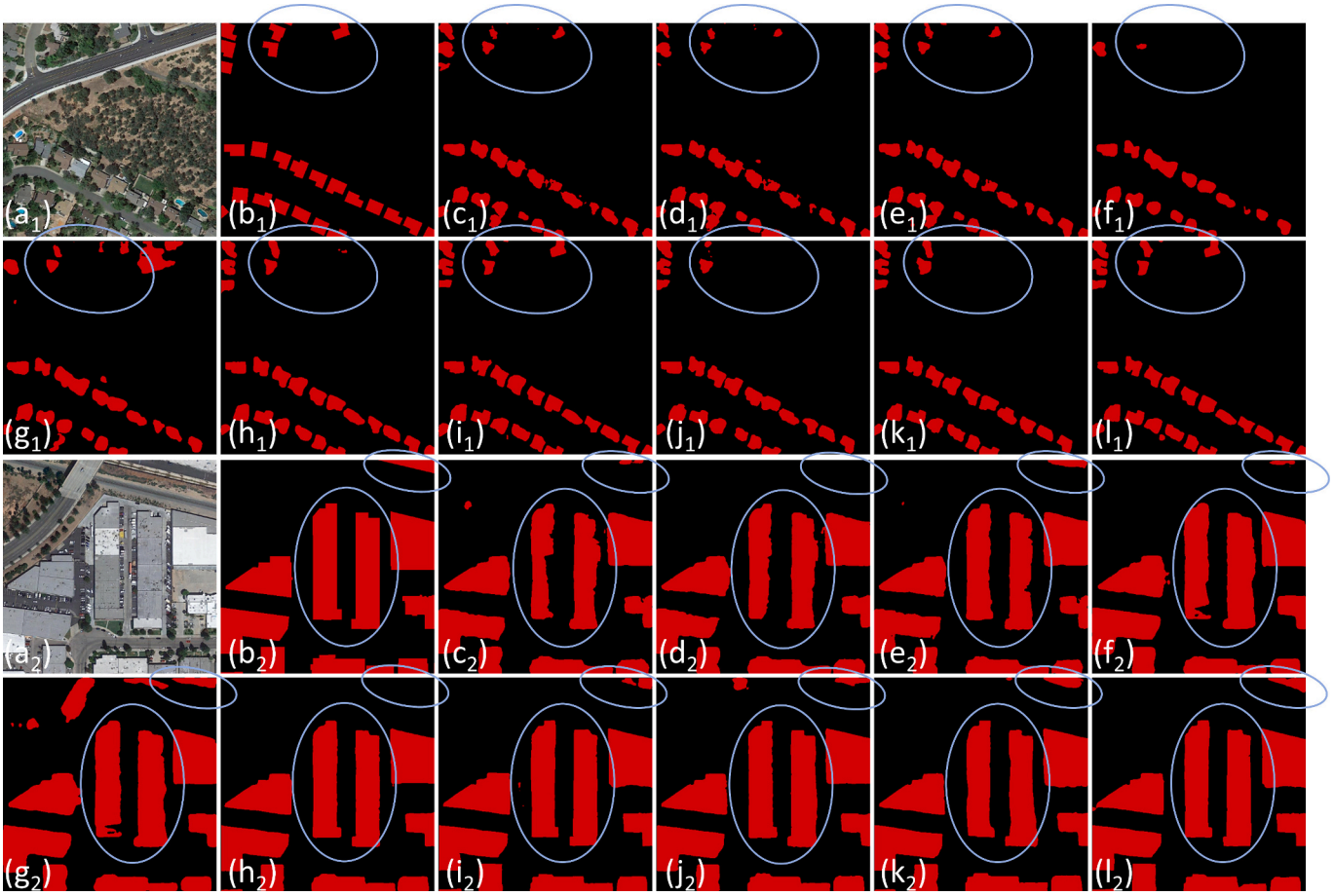

Fig. 20. Visualization results of different models on the test dataset. (a) Image. (b) Label. (c) TopFormer-B. (d) EdgeFormer-S. (e) Segformer-B3. (f) MobileNet. (g) EfficientNet. (h) ST-UNets. (i) HRFormer. (j) UNet. (k) HRNet. (l) BldgNet(ours). The numbers 1 and 2 indicate the first and second examples, respectively.

Table 6
Results of the module ablation experiment.

