ISPRS Journal of Photogrammetry and Remote Sensing 195 (2023) 129–152

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach

Shenglong Chen [a,*], Yoshiki Ogawa [b], Chenbo Zhao [a], Yoshihide Sekimoto [b]

[a] Department of Civil Engineering, The University of Tokyo, 153-8505 Tokyo, Japan
[b] Center for Spatial Information Science (CSIS), The University of Tokyo, 153-8505 Tokyo, Japan

ARTICLE INFO

Keywords:
Building extraction
Instance segmentation
Super-resolution
Transformer
Remote sensing
Open data

ABSTRACT

Building footprint is a primary dataset of an urban geographic information system (GIS) database. Therefore, it is essential to establish a robust and automated framework for large-scale building extraction. However, the characteristic of remote sensing images complicates the application of the instance segmentation method based on the Mask R-CNN model, which ought to be improved toward extracting and fusing multi-scale features. Moreover, open-source satellite image datasets with wider spatial coverage and temporal resolution than high-resolution images may exhibit different coloration and resolution. This study proposes a large-scale building extraction framework based on super-resolution (SR) and instance segmentation using a relatively lower-resolution (>0.6 m) open-sourced dataset. The framework comprises four steps: color normalization and image super-resolution, scene classification, building extraction, and scene mosaicking. We took Hyogo Prefecture, Japan (19,187 km^2) as a test area and extracted 1,726,006 (29.12 km^2) of the 3,301,488 buildings (32.46 km^2), where the number of buildings and footprint area increased by 3.0 % and 5.0 % respectively. The result indicated that the color normalization and image super-resolution could improve the visual quality of open-source satellite images and contribute to building extraction accuracy. Moreover, the improved Mask R-CNN based on Multi-Path Vision Transformer (MPViT) backbone achieved F1 scores of 0.71, 0.70, 0.81, and 0.67 for non-built-up, rural, suburban, and urban areas, respectively, which is better than those of the baseline model and other mainstream instance segmentation approaches. This study demonstrates the potential of acquiring acceptable building footprint maps from open-source satellite images, which has significant practical implications.

1. Introduction

Building footprint is a primary dataset in an urban geographic information system (GIS) database, which varies frequently. The use of remote sensing imagery with rich features to dynamically extract and grasp information on buildings is essential for urban planning, map production, population estimation, and disaster response (Gupta and Shah, 2021; Hong et al., 2020; Huang et al., 2019a; Wei et al., 2021; Zhai and Chen, 2020). However, there is a vast variation in the size, texture, color, and shape of buildings in different regions. Moreover, owing to the complexity of the background (e.g., shadows and other artificial objects such as buildings) and diversity in the remote sensing image sources (e.g., different spatial resolution and capture time), establishing a reliable building database for a wide area poses significant challenges. The building footprints used in GIS systems are primarily created manually by surveyors or derived from voluntary geographic information (VGI) (Haklay and Weber, 2008). Therefore, it is essential to establish a robust and automated framework for generating large-scale building footprints.

In recent years, with the rapid development of deep learning technology in the field of computer vision, deep learning models represented by convolutional neural networks (CNN) are gradually introduced into the field of remote sensing and show exciting potential for tasks such as image classification and object detection (Cheng et al., 2020; O'Shea and Nash, 2015; Paisitkriangkrai et al., 2015). Deep learning models can automatically learn and interpret more abstract high-dimensional features from low-dimensional features present in the original image, which undoubtedly provides a massive advantage in acquiring the complex spectral, textural, and geometric features in remote sensing images (Yuan et al., 2020). Many researchers have used semantic

* Corresponding author at: Department of Civil Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
E-mail address: chen-sl@csis.u-tokyo.ac.jp (S. Chen).

https://doi.org/10.1016/j.isprsjprs.2022.11.006
Received 11 April 2022; Received in revised form 11 October 2022; Accepted 12 November 2022
Available online 30 November 2022

segmentation approaches to achieve efficient and automatic building extraction from remote sensing images (Chen et al., 2021; Ji et al., 2018; Shao et al., 2020). However, the study of building extraction methodologies is not limited to semantic segmentation. Most practical applications focus not only on whether a pixel is a building but also on classifying different buildings individually to obtain the distribution and number of buildings, which is a typical instance segmentation task (Wagner et al., 2020).

Most state-of-the-art instance segmentation methods are based on the two-stage object detection model, such as Mask R-CNN, MS R-CNN, and RefineMask (He et al., 2017; Huang et al., 2019b; Zhang et al., 2021a). Mask R-CNN is a classical model and standard framework proven to be one of the most potent and adaptable deep learning models for different domains (Khan et al., 2021; Yu et al., 2019). The primary idea is to add a mask branch to the Faster R-CNN to predict the class of each pixel and use the region of interest (RoI) align, instead of ROI pooling to align the feature maps (Ren et al., 2015). However, remote sensing images have characteristics that are uncommon in other types of images. For example, satellite images of urban areas can easily contain many buildings concentrated in a single area. This characteristic complicates the application of ROI-based Mask R-CNN, primarily because: (1) Feature extraction models are required to extract complex image features owing to the large variability among buildings. (2) Buildings with large dimensions are easy to be extracted, whereas buildings with small sizes are prone to be omitted owing to the existence of multi-scale features of buildings on satellite images.

Moreover, with respect to data sources, high-resolution (HR) remote sensing imagery, such as Worldview, Quickbird, or UAV aerial images, were considered essential sources for building extraction (Latha et al., 2019; Sirko et al., 2021; Yang and Tang, 2020). HR images tend to exhibit richer texture detail and usually facilitate better separation between building and background, which is critical for producing high-quality building footprints (Haut et al., 2019). However, HR images are limited by spatial coverage, posing challenges for large-scale data use. Moreover, licensing agreements and temporal resolution limitations make it impossible to meet the usage requirements of some applications highly relevant to time continuity (e.g., long-term building change detection studies) (Dong and Shan, 2013). Alternatively, some open-sourced satellite images with the relatively lower spatial resolution are free to use and have better spatial and temporal accessibility (Huang and Jin, 2022; Touzani and Granderson, 2021). The availability of data is essential for large-scale building extraction. If these open data can be used to generate building footprint maps, the difficulty of obtaining HR images can be avoided. However, two challenges impede the full use of open-source data. First, lower image resolution yields lower accuracy and approximation of edges in segmentation results (Belgiu and Drăguţ, 2014). Second, open-source datasets may be multi-sourced, resulting in differences in capture time, color, and resolution between different images. Therefore, to extract individual building footprints from open-source datasets over broad areas with high accuracy, the image resolution must be improved, and the multi-sourced images should be homogenized.

This study proposes a framework for large-scale building extraction using a super-resolution-based instance segmentation algorithm to solve the application in open-source satellite images. The proposed framework comprises the following steps: First, we filter the images without buildings through an image classification model to improve the efficiency of large-scale building extraction. Subsequently, considering the relatively lower resolution (>0.6 m) and the tone diversity of the images in the open-source dataset, we apply color normalization and a super-resolution algorithm to mitigate the color differences between images from different sources and to increase the texture details (Cui et al., 2021; Wang et al., 2021b). Further, to improve the efficiency of large-scale building extraction, we classify the images with or without buildings prior to segmentation and optimize the resulting post-processing flow based on a modified scene mosaicking technique (Carvalho et al., 2020). Referring to the ideas of multi-scale feature enhancement and self-attention mechanism of the MPViT model, we improve the Mask R-CNN model to enhance the ability of multi-scale instance segmentation of buildings under different scenarios (He et al., 2017; Lee et al., 2021). To demonstrate the validity of the proposed framework, we considered the Hyogo prefecture, Japan, as the study area (19,187 km^2) and evaluated the extraction accuracy according to the building type and region, as the distribution of buildings is different for various areas.