LWA	EA	DASCI	IoU(%)	FPS
×	×	×	67.77	103
√	×	×	69.81	94
√	√	×	70.75	88
√	√	√	72.84	88

Table 7
Accuracy of prediction results with different training strategies. M_1 represents the model trained on ASB data, M_2 represents the model fine-tuned in the study area, and M_3 represents the model with the semi-supervised training.

Model	IoU(%)
M_1	71.36
M_2	71.92
M_3	73.15
M_3 + Post-processing	73.98

quently, this study designs a semi-supervised sample construction scheme. Specifically, an ISB sample can be divided into two regions, i.e., the regions with building footprints and the ones without building footprints. In the building regions, the footprints are visually inspected to ensure completeness and absence of offset and distortion. In the regions without building footprints, we utilize the prediction results of BldgNet to label these areas as missing regions, enabling the model not

Table 8
Results of ASB accuracy evaluation in different regions.

Region	IoU(%)	Precision(%)	Recall(%)	F1 score(%)
Europe	94.48	95.07	94.60	94.83
North America	95.22	95.55	95.74	95.65
South America	92.66	95.57	96.83	96.20
Africa	94.63	90.71	94.50	92.56
Asia	89.93	90.18	96.09	93.04
Oceania	94.60	96.04	96.00	96.02

Table 9
Accuracy of ISB and SST samples.

Dataset	IoU(%)	Precision(%)	Recall(%)	F1 score(%)
ISB	35.27	90.04	36.71	52.15
SST Samples	86.05	90.04	95.10	92.50

to learn from these regions. As the prediction results of BldgNet are solely used for masking samples but not for network learning, and it holds the lowest priority in the semi-supervised sample generation process (compared to $region_1$ and $region_2$), the impact of the errors from its prediction results can be effectively suppressed.

To quantitatively assess the accuracy of ISB samples and semi-supervised training samples (SST), we randomly selected and manually annotated 300 samples (in patches) for accuracy assessment. The results are presented in Table 9. The high precision of SST suggests that the proposed semi-supervised sample generation method can effectively

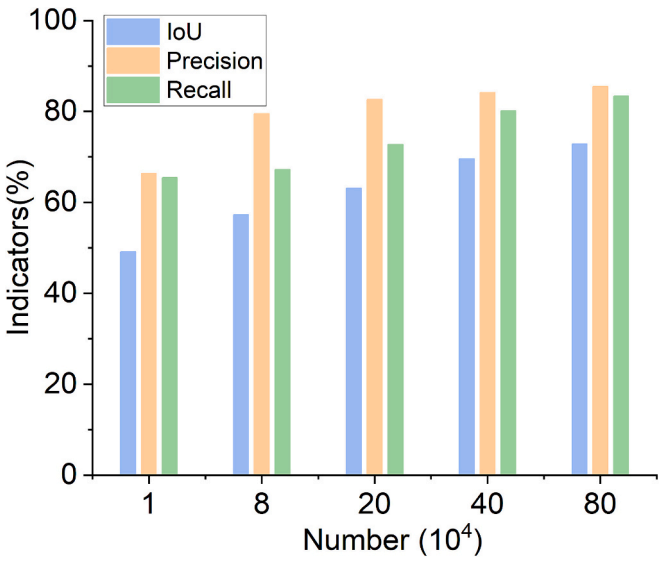

Fig. 21. Accuracy of the model with different sample sizes.

Table 10
The efficacy of different models with different scaling-up strategies.

Method	IoU(%)	Params(M)	FPS
BldgNet-small	70.16	29.7	102
BldgNet-base	72.84	49.5	88
BldgNet-large	73.01	66.3	69

extract accurate building samples from ISB. Meanwhile, the recall of SST samples is 95.10%, which is 58.39% higher than the original ISB, signifying that the semi-supervised method can effectively mask the building omission regions of ISB samples by leveraging the BldgNet predictions.

6.6. Robustness of the proposed model

To evaluate the robustness and fitting ability of the proposed model, we varied the number of training samples from the ASB dataset and assessed the model's accuracy on the test set (the same dataset used for evaluating CBF). The results are depicted in Fig. 21. It can be observed that as the number of samples increases, the model accuracy rises. However, the accuracy increments diminish as the number of samples continues to increase. This result can be served as a reference for constrcuting the large-scale building sample set.