The main contributions of this study:

(1) A proposed enhancement strategy for open-source satellite images based on color normalization and a super-resolution algorithm to improve image quality and instance-segmentation accuracy.
(2) Improvement of the Mask R-CNN model based on the MPViT backbone. It was experimentally proven to have better object-wise performance than the baseline model (ResNeXt-101 backbone), and other mainstream instance segmentation approaches. The experiment results of the test area indicated that the proposed framework could improve the performance of building instance segmentation from open-source images for a wide area. This can broaden the application of open-source data and benefit related downstream tasks.

The remainder of this paper is organized as follows: Section 2 introduces the related works. Sections 3 and 4 present the data we used in this study and the details of the proposed framework. In Section 5, we describe our experiments and the results obtained. Section 6 discusses the interpretations and implications of the results. Finally, Section 7 presents the conclusions drawn from the study.

2. Related work

2.1. Building extraction via instance segmentation

Since Ji et al. (2018) realized the extraction of >180,000 buildings for a wide area based on a Mask R-CNN model and achieved better pixel-wise accuracy than the FPN, the instance-based segmentation approach has proven its effectiveness for building extraction. There are three primary approaches for extract building at the instance level: (1) semantic segmentation, (2) contour based, and (3) end-to-end instance segmentation. In (1), semantic segmentation first classifies each image pixel as either a building or non-building; the subsequent post-processing extracts the instances (Wagner et al., 2020). Xu et al. (2021a) introduced the holistically nested network and attention mechanism into U-Net (HA U-Net). They also applied the watershed algorithm in post-processing to improve the problem of building adhesion. Girard et al. (2021) innovatively added a frame field output to the deep-segmentation model to improve segmentation via multi-task learning and promote the polygonization of footprints. However, semantic segmentation is trained pixel-wise, which tends to misidentify features similar to buildings. Unlike semantic segmentation, the emerging (2) contour-based approach directly extracts the edges and corners of buildings. In earlier works, some scholars tried to use recurrent neural networks (RNNs) to represent graph structures as closed polygons to predict the locations of key points in target objects (Acuna et al., 2018; Castrejon et al., 2017; Li et al., 2019). Wei and Ji (2021) first introduced the graph neural network (GCN) to directly predict and adjust building boundaries. These methods are difficult to train compared to CNN and can only output simple polygons without voids. This limits the extraction of complex buildings. In (3), the problem is decomposed into feature extraction, boundary box regression, and mask prediction stages. The end-to-end instance segmentation approach is direct and flexible, allowing the algorithm to obtain boundary boxes and segmentation masks and perform better (Wen et al., 2019). Therefore,

some researchers have designed novel algorithms to extract buildings from remote sensing images based on state-of-the-art instance segmentation methods. Wu et al. (2020) proposed an anchor-free instance segregation method based on CenterMask and balanced building extraction accuracy and efficiency. Wang et al. (2022) fused RGB images from an unmanned aerial vehicle (UAV) with visible light difference vegetation index (VDVI) features and Sobel edge detection features to improve the recognition accuracy of the Mask R-CNN model for rural building roofs. Liu et al. (2022a) proposed an MS-CNN model combing ResNeSt and Mask R-CNN, improving multiscale feature extraction capability using a fusion enhancement strategy and feature segmentation mechanism. However, most of the studies have focused on improving model performance. There are few frameworks for large-scale building extraction at the instance level. Moreover, these studies used HR remote sensing images (less than0.5 m) as the source data. Thus, the extraction and fusing of multi-scale features, on relatively lower resolution open-source datasets, requires further investigation to achieve high-accuracy instance segmentation.

2.2. Super-resolution-based image segmentation

Super-resolution (SR) techniques have emerged as practical strategies to address the problem of low-resolution images (LR) in open-source datasets (Glasner et al., 2009). Increasing the image resolution facilitates the use of finer spatial details than the original image. However, traditional interpolation-based SR methods, such as bicubic interpolation, tend to lose high-frequency spatial information, generating blurred HR images (Keys, 1981; Ledig et al., 2017). SR is also considered a one-to-many mapping from LR to HR space with multiple possible solutions. The recent advancements in the SR approach based on the generative adversarial network (GAN) are highly significant. GAN models that use perceptual loss and discriminator loss to improve and recover image authenticity, demonstrating the remarkable capability of deep learning techniques in SR applications (Ledig et al., 2017; Liang et al., 2021; Wang et al., 2018). However, the rich information from remote sensing images characterizes the spatial relationships of features. The classical degradation model is challenging to extend to the complex degradation process of remote sensing images, which easily cause artifacts (Xu et al., 2022). Zhang et al. (2021b) and Wang et al. (2021b) have introduced a more complex higher-order degradation modelling process, based on the ESRGAN network to improve the practicality of deep super-resolvers.

Additionally, some scholars have tried to integrate the SR algorithm into the building extraction process, and explored the impact of SR on downstream tasks. In general, there are two methods of super-resolution-based building extraction: (1) the end-to-end approach (Zhang et al., 2021d) and (2) the two-stage approach (Guo et al., 2019). The end-to-end approach integrates super-resolution and building extraction models in the same network, which enables better collaboration between two tasks. Zhang et al. (2021d) proposed an end-to-end SR semantic segmentation network (FSRSS-Net), which can generate a 2.5 m building distribution map from 10 m resolution images by upsampling the low- and high-level features through the deconvolution layer. Xu et al. (2021b) compared the end-to-end model with different two-stage models, which possess better accuracy when the image resolution is higher than 1.2 m. In the two-stage approach, the stages perform independently, usually obtaining the HR images or features with the SR module. They are used as inputs to the building extraction module to obtain the final HR segmentation maps. Guo et al. (2019) demonstrated that the first stage of the super-resolution module could mitigate the problems caused by the multi-source image, with different resolutions between the training and test data. Zhang et al. (2021c) proposed a two-stage framework, SRbuildingSeg, which can completely utilize the information of a provided LR image without relying on any external HR image. In contrast, the two-stage approach in this paper has a higher degree of freedom, and is not affected by model complexity. However, all the above studies are based on semantic segmentation. The SR methods for object-wise building extraction still lack quantitative evaluation and discussion. Thus, further research will be necessary.

3. Data

3.1. Study area

Several open-source datasets are widely used as the benchmark of building extraction by researchers, such as the WHU building dataset (Ji et al., 2018), the Massachusetts Buildings Dataset (Mnih, 2013) and the Inria Aerial Dataset (Maggiori et al., 2017). However, despite the WHU dataset covering 550 km^2, the most expansive area, it cannot meet the application needs of building extraction on a large scale. Moreover, owing to these datasets using HR images, the field of view of the down-sampled images differs considerably from the open-sourced image in actual use. Therefore, we selected Hyogo Prefecture, Japan, as the study area and open-sourced images as test data to demonstrate the validity of the proposed method in this study. Hyogo Prefecture is in the Kansai region, with a geographical area of 8400 km^2. The capital city, Kobe, is the second-largest urban area in Japan. Hyogo Prefecture envelopes various land use categories, including agricultural, residential, commercial, and industrial areas, with great diversity and complexity, which poses a significant challenge to building extraction. Fig. 1 shows the selected study areas and associated details. To facilitate downloading remote sensing images from open-source datasets, we selected the smallest rectangle surrounding Hyogo prefecture (as indicated in the green box), covering 19,187 km^2, as the test area. In addition, we selected representative areas from other parts of Japan as the training set (as indicated in red boxes), primarily located in Shinjuku, Setagaya, Hachioji, and Susono City.