Here we discuss the effects of the scaling-up strategy. Specifically, we vary the number of layers of the Transformer Block to obtain 3 models of different sizes. BldgNet-small indicates that the number of layers of the four Transformer Blocks in the encoder is 3, 4, 6, and 3, respectively. BldgNet-base represents that the number of layers is 3, 4, 18, and 3, and BldgNet-large indicates that the number of layers is 3, 8, 27, and 3 in order. The results (see Table 10) show that model scaling-up can improve the model accuracy at the expense of efficiency. BldgNet-base can significantly enhance the model accuracy (an increment of 2.7% for the IoU) relative to BldgNet-small, although the former also increases the model parameters and running time. On the other hand, however, BldgNet-large substantially increases the model complexity and runtime relative to BldgNet-base, but does not significantly improve the model accuracy (<0.2%). Therefore, in this study, we adopt the BldgNet-base by comprehensively considering the efficacy of the models.

6.7. Limitations and future directions

Although we use sub-meter imagery (0.5 m) as input, there are still a few closely connected urban villages identified as a whole (Fig. 22(a)). We attempted to segment them into individual buildings by considering the variations of image gradients(Comaniciu and Meer, 2002), as shown in Fig. 22(b). However, this method requires manual adjustment of parameters and is difficult to apply on a large scale. This issue might result in a slight underestimation of the number of buildings in CBF. Moreover, based on the CBF dataset, building attributes, such as time and height, can be estimated with additional relevant datasets (e.g., GISA(Global Impervious Surface Area(Huang et al., 2021)), CNBH (Chinese Building Height(Wu et al., 2023)). In the future, it is possible to retrieval high-precision building time and height information.

In rural areas, extracting scattered, small-sized buildings is challenging. As discussed in Section 5.1, the accuracy of CBF is lower in rural and Type III cities compared to other areas. This can be attributed to the fact that OSM data predominantly covers urban regions, resulting in fewer samples available for rural and Type III cities. In future research, we aim to strengthen the quantity and quality of samples in these regions. Moreover, in this study, we utilized an existing impervious surface product (GISA) to mask non-building areas (e.g., deserts, lakes) to expedite the mapping. However, this approach may overlook a small number of buildings. Therefore, in future, we also plan to involve all the high-resolution imagery of the study area, in order to eliminate the potential omissions.

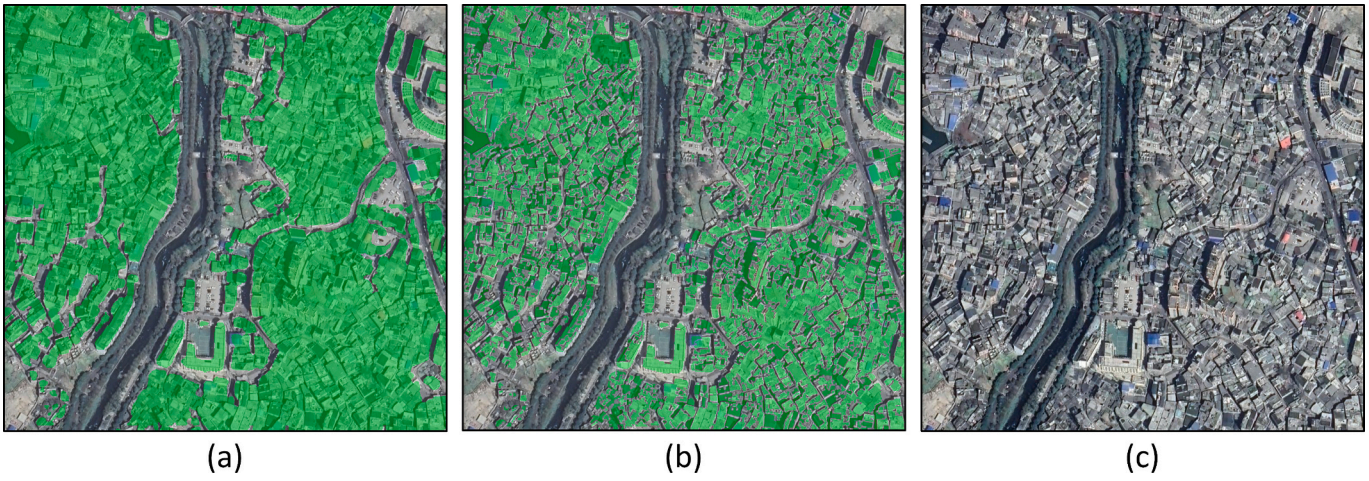

Fig. 22. Examples of buildings that are misconnected in the prediction results. (a) Original prediction result. (b) Processed result. (c) Google imagery.