3.2. Data preparation

Spatial resolution measures the richness of features contained in remote sensing images, which is vital for distinguishing buildings from the background. Lower resolutions usually result in less segmentation accuracy and rougher footprints. Furthermore, deep learning models trained on LR images are challenging to apply to images with higher resolutions (Hamaguchi and Hikosaka, 2018). Images from the same source for model training and testing can enhance the model's generalization ability. However, considering the resolution of the images in test data and the super-resolution step in the proposed framework, the satellite images from Google Earth (GE) with a spatial resolution of 0.3 m were selected as training data. All GE images were acquired on July 20, 2020. A polygon-based approach was used to annotate in ArcGIS Pro to best represent the building footprint, and any adjacent buildings with undetermined boundaries were annotated as a single building. Due to manual discrimination limitations, minor errors were inevitable, particularly in high-density building areas in cities. Finally, we collected 1222 aerial *ortho*-color images of dimensions 1024 × 1024 pixels and manually annotated over 186,000 high-quality building footprints. All images were in TIF format with three channels (RGB), and the corresponding labels were annotated according to the COCO annotation format (Lin et al., 2014). The data were assigned to the training and validation sets in the ratio of 70 %:30 %, respectively. The examples of training data for different areas are shown in Fig. 2. The number of image data for different regions is shown in Table 1.

For the test area, we selected the open-sourced seamless satellite image dataset from the Geospatial Information Authority of Japan (GSI)[1], as the cloud coverage of the dataset is less than 10 % and includes the entire region of Japan. The GSI seamless photo dataset contains PNG images obtained from multiple sources with a resolution ranging from 0.6 to 1 m. The dimensions of a single image are standardized to 256 ×

[1] URL: https://www.gsi.go.jp/tizu-kutyu.html (Accessed on 03/25/ 2022).

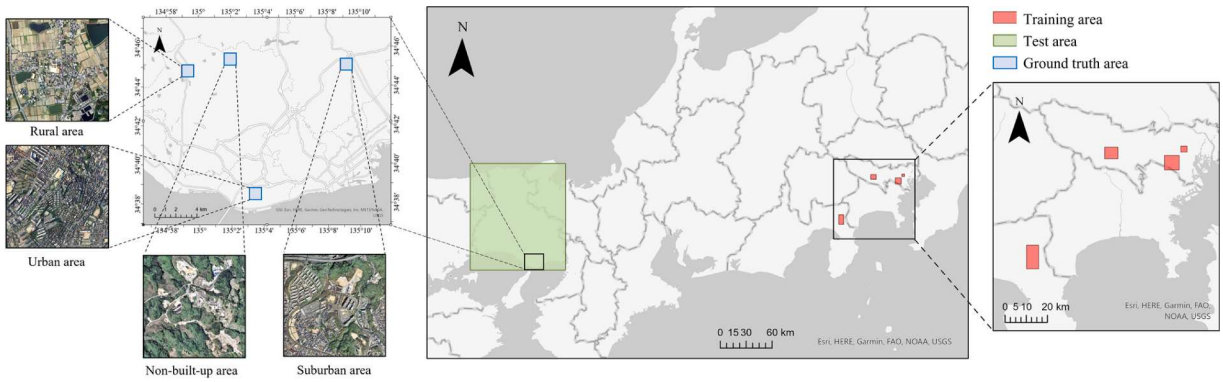

Fig. 1. Scope of the study area. The green box encloses the test area (Hyogo Prefecture, Japan). The red boxes enclose the training areas (Shinjuku, Setagaya, Hachioji, and Susono City). The blue boxes enclose the areas with ground truth for accuracy evaluation (non-built-up, rural, urban, and suburban areas, each selection of area 1 km^2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

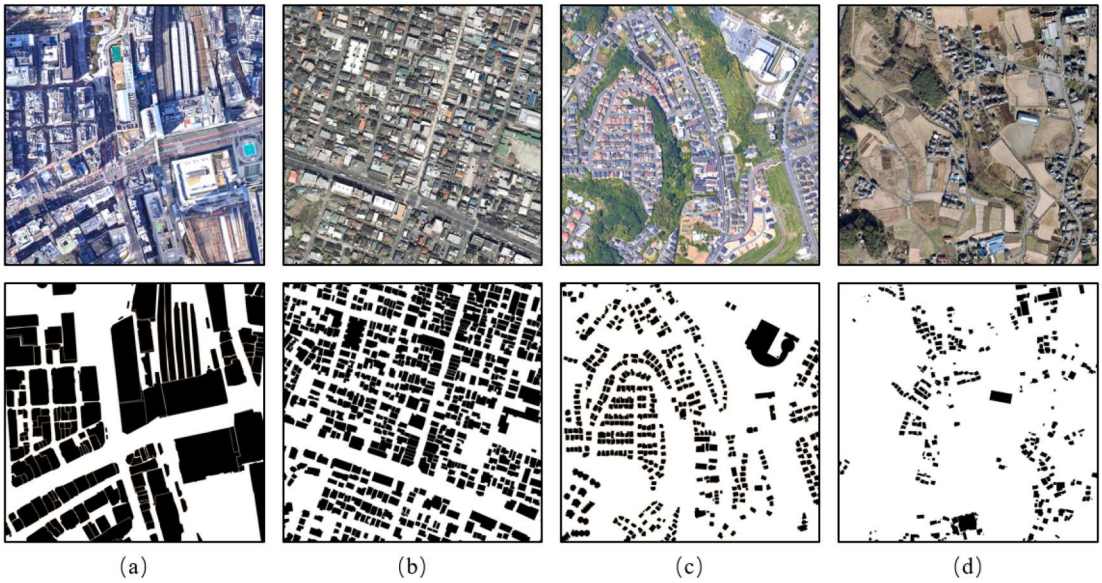

Fig. 2. Example of training data in different regions, including a) high-rise areas, b) urban area, c) suburban area and d) rural area. The first row is the GE satellite images, and the second is the ground truth of building footprints.

Table 1
Number of image data in different regions.

Area	Shinjuku	Setagaya	Hachioji	Susono
Training set	32	240	269	319
Validation set	8	103	115	136

256 pixels. The original GSI aerial imagery was captured by the digital aerial camera VEXCEL Ultracam D mounted on the aerial survey aircraft Kunikaze III (Sango, 2009). The Ultracam D camera can capture RGB and NIR 4-channel images with the highest resolution of 3 cm at a flying height of 300 m. As the dataset uses a tile coordinate system and does not contain geographic coordinates, georeferencing is necessary. The tile coordinates of the top-left point can be converted to latitudinal and longitudinal coordinates according to Equation (1):

$$\lambda = 180\left(\frac{x}{2^{z+7}} - 1\right), \tag{1}$$

$$\varphi = \frac{180}{\pi}\left(\sin^{-1}\left(\tanh\left(-\frac{\pi}{2^{z+7}}y + \tanh^{-1}\left(\sin\left(\frac{\pi}{180}L\right)\right)\right)\right)\right),$$

where x and y are the tile coordinates of the images; z is the zoom level; λ and φ are the longitudinal and latitudinal coordinates, respectively; L is a constant, equal to 85.05112878.

To satisfy the input requirements of neural networks and the efficient use of video memory, the size of remote sensing images in the dataset required to be unified. In this study, all test images were mosaicked together and cropped to the dimensions of 1024 × 1024 pixels with 20 % overlapping to avoid splitting buildings at the image edges. The defective parts were filled using black pixels. After cropping, the entire test area contained 81,438 patches. Moreover, buildings and other essential features in different areas showed different grey values, texture, density, and size characteristics. Therefore, to evaluate the accuracy and robustness of the prediction results, we randomly selected three 1-km^2 areas for non-built-up, rural, suburban, and urban regions in Hyogo Prefecture (the blue boxes in Fig. 1) and manually created ground truth to create a multi-scene test dataset. The prepared test dataset of Hyogo Prefecture is available at https://github.com/zensenlon/Building_extrction_Hyogo.

4. Method

The framework of large-scale building extraction, based on SR and instance segmentation, was proposed in this study. Fig. 3 shows the four

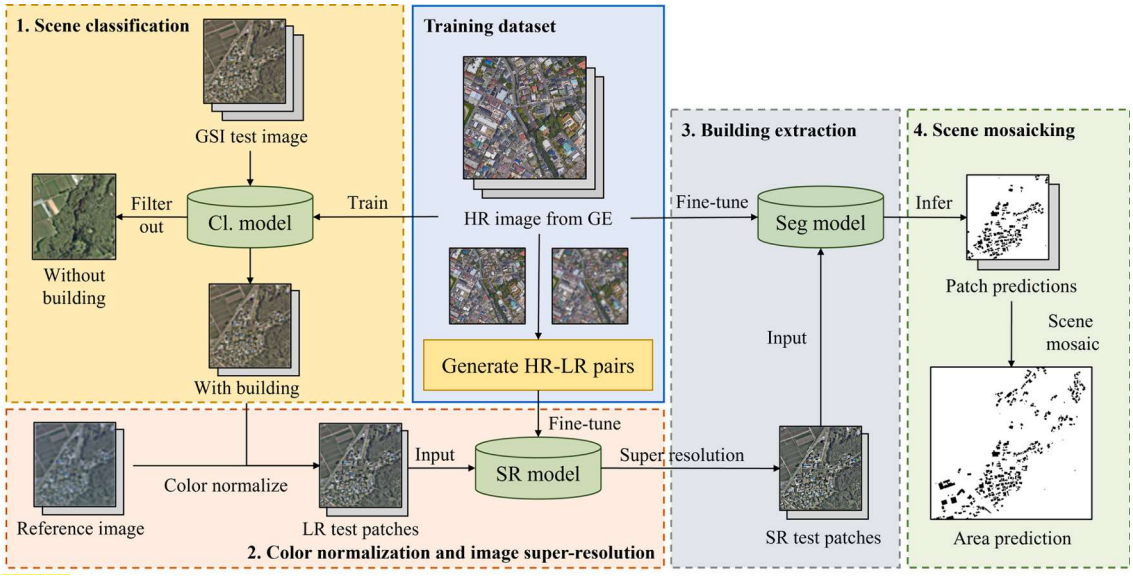

Fig. 3. The proposed building extraction framework, where the Cl., SR, and Seg models represent the scene classification, super-resolution (SR), and instance segmentation models, respectively. The HR, LR, GSI and GE represent the high-resolution, low-resolution, Geospatial Information Authority of Japan and Google Earth, respectively.

procedures constituting the proposed framework: (1) scene classification, (2) color normalization and image super-resolution, (3) building extraction, and (4) scene mosaicking. The model training part in the first three procedures can be considered independent. After acquiring and processing the satellite images from GE, the datasets for training scene classification, SR, and instance segmentation models can be created. Among them, only HR images are required for scene classification and the instance segmentation model, whereas the fine-tuning of SR models requires the generation of HR and LR image pairs. For testing, GSI images were first input into the scene classification model to filter the images containing only background. Subsequently, the image color was normalized based on the reference image, and the texture details were enhanced using the fine-tuned SR model. The SR test patches were subsequently input into the instance segmentation model to predict the building footprint. Finally, the patch predictions were stitched to obtain the footprint segmentation maps.

4.1. Scene classification

The target areas for building extraction were built-up areas, settlements, and other regions with human activity. However, such developed areas constitute only a tiny portion of the land. For example, in Hyogo Prefecture, 77 % of land cover is forest and agricultural areas, and only 7 % of land use is residential. Although some techniques, such as data augmentation and loss re-weighting, handle the effect of class imbalance, they are inefficient at performing inference on all remote sensing images of the target area. Furthermore, it is difficult to quickly and accurately determine the built-up area for areas where land-classification data are lacking. Therefore, we first filtered the images not containing buildings through scene classification to improve the inference efficiency.

Scene classification is one of the fundamental and challenging tasks in remote sensing. As large-scale remote sensing scene classification benchmarks emerged in recent years, CNN-based models have become the primary solutions because of their high accuracy (Cheng et al., 2020; Ma et al., 2021; Tang et al., 2021). Although the features of remote sensing images differ from natural images, existing studies have shown that transfer learning from the pre-trained model on large datasets can help improve the scene classification performance of CNN models (Das and Chandran, 2021; Pires de Lima and Marfurt, 2019). Therefore, in this study, we trained a two-class scene classification model based on ConvNeXt pre-trained on ImageNet to identify the presence or absence of buildings in images (Liu et al., 2022b).

ConvNeXt is one of the state-of-the-art backbone models that achieved a top-1 accuracy of 87.8 % on the ImageNet dataset. The model is built entirely from convolutional networks, with competitive results of Transformer in terms of accuracy and scalability while maintaining the simplicity and effectiveness of standard CNN (Vaswani et al., 2017). In terms of the network architecture, ConvNeXt develops from the original ResNet network and improves the model step by step by borrowing the design of Swin Transformer (Liu et al., 2021b). Fig. 4 shows the four stages of the ConvNeXt-based network, each providing output features at different scales. The ratio of the number of ConvNeXt blocks in each stage is 1:1:9:1. Furthermore, the initial down-sampling module uses a convolutional layer with a kernel size of 4×4 to form the Patchify layer. The ConvNeXt block uses a larger 7×7 kernel size and a different inverted bottleneck structure relative to the ResNet. In the microscopic design, the commonly used batch norm (BN) and ReLU activation function in CNN are replaced by the layer normalization (LN) layer and Gaussian error linear units (GELU) function (Equation (2)), respectively, used in Transformer, resulting in the reduced number of layers. Only one activation layer is retained between the block's two 1×1 convolution layers and the one LN layer before the convolutions layer. Further, a separate 2×2 convolutional downsampling layer with a stride of 2 is used, and an LN layer is added later to stabilize the training.

$$\mathrm{GELU}(x) = x^{*}\Phi(x) \tag{2}$$

In Equation (2), $\Phi(x)$ is the cumulative distribution function for Gaussian distribution.

In terms of the training data, although many researchers have developed and published scene recognition datasets for remote sensing images in recent years, such as the Million-AID (Long et al., 2021) and GID dataset (Tong et al., 2020), these scenes often contain multiple categories that are not suitable for our purpose. For example, the farmland scene is judged based on the appearance of the field and thus cannot be directly categorized as containing buildings or not. Therefore, we selected and created the scene classification dataset from existing data to completely use the available resources. As most of the existing images contain buildings (Fig. 5a), we supplemented the images from Google Earth which contain only background information (Fig. 5b). In

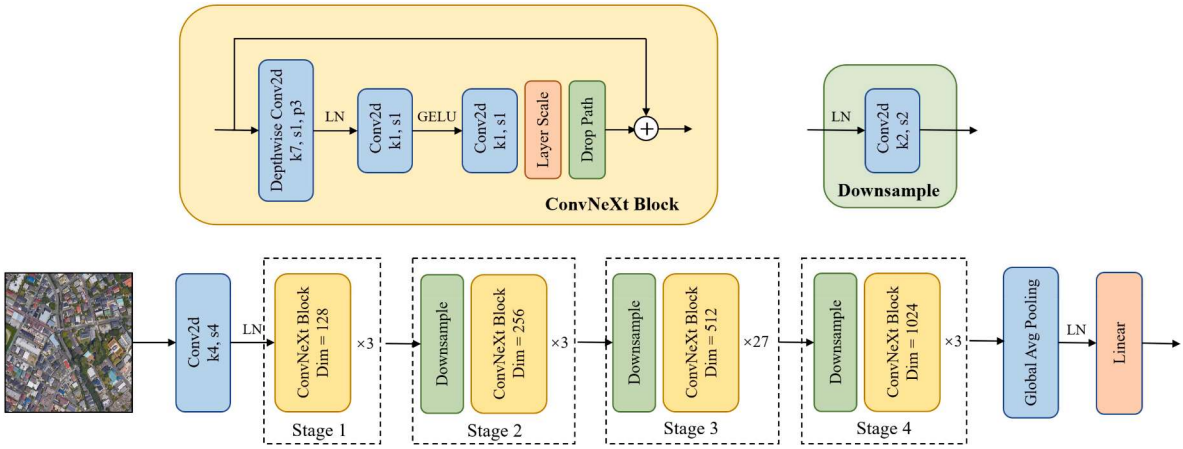

Fig. 4. Architecture of the ConvNeXt network, a four-stage feature hierarchy, was built to extract features of different scales. The downsample layer and ConvNeXt block are stacked at each stage in the ratio of 3:3:27:3. The LN and GELU represent the layer normalization and Gaussian error linear units.