7. Conclusion

In this paper, we generated the CBF (China Building Footprint) dataset, which is the first open-sourced sub-meter building footprint data of China. A large amount of diverse and high-quality training samples are crucial for achieving accurate building extraction over a large-scale study area (e.g., China). Therefore, in this study, a semi-automated procedure was proposed for constructing a global building sample dataset (GBD). This dataset can serve as a valuable sample resource for building mapping worldwide.

From the perspective of technologies, we proposed a framework for accurate and robust building extraction, and offered applicable solutions for the difficulties and challenges faced by deep learning based building extraction over a large scale. Therefore, the proposed BldgNet included a series of novel modules: The LWA (Large Window Attention) module improves the acquisition of global and contextual information, thereby improving the extraction performance of buildings with different sizes. The EA (Edge Attention) module improves the extraction of building boundaries, and the DASCI (Distribution Alignment Module with consideration of spatial contextual information) alleviates the issue of foreground-background imbalance. The experimental results also demonstrated that the proposed BldgNet can achieve a balance between accuracy and efficiency.

CRediT authorship contribution statement

Xin Huang: Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Zhen Zhang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Jiayi Li:** Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, 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.

Data availability

The GBD and CBF datasets can be publicly available and downloadable via https://zenodo.org/doi/10.5281/zenodo.10043351.

Acknowledgments

The research was supported by the National Natural Science Foundation of China (under Grants 42271328 and 42071311).

References

Ahn, J., Kwak, S., 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 4981–4990. https://doi.org/10.1109/CVPR.2018.00523.

Appolloni, E., Orsini, F., Specht, K., Thomaier, S., Sanyé-Mengual, E., Pennisi, G., Gianquinto, G., 2021. The global rise of urban rooftop agriculture: a review of worldwide cases. J. Clean. Prod. 296 https://doi.org/10.1016/j.jclepro.2021.126556.

Borck, R., 2016. Will skyscrapers save the planet? Building height limits and urban greenhouse gas emissions. Reg. Sci. Urban Econ. 58, 13–25. https://doi.org/10.1016/j.regsciurbeco.2016.01.004.

Buda, M., Maki, A., Mazurowski, M.A., 2018. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw. 106, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011.

Byrne, J., Taminiau, J., Kurdgelashvili, L., Kim, K.N., 2015. A review of the solar city concept and methods to assess rooftop solar electric potential, with an illustrative application to the city of Seoul. Renew. Sust. Energ. Rev. 41, 830–844. https://doi.org/10.1016/j.rser.2014.08.023.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Adv. Neural Inf. Proces. Syst. 32.

Cao, Y., Huang, X., 2021. A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: a case study of 42 Chinese cities. Remote Sens. Environ. 264, 112590.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, P., Huang, H., Liu, J., Wang, J., Liu, C., Zhang, N., Su, M., Zhang, D., 2023a. Leveraging Chinese GaoFen-7 imagery for high-resolution building height estimation in multiple cities. Remote Sens. Environ. 298 https://doi.org/10.1016/j.rse.2023.113802.

Chen, S., Ogawa, Y., Zhao, C., Sekimoto, Y., 2023b. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. ISPRS J. Photogramm. Remote Sens. 195, 129–152. https://doi.org/10.1016/j.isprsjprs.2022.11.006.

Comaniciu, D., Meer, P., 2002. Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619. https://doi.org/10.1109/34.1000236.

Conn, B., Arandjelovic, O., 2017. Towards computer vision based ancient coin recognition in the wild - automatic reliable image preprocessing and normalization. In: Proc. Int. Jt. Conf. Neural Networks 2017-May, pp. 1457–1464. https://doi.org/10.1109/IJCNN.2017.7966024.

Corbane, C., Syrris, V., Sabo, F., Politis, P., Melchiorri, M., Pesaresi, M., Soille, P., Kemper, T., 2021. Convolutional neural networks for global human settlements mapping from Sentinel-2 satellite imagery. Neural Comput. & Applic. 33, 6697–6720. https://doi.org/10.1007/s00521-020-05449-7.

Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S., 2019. Class-balanced loss based on effective number of samples. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2019-June, pp. 9260–9269. https://doi.org/10.1109/CVPR.2019.00949.

Dong, X., Bao, J., Chen, Dongdong, Zhang, W., Yu, N., Yuan, L., Chen, Dong, Guo, B., 2021. CSWin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12124–12134.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. ICLR 2021 - 9th Int. Conf. Learn. Represent.

Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L., 2023. Teaching Structured Vision & Language Concepts to Vision & Language Models, pp. 2657–2668. https://doi.org/10.1109/cvpr52729.2023.00261.

Frantz, D., Schug, F., Okujeni, A., Navacchi, C., Wagner, W., van der Linden, S., Hostert, P., 2021. National-scale mapping of building height using Sentinel-1 and Sentinel-2 time series. Remote Sens. Environ. 252, 112128 https://doi.org/10.1016/j.rse.2020.112128.

Frolking, S., Mahtta, R., Milliman, T., Seto, K.C., 2022. Three decades of global trends in urban microwave backscatter, building volume and city GDP. Remote Sens. Environ. 281, 113225 https://doi.org/10.1016/j.rse.2022.113225.

Ge, W., Huang, W., Guo, S., Scott, M., 2019. Label-PEnet: sequential label propagation and enhancement networks for weakly supervised instance segmentation. In: Proc. IEEE Int. Conf. Comput. Vis. 2019-Octob, pp. 3344–3353. https://doi.org/10.1109/ICCV.2019.00344.

Gong, P., Li, X., Wang, J., Bai, Y., Chen, B., Hu, T., Liu, X., Xu, B., Yang, J., Zhang, W., Zhou, Y., 2020. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. Remote Sens. Environ. 236, 111510 https://doi.org/10.1016/j.rse.2019.111510.

Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing. Springer, pp. 878–887.

He, H., Li, X., Cheng, G., Shi, J., Tong, Y., Meng, G., Prinet, V., Weng, L. Bin, 2021. Enhanced boundary learning for glass-like object segmentation. Proc. IEEE Int. Conf. Comput. Vis. 15839–15848. https://doi.org/10.1109/ICCV48922.2021.01556.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., Zipf, A., 2023. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap. Nat. Commun. 14, 1–14. https://doi.org/10.1038/s41467-023-39698-6.

Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T., 2019. Axial attention in multidimensional transformers. arXiv preprint. .arXiv1912.12180.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., 2019. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1314–1324.

Hsu, C.C., Hsu, K.J., Tsai, C.C., Lin, Y.Y., Chuang, Y.Y., 2019. Weakly supervised instance segmentation using the bounding box tightness prior. Adv. Neural Inf. Proces. Syst. 32, 1–12.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 7132–7141.

Huang, C., Loy, C.C., Tang, X., 2016. Learning deep representation for mood classification in microblog. In: 2016 IEEE Conf. Comput. Vis. Pattern Recognit.

Huang, S., Shen, Z., Huang, Z., Ding, Z., Dai, J., Han, J., Wang, N., Liu, S., 2023. Anchor3DLane: Learning to Regress 3D Anchors for Monocular 3D Lane Detection 17451–17460. https://doi.org/10.1109/cvpr52729.2023.01674.

Huang, X., Li, J., Yang, J., Zhang, Z., Li, D., Liu, X., 2021. 30 m global impervious surface area dynamics and urban expansion pattern observed by Landsat satellites: from 1972 to 2019. Sci. China Earth Sci. 64, 1922–1933. https://doi.org/10.1007/s11430-020-9797-9.

Huang, X., Yang, J., Wang, W., Liu, Z., 2022. Mapping 10 m global impervious surface area (GISA-10m) using multi-source geospatial data. Earth Syst. Sci. Data 14, 3649–3672. https://doi.org/10.5194/essd-14-3649-2022.

Jampani, V., Sun, D., Liu, M.Y., Yang, M.H., Kautz, J., 2018. Superpixel sampling networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11211 LNCS, 363–380. https://doi.org/10.1007/978-3-030-01234-2_22.

Ji, S., Wei, S., Lu, M., 2019. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57, 574–586. https://doi.org/10.1109/TGRS.2018.2858817.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y., 2020. Decoupling representation and classifier for long-tailed recognition. In: 8th Int. Conf. Learn. Represent. ICLR 2020, pp. 1–16.

Kang, W., Xiang, Y., Wang, F., You, H., 2019. EU-net: an efficient fully convolutional network for building extraction from optical remote sensing images. Remote Sens. 11, 2813.