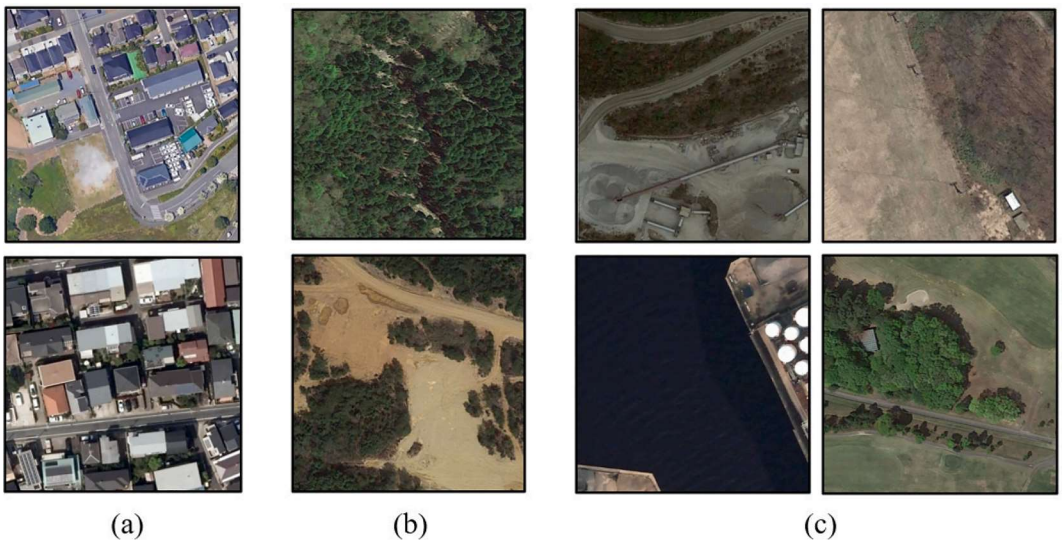

Fig. 5. Examples of the scene classification dataset. (a) Scenes with buildings; (b) scenes without buildings; (c) exceptional cases classified as scenes with buildings, including a construction site, LPG storage tank, and a cabin in the forest or meadow.

addition, although our dataset contained only two classes, the classification model was easy to confuse for some exceptional cases (e.g., features with similar semantic information to buildings such as liquefied petroleum gas (LPG) storage tanks and single cabins in a natural environment, as shown in Fig. 5c). As scene classification needed not to lead to omissions in building extraction, the exceptional scenes were

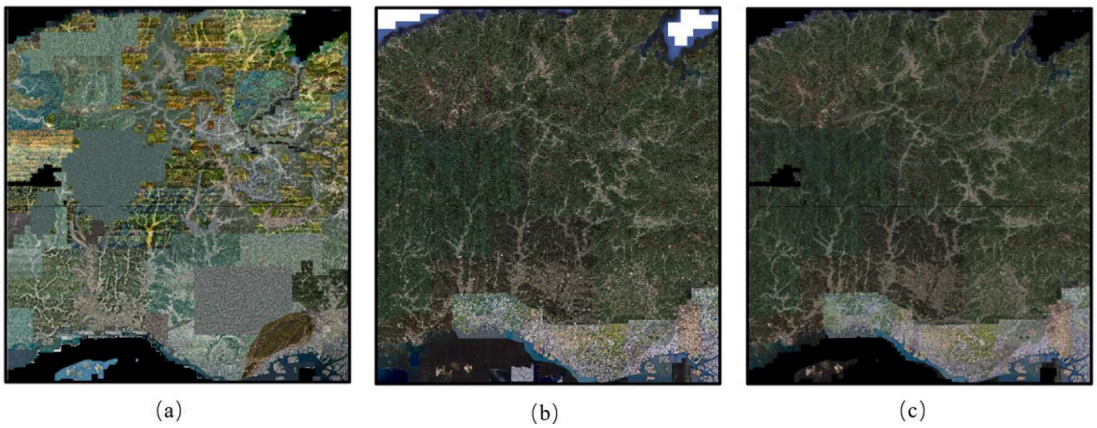

Fig. 6. Color normalization of the stitched image. The black and white pixels represent the missing data; (a) original GSI image; (b) reference GE image with 3 m resolution; (c) color-normalized image with qualitatively improved color transition.

classified as with buildings, and the number of images for unique scenes was maximized. The final dataset contained 4055 and 3974 images for the class with and without buildings, respectively, assigned to the training set, validation set, and test set in the ratio of 70:20:10.

4.2. Color normalization and image super-resolution

4.2.1. Color normalization

As described in Section 3.2, the GSI seamless photo dataset used as the test image data source comprised satellite images captured by various sensors simultaneously. Consequently, there were significant color differences in the test data of Hyogo Prefecture after mosaicking (Fig. 6a). The color difference disrupts the visual continuity of the images and impedes the generalization ability of the building extraction model. Moreover, the color differences caused by the various image sources of the test and training set can harm the building extraction. Therefore, performing color normalization for images is necessary. It helps shorten the difference in color distribution between the target and source domains, without damaging the image texture information or affecting the feature interpretation.

Color normalization approaches for remote sensing images include the parametric approach, GAN-based style transfer algorithm and reference-image-based approach. The parametric method recovers the color relationships between images by building a parametric regression model. However, it is not suitable for a large area because of potential color deviation when the texture and color of the object differ significantly (Cresson and Saint-Geours, 2015; Xie et al., 2018). GAN-based style transfer algorithm is an emerging approach for color normalization (Gatys et al., 2015; Gupta et al., 2019). Compared with the parametric algorithms, the CNN layer can be better for extracting deep features of images, but the computation time is longer and the details are easily distorted (Xiao et al., 2021; Xue et al., 2020). It is not conducive to the generalization ability of the building extraction model. In contrast, the reference-image-based approach enables simple and fast color-mapping relationships between different remote sensing images (Cui et al., 2021). Therefore, in the study, we selected the GE image as a reference map to correct the tone of GSI images.

Analyzed from a spectral perspective, remote sensing images can be divided into high- and low-frequency information. High-frequency information is caused by the sharp excess of grayscale, primarily including the texture and edges of buildings. The low-frequency information is related to the grayscale component of image variations, which reflects the trend of image color variations (Hao et al., 2017). As the low-frequency component of the image is smooth, the replacement of the low-frequency information will not interfere with the high-frequency information of the image. Therefore, we considered the low-frequency information of the reference image as a standard to correct the tones of the test image (Cui et al., 2021). Adequate color consistency is necessary for selecting reference images, and the similarity of the tone styles and training dataset is desirable. Additionally, the resolution of the reference map should not be too high, to ensure a relatively small amount of data and a high processing efficiency. Thus, we selected the 3 m resolution GE image as the reference (Fig. 6b).

Fig. 7 illustrates the framework of the color normalization algorithm with the following main procedure: The GSI image I_{src} was first down sampled to the same resolution as the reference image I_{ref}, and the high- and low-frequency information of the down-sampled GSI image $I_{srcdown}$ and reference images I_{ref} were separated through Gaussian filtering.

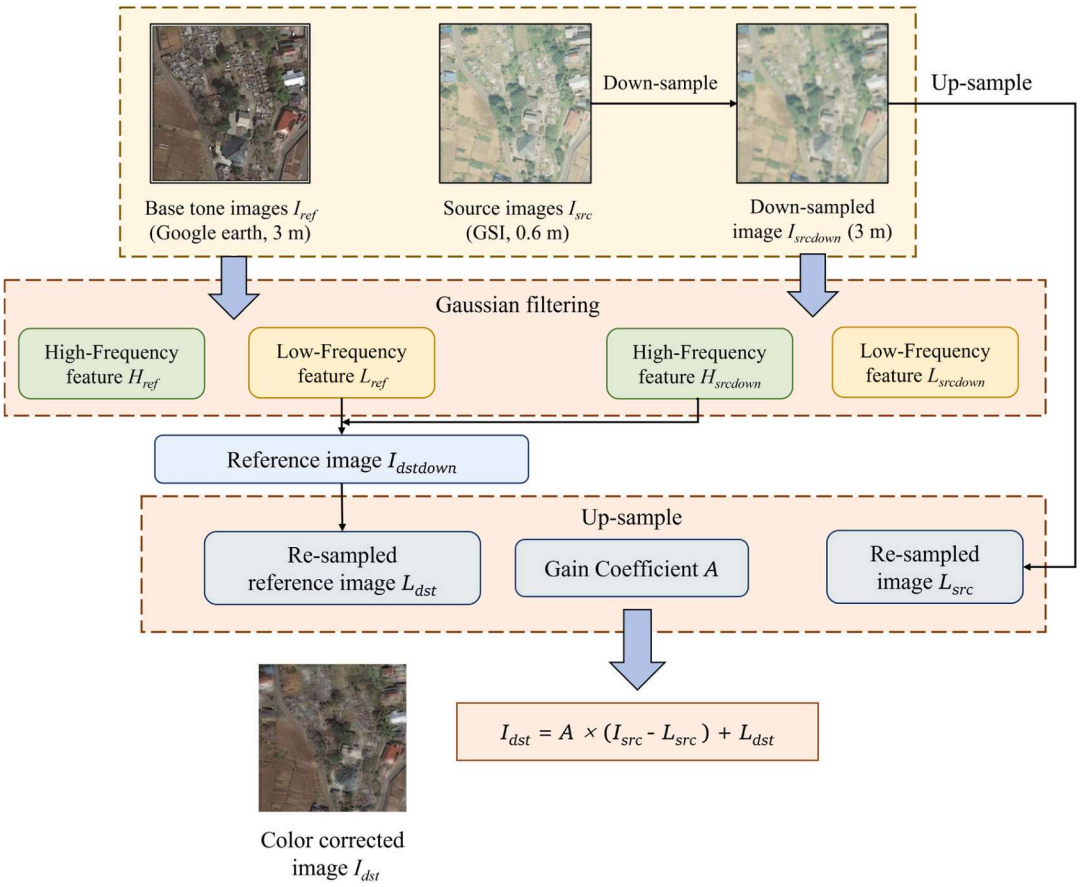

Fig. 7. Flow of the color normalization algorithm based on reference images. The source and reference images' high- and low-frequency information are separated by Gaussian filtering. Then the source image's high-frequency information and the reference image's low-frequency information are combined to generate the color-corrected image.

Subsequently, the low-frequency information of the down-sampled GSI image $L_{srcdown}$ was replaced with the low-frequency information of the reference image L_{ref} to reconstitute the toned image $I_{dstdown}$. Thus, $I_{dstdown}$ exhibits both the tones of the reference image and the texture information of the GSI test image. Finally, $I_{srcdown}$ and $I_{dstdown}$ were restored to their original resolution through bilinear interpolation. The up-sampled $I_{srcdown}$ and the smooth $I_{dstdown}$ fitting the original images well, both could be considered to represent the low-frequency information of the source and toned images, and denoted as L_{src} and L_{dst}. The color-corrected image I_{dst} (Fig. 6c) can be obtained through Equation (3), where A is the gain coefficient, calculated from the luna components of I_{ref} and $I_{srcdown}$. The overall color consistency of the color-corrected image was significantly improved, with no significant tonal differences in adjacent edges and with consistency with the reference image tones, which proves the algorithm's effectiveness.

$$I_{dst} = A \times (I_{src} - L_{src}) + L_{dst} \tag{3}$$

4.2.2. Super-resolution

Although models trained on high-resolution images exhibit high performance, the difference in resolution between the training and test sets can generally affect instance segmentation. First, the resolution defines the size of individual pixels. Consequently, the size of the same building is different in the remote sensing images of different resolutions. However, it is difficult for a model trained for a particular resolution to accurately represent the features of a test dataset containing a different resolution, which will result in the poor generalization ability of instance segmentation. (Hamaguchi and Hikosaka, 2018). In addition, small-sized buildings in low-resolution images are easily omitted because of the limited resolution (Kisantal et al., 2019). We adopted SR techniques to overcome the limitation of resolution differences in multi-source remote sensing images.

Supervised SR algorithms perform more satisfactorily than unsupervised ones in high-frequency regions (Chen et al., 2020). HR-LR image pairs should be input as training data to train a supervised SR model. Acquiring HR and LR images for specific areas is difficult; therefore, LR images are usually obtained by downsampling from HR images. However, owing to the complex degradation process, the down-sample-based data synthesis method cannot simulate the remote sensing images. Thus, we adopted the Real-ESRGAN network as the SR model (Wang et al., 2021b), which proposes a high-order degradation approach by introducing two times the number of degradation parameters as traditional degradation (blur, noise, resize, and image compression). Each degradation uses a different hyperparameter to simulate a real-world low-resolution image blurring. Using the higher-order degradation model described, we down-sampled the GE images with 0.3 m resolution and constructed 6000 sets of HR-LR synthetic image pairs (0.3 m-0.6 m). Each HR image had a size of 512*512 pixels and was assigned to the training set, validation set and test set in a ratio of 80 %:20 %:20 %.

Fig. 8 shows the architecture of the Real-ESRGAN network, which contains a generation network and a u-net discriminator. The generative network is the same as that of ESRGAN, with the addition of image resolution enhancement for 2 × and 1 × in function (Wang et al., 2018). Compared with the generator structure of the classical SRGAN network, ESRGAN replaces the original basic block with a residual in residual dense block (RRDB), which combines a multilayer residual network and dense connect. Moreover, the Batch Normalization (BN) layers are removed to improve the model generalization ability and reduce the model complexity. As the generation network of ESRGAN only supports 4 × SR, to adapt to 2 × and 1 × SR, the network first performs pixel-unshuffle to expand the number of image channels with a reduced image resolution, and the processed image is then input into the network for SR reconstruction (Shi et al., 2016). In addition, the U-Net discriminator can judge the authenticity of generated images from the pixel perspective, focusing on the details of the generated image while ensuring overall authenticity. Moreover, spectral normalization is added to alleviate the training instability problem due to complex data and networks.

4.3. Instance segmentation

Mask R-CNN is a widely used, powerful deep learning model (He et al., 2017). The network is derived from Faster R-CNN and a fully convolutional network (FCN), to which a new task branch is added to complete the pixel instance segmentation of the target object. Fig. 9 shows the network architecture of Mask R-CNN. The input images are first sent to ResNet for feature extraction. The backbone feature map is passed through the region proposal network (RPN) to extract the possible ROI (Ren et al., 2015). This ROI is mapped into fixed dimensional feature vectors by the ROIAlign layer. Two branches are for classification and regression of the target boundary box through the fully connected layer—the fully convolutional layer upsamples the other branch to obtain the segmented region image.

Although encouraging, the performance of deep residual networks (Resnet), a vital feature extraction backbone of Mask R-CNN, still has

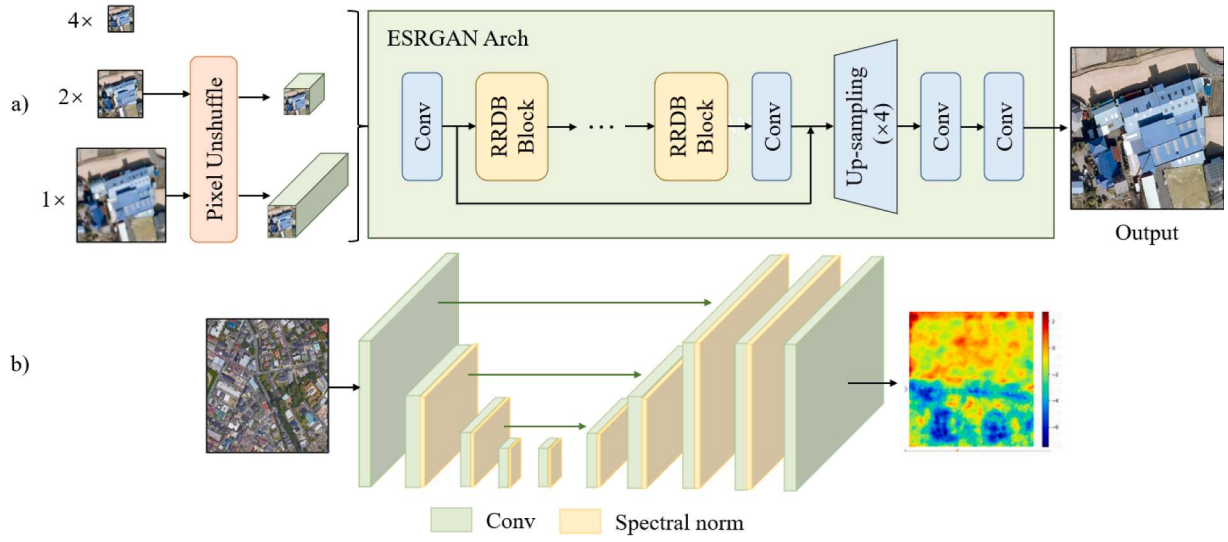

Fig. 8. Architecture of the Real-ESRGAN network. a) Generator; b) U-net discriminator with spectral normalization introduced to stabilize the training (Wang et al., 2021b). The RRDB block represents the Residual-in-Residual dense block, a module of ESRGAN (Wang et al., 2018).

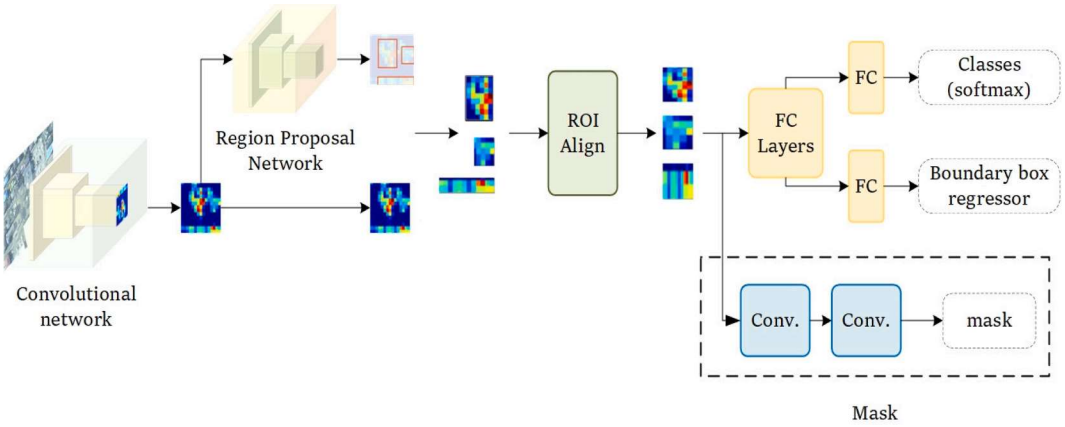

Fig. 9. Architecture of Mask R-CNN consists of a convolutional backbone network used to extract features, the Region Proposal Network (RPN) used to detect candidate regions proposals, and head networks to identify boundary boxes (classification + regression) and predict masks (He et al., 2017).

limitations. Deep CNN (DCNN) with small and fixed fields of view can only focus on a small range of neighbourhood features (Yan et al., 2021). This local feature representation limits the network from recognizing buildings from complex backgrounds in remote sensing images. Moreover, for intensive prediction tasks such as building extraction, it is crucial to represent features at multiple scales to distinguish between objects of different sizes (Wang et al., 2022). In summary, the primary idea to improve the performance of Mask R-CNN is to improve the multi-scale feature representation and the ability to fuse features. Based on this, the attention-based Transformer has certain advantages over DCNN and has achieved state-of-the-art performance in several vision tasks (Liu et al., 2021b; Wang et al., 2021a). Therefore, this study combines Mask R-CNN and a multi-path vision transformer (MPViT) as the instance segmentation model for building extraction (Lee et al., 2021). The effectiveness of the improved Mask R-CNN model used in this paper is based on multi-scale feature extraction and global and local feature fusion.

MPViT follows a unique approach to multi-scale patch embedding and multi-path structures compared to other Transformers (Fig. 10). A four-stage feature hierarchy to generate feature maps at various scales is constructed. The blocks of the proposed multi-scale patch embedding (MS-PatchEmbed) and multi-path Transformer (MP-Transformer) are stacked at each stage. By applying convolution operations with overlapping patches, the MS-PatchEmbed layer uses fine and coarse-grained visual tokens at the feature level. This enables the features of the exact resolution with a different sequence length of tokens by changing padding and stride. Then, the tokens embedding features of varied sizes are fed into the MP-transformer encoder separately. In this transformer block, global self-attention is performed by each transformer encoder with various patch sizes. The generated features are aggregated to allow for fine and coarse feature representations. In the feature aggregation step, a global-to-local feature interaction (GLI) process is introduced to connect the local convolutional features to the global features of the Transformer, leveraging the local connectivity convolution and the global connectivity context of the Transformer. Moreover, to alleviate the computational burden associated with the multi-path structure, the efficient factorized self-attention of CoaT is applied (Xu et al., 2021c):

$$\mathrm{FactorAtt}(Q,K,V)=\frac{Q}{\sqrt{C}}\left(\mathrm{softmax}(K)^{T}V\right), \tag{4}$$

where Q, K, and $V \in \mathbb{R}^{N\times C}$ denote linearly projected queries, keys, and values, respectively; N, and C are the number of tokens and embedding dimension, respectively.

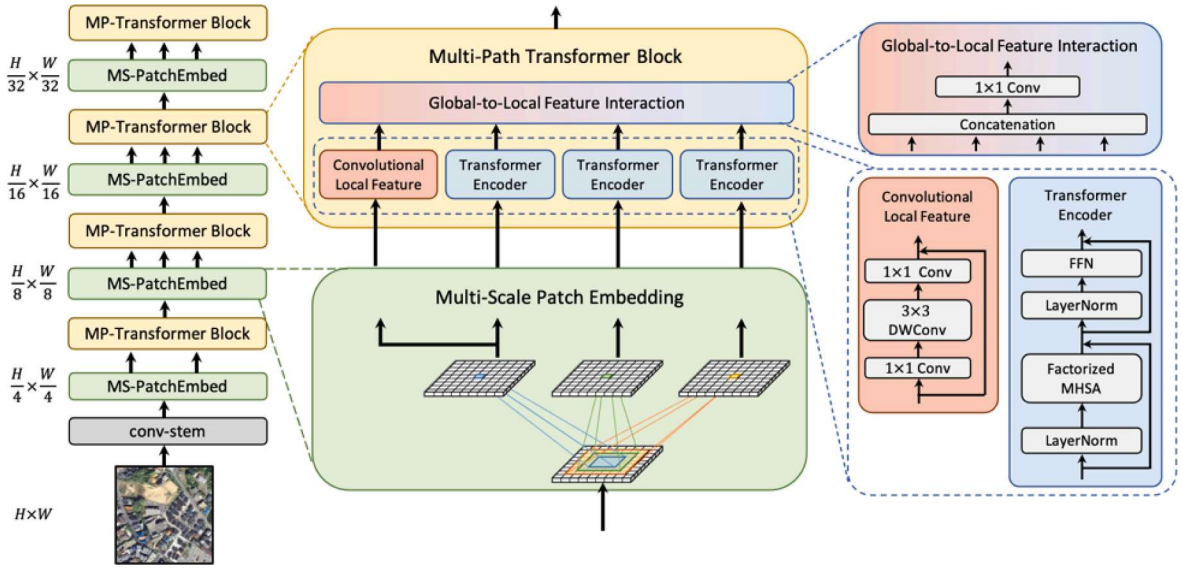

Fig. 10. Architecture of the MPViT. Multi-path transformer (MP-Transformer) block and multi-scale patch embedding (MS-PatchEmbed) block are stacked in each stage (Lee et al., 2021). Besides, factorized self-attention is used in Transformer encoders.

4.4. Scene mosaicking

A single remote sensing image encloses a large area. However, the size usually exceeds the image input limit for deep learning models owing to computational limitations. A widely used remote sensing image segmentation strategy is subdividing an image into equal-sized patches based on the sliding window method (Yi et al., 2019). The sliding window method runs through the image at a specific size in horizontal and vertical directions in certain steps. An overlap is created between two consecutive patches when the step size is smaller than the window size. The error of instance segmentation occurs at the edges of the patch. For example, as shown in Fig. 11a, if there is no overlap between adjacent patches, the building at the edge may be split into two parts, eventually producing two separate footprints. This edge effect can be eliminated or minimized by the overlap of patches, as shown in Fig. 11b. However, the size of the sliding window and the image overlap rate ought to completely account for the scale of the target feature without excessively increasing the computational load (de Bem et al., 2020). Considering the actual situation of the study, we set the dimension of each patch to 1024 and the overlap rate to 20 %.

Furthermore, the predictions of each patch ought to be mosaicked to attain a complete building footprint map of the target area. The multiple masks generated for the same building in the overlapping regions should also be merged. A modified Mask non-maximum suppression (Mask-NMS) algorithm was applied according to the mask area to exclude excessive incomplete masks. The method first sorts the masks in the overlap according to their areas, calculates the intersection-over-union (IoU) between them, and eliminates the masks with smaller areas if the IoU is more significant than a threshold. The IoU is calculated from the ratio of the intersection area $A_{overlap}$ and the concatenation area A_{Union} of two masks (Equation (5)). This study considered an IoU of 0.2 to exclude excessive masks. Compared to the traditional NMS algorithm that suppresses the mask based on the score, an area-based ranking is more appropriate because there is no guarantee that the mask with the highest score to encompass the entire building (Neubeck and Van Gool, 2006).

$$IoU = \frac{A_{overlap}}{A_{Union}} \tag{5}$$

Fig. 12 visually demonstrates the specific process to illustrate the proposed scene mosaicking method. Fig. 12a illustrates the initial prediction results of the two overlapping patches, and the yellow masks were the same footprints generated in the overlapping area. The red masks in Fig. 12b are the footprints in the center of the overlap area. As these masks were completely extracted in both patches I and II, the patch with the smaller area would be eliminated after mosaicking. Fig. 12c demonstrates the footprint at the edge of the overlap region, where the green and blue masks belong to patches I and II, respectively, which can only be predicted partially. As these masks were extracted completely in the adjacent patches, they would be suppressed directly through NMS, based on footprint area. Fig. 12d shows the footprint map after scene mosaicking, where all overlapping masks have been removed.

5. Experiment

5.1. Implement details

Three experiments were conducted to verify the validity of the proposed framework. In the first experiment, we compared the results of different backbones for scene classification. In the second and third experiments, we compared the effects of different super-resolution (SR) and instance segmentation methods on the building extraction performance. The experiments were implemented on a large-scale platform called mdx in 4 Nvidia A100 GPUs (40 GB) (Suzumura et al., 2022). The detailed setting of each experiment is described in Table 2. The hyperparameters of optimizers and data augmentation are kept the same as the optimal parameter settings in the original paper of each model (He et al., 2017; Lee et al., 2021; Liu et al., 2022b; Wang et al., 2021b); the training strategy is set according to the hardware environment and characteristics of the training set through the empirical method of hyperparameter optimization (Feurer and Hutter, 2019).

In the training phase of the scene classification model (ConvNeXt-base), the network was trained by an AdamW optimizer with a momentum $\beta_1, \beta_2 = 0.9, 0.999$ and a weight decay of 0.05 (Loshchilov and Hutter, 2017) based on MMclassification (Contributors, 2020). The learning rate was initialized as 0.0005, and we utilized a cosine decay schedule with 220 warmup iterations. Furthermore, to prevent overfitting and to increase model performance, RandAugment (N = 9, M = 0.5), mixup (prob = 0.8) and cutmix (prob = 1) were applied for data augmentation (Cubuk et al., 2020; Yun et al., 2019; Zhang et al., 2017). The network converged at 1100 iterations, and the batch size was set to 128.

For SR, the experiment (Real-ESRGAN) was based on the BasicSR (Wang, Yu, Chan, et al. 2018) with a batch size of 16. The network was trained for 5000 iterations with a learning rate of 0.0001 for both the generator and discriminators. The Adam optimizer (betas = [0.9, 0.99], decay = 0) and exponential moving average (EMA) was used for better performance and stable training. The horizontal flip and random rotation were performed for data augmentation.

The training of the instance segmentation model (Mask R-CNN with an MPViT backbone) was implemented through Detectron2 (Wu et al. 2020). As the dataset was not sufficiently large to train a Mask R-CNN end-to-end model from the start, a pre-trained model on the ImageNet dataset was used for transfer learning. Transfer learning reduces the training data and improves the model's overall accuracy and generalization ability. The dataset with GE images described in Section 3.2 was used for training. Following the standard settings and the training recipe of Swin-Transformer, the maximum number of iterations was set to 30,000 using a multi-scale training strategy, which resizes the input such that the shorter side is between 900 and 1100, whereas the longer side is $\leq$ 1333. Additionally, the AdamW optimizer was used for network optimization with an initial learning rate of 0.0001 and weight decay of 0.05 (Loshchilov and Hutter 2017). We selected a batch size of 16 and an initial learning rate of 0.0001, which decays 10 times at the iterations of 24,000 and 27,000. All the parameters were initialized according to the orthogonal distribution. Furthermore, we employed random cropping, flipping, and rotation to avoid overfitting the augmentation model.

5.2. Scene classification

Table 3 shows the performance evaluation results of three SOTA backbones on the test set, which are ResNeSt-101 (Zhang et al., 2022), Swin-Transformer-base (Liu et al., 2021b) and ConvNeXt-base (Liu et al., 2022b). We used precision, recall, F1, and accuracy as the evaluation metrics. The definition of each accuracy metric is shown in Equation (6), where TP, FP, TN, and FN refer to true positive, false positive, true negative, and false negative, respectively. The performance of the ConvNeXt-base model outperforms other models, achieving exceptionally high accuracy and an F1 value of 0.993. Although the Transformer architecture has recently surpassed the CNN on image-classification tasks in recent years, the two CNN-based models perform better in our study. This is because the Transformer architecture model lacks convolution's inductive bias and needs to learn features from a large dataset (Liu et al., 2021a). Therefore, in this study, the Swin-Transformer did not perform well for a small dataset