Khan, S.H., Hayat, M., Bennamoun, M., Sohel, F.A., Togneri, R., 2018. Cost-sensitive learning of deep feature representations from imbalanced data. IEEE Trans. Neural Networks Learn. Syst. 29, 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: weakly supervised instance and semantic segmentation. In: Proc. -30th IEEE Conf. Comput. Vis. Pattern recognition, CVPR 2017 2017-Janua, pp. 1665–1674. https://doi.org/10.1109/CVPR.2017.181.

Kirillov, A., Wu, Y., He, K., Girshick, R., 2020. Pointrend: image segmentation as rendering. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, pp. 9796–9805. https://doi.org/10.1109/CVPR42600.2020.00982.

Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., Jia, J., 2021. Semi-supervised semantic segmentation with directional context-aware consistency. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1205–1214. https://doi.org/10.1109/CVPR46437.2021.00126.

Li, L., Liang, J., Weng, M., Zhu, H., 2018b. A multiple-feature reuse network to extract buildings from remote sensing imagery. Remote Sens. 10, 1350.

Li, Q., Arnab, A., Torr, P.H.S., 2018a. Weakly- and semi-supervised panoptic segmentation. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11219 LNCS, 106–124. https://doi.org/10.1007/978-3-030-01267-0_7.

Li, Xiangtai, Li, Xia, Zhang, L., Cheng, G., Shi, J., Lin, Z., Tan, S., Tong, Y., 2020. Improving semantic segmentation via decoupled body and edge supervision. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 12362 LNCS, 435–452. https://doi.org/10.1007/978-3-030-58520-4_26.

Li, Xuecao, Zhou, Y., Gong, P., Seto, K.C., Clinton, N., 2020a. Developing a method to estimate building height from Sentinel-1 data. Remote Sens. Environ. 240, 111705 https://doi.org/10.1016/j.rse.2020.111705.

Li, Xuecao, Gong, P., Zhou, Y., Wang, J., Bai, Y., Chen, B., Hu, T., Xiao, Y., Xu, B., Yang, J., Liu, X., Cai, W., Huang, H., Wu, T., Wang, X., Lin, P., Li, Xun, Chen, J., He, C., Li, Xia, Yu, L., Clinton, N., Zhu, Z., 2020b. Mapping global urban boundaries from the global artificial impervious area (GAIA) data. Environ. Res. Lett. 15 https://doi.org/10.1088/1748-9326/ab9be3.

Lin, D., Dai, J., He, K., 2016. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation the Chinese Univeristy of Hong Kong. Cvpr 3159–3167.

Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C., 2022. Adaptive early-learning correction for segmentation from noisy annotations. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2022-June, pp. 2596–2606. https://doi.org/10.1109/CVPR52688.2022.00263.

Liu, X., Huang, Y., Xu, X., Li, Xuecao, Li, Xia, Ciais, P., Lin, P., Gong, K., Ziegler, A.D., Chen, A., Gong, P., Chen, J., Hu, G., Chen, Y., Wang, S., Wu, Q., Huang, K., Estes, L., Zeng, Z., 2020. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. Nat. Sustain. 3, 564–570. https://doi.org/10.1038/s41893-020-0521-x.

Liu, Y.J., Yu, M., Li, B.J., He, Y., 2018. Intrinsic manifold SLIC: a simple and efficient method for computing content-sensitive superpixels. IEEE Trans. Pattern Anal. Mach. Intell. 40, 653–666. https://doi.org/10.1109/TPAMI.2017.2686857.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Liu, Z., Tang, H., Feng, L., Lyu, S., 2023. China building rooftop area: the first multi-annual (2016-2021) and high-resolution (2.5â€¯m) building rooftop area dataset in China derived with super-resolution segmentation from Sentinel-2 imagery. Earth Syst. Sci. Data 15, 3547–3572. https://doi.org/10.5194/essd-15-3547-2023.

Ma, Xiao, Zheng, G., Chi, X., Yang, L., Geng, Q., Li, J., Qiao, Y., 2023. Mapping fine-scale building heights in urban agglomeration with spaceborne lidar. Remote Sens. Environ. 285, 113392 https://doi.org/10.1016/j.rse.2022.113392.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In: Int. Geosci. Remote Sens. Symp. 2017-July, pp. 3226–3229. https://doi.org/10.1109/IGARSS.2017.8127684.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., van der Maaten, L., 2018. Exploring the limits of weakly supervised Pretraining. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 11206 LNCS, 185–201. https://doi.org/10.1007/978-3-030-01216-8_12.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E., 2020. Outlining where humans live, the world settlement footprint 2015. Sci. Data 7, 1–14. https://doi.org/10.1038/s41597-020-00580-5.

Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S., 2020. Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314.

Microsoft, 2023. GlobalMLBuildingFootprints [WWW Document]. URL. https://github.com/microsoft/GlobalMLBuildingFootprints (accessed 11.19.23).

Nadal, A., Alamús, R., Pipia, L., Ruiz, A., Corbera, J., Cuerva, E., Rieradevall, J., Josa, A., 2017. Urban planning and agriculture. Methodology for assessing rooftop greenhouse potential of non-residential areas using airborne sensors. Sci. Total Environ. 601–602, 493–507. https://doi.org/10.1016/j.scitotenv.2017.03.214.

National Bureau of Statistics, 2021. Statistical Tables on Economic and Social Development [WWW Document]. URL. https://www.stats.gov.cn/xw/tjxw/spxw/202302/t20230221_1914279.html (accessed 9.29.21).

Oh, Y., Kim, B., Ham, B., 2021. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit, pp. 6909–6918. https://doi.org/10.1109/CVPR46437.2021.00684.

Pu, M., Huang, Y., Liu, Y., Guan, Q., Ling, H., 2022. EDTER: edge detection with transformer. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2022-June, pp. 1392–1402. https://doi.org/10.1109/CVPR52688.2022.00146.

Ren, M., Zeng, W., Yang, B., Urtasun, R., 2018. Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning. PMLR, pp. 4334–4343.

Resch, E., Bohne, R.A., Kvamsdal, T., Lohne, J., 2016. Impact of urban density and building height on energy use in cities. Energy Procedia 96, 800–814. https://doi.org/10.1016/j.egypro.2016.09.142.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Shen, L., Lin, Z., Huang, Q., 2016. Relay backpropagation for effective learning of deep convolutional neural networks. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 9911 LNCS, 467–482. https://doi.org/10.1007/978-3-319-46478-7_29.

Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X., Tian, Q., 2023. A survey on label-efficient deep image segmentation: bridging the gap between weak supervision and dense prediction. IEEE Trans. Pattern Anal. Mach. Intell. 45, 9284–9305. https://doi.org/10.1109/TPAMI.2023.3246102.

Shen, Y., Cao, L., Chen, Z., Lian, F., Zhang, B., Su, C., Wu, Y., Huang, F., Ji, R., 2021. Toward joint thing-and-stuff mining for weakly supervised panoptic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 16689–16700. https://doi.org/10.1109/CVPR46437.2021.01642.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y.S.E., Dauphin, Y., Keysers, D., Neumann, M., Cisse, M., Quinn, J., 2021. Continental-Scale Building Detection from High Resolution Satellite Imagery 1–15.

Stutz, D., Hermans, A., Leibe, B., 2018. Superpixels: an evaluation of the state-of-the-art. Comput. Vis. Image Underst. 166, 1–27. https://doi.org/10.1016/j.cviu.2017.03.007.

Sun, X., Yin, D., Qin, F., Yu, H., Lu, W., Yao, F., He, Q., Huang, X., Yan, Z., Wang, P., Deng, C., Liu, N., Yang, Y., Liang, W., Wang, R., Wang, C., Yokoya, N., Hänsch, R., Fu, K., 2023. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. Nat. Commun. 14, 1–13. https://doi.org/10.1038/s41467-023-37136-1.

Tan, M., Le, Q., 2019. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tang, K., Huang, J., Zhang, H., 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. Adv. Neural Inf. Proces. Syst. 33, 1513–1524.

Tarzanagh, D.A., Li, Y., Thrampoulidis, C., Oymak, S., 2023. Transformers as Support Vector Machines, 3, pp. 1–62.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., Dosovitskiy, A., 2021. MLP-mixer: an all-MLP architecture for vision. Adv. Neural Inf. Proces. Syst. 29, 24261–24272.

Wang, H., Wang, Q., Yang, F., Zhang, W., Zuo, W., 2019. Data Augmentation for Object Detection Via Progressive and Selective Instance-Switching.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., 2020c. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43, 3349–3364.

Wang, R., Qin, J., Li, K., Li, Y., Cao, D., Xu, J., 2022. BEV-LaneDet: A Simple and Effective 3D Lane Detection Baseline.

Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J., 2020b. The devil is in classification: a simple framework for long-tail instance segmentation. In: European Conference on Computer Vision. Springer, pp. 728–744.

Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F., 2020a. Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957.

Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X., 2020d. Self-supervised Equivariant attention mechanism for weakly supervised semantic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 12272–12281. https://doi.org/10.1109/CVPR42600.2020.01229.

Wang, Y.-X., Ramanan, D., Hebert, M., 2017. Learning to Model the Tail. Adv. Neural Inf. Proces. Syst. 30.

Wang, Y.-X., Girshick, R., Hebert, M., Hariharan, B., 2018. Low-shot learning from imaginary data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7278–7286.
Wu, G., Shao, X., Guo, Z., Chen, Q., Yuan, W., Shi, X., Xu, Y., Shibasaki, R., 2018. Automatic building segmentation of aerial imagery using multi-constraint fully convolutional networks. Remote Sens. 10, 407.
Wu, J., Zhou, C., Zhang, Q., Yang, M., Yuan, J., 2020b. Self-mimic learning for small-scale pedestrian detection. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2012–2020.
Wu, T.-Y., Morgado, P., Wang, P., Ho, C.-H., Vasconcelos, N., 2020a. Solving long-tailed recognition with deep realistic taxonomic classifier. In: European Conference on Computer Vision. Springer, pp. 171–189.
Wu, W. Ben, Ma, J., Banzhaf, E., Meadows, M.E., Yu, Z.W., Guo, F.X., Sengupta, D., Cai, X.X., Zhao, B., 2023. A first Chinese building height estimate at 10 m resolution (CNBH-10 m) using multi-source earth observations and machine learning. Remote Sens. Environ. 291, 113578 https://doi.org/10.1016/j.rse.2023.113578.
Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Proces. Syst. 15, 12077–12090.
Yan, H., Zhang, C., Wu, M., 2022. Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations Via Large Window Attention.
Yang, C., Zhao, S., 2022. A building height dataset across China in 2017 estimated by the spatially-informed approach. Sci. Data 9, 1–11. https://doi.org/10.1038/s41597-022-01192-x.
Yin, D., Gao, F., Thattai, G., Johnston, M., Chang, K.-W., 2023. GIVL: Improving Geographical Inclusivity of Vision-Language Models with Pre-Training Methods 10951–10961. https://doi.org/10.1109/cvpr52729.2023.01054.
Yuan, Y., Fu, R., Huang, L., Lin, W., Zhang, C., Chen, X., Wang, J., 2021. Hrformer: high-resolution vision transformer for dense predict. Adv. Neural Inf. Proces. Syst. 34, 7281–7293.
Zhang, S., Li, Z., Yan, S., He, X., Sun, J., 2021. Distribution alignment: a unified framework for long-tail visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2361–2370.
Zhang, W., Huang, Z., Luo, G., Chen, T., Wang, X., Liu, W., Yu, G., Shen, C., 2022a. TopFormer: token pyramid transformer for mobile semantic segmentation. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2022-June, pp. 12073–12083. https://doi.org/10.1109/CVPR52688.2022.01177.
Zhang, Z., Qian, Z., Zhong, T., Chen, M., Zhang, K., Yang, Y., Zhu, R., Zhang, F., Zhang, H., Zhou, F., Yu, J., Zhang, B., Lü, G., Yan, J., 2022b. Vectorized rooftop area data for 90 cities in China. Sci. Data 9, 3–5. https://doi.org/10.1038/s41597-022-01168-x.
Zhang, Z., Huang, X., Li, J., 2023. DWin-HRFormer: a high-resolution transformer model with directional windows for semantic segmentation of urban construction land. IEEE Trans. Geosci. Remote Sens. 61 https://doi.org/10.1109/TGRS.2023.3241366.
Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890.
Zhou, Y., Zhu, Y., Ye, Q., Qiu, Q., Jiao, J., 2018. Weakly supervised instance segmentation using class peak response. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3791–3800. https://doi.org/10.1109/CVPR.2018.00399.
Zhou, Y., Li, X., Chen, W., Meng, L., Wu, Q., Gong, P., Seto, K.C., 2022. Satellite mapping of urban built-up heights reveals extreme infrastructure gaps and inequalities in the global south. Proc. Natl. Acad. Sci. USA 119, 1–9. https://doi.org/10.1073/pnas.2214813119.
Zou, Y., Zhang, Z., Zhang, H., Li, C.L., Bian, X., Huang, J. Bin, Pfister, T., 2021. Pseudoseg: Designing Pseudo Labels for Semantic Segmentation. In: ICLR 2021 - 9th Int. Conf. Learn. Represent, vol. 2, pp. 1–18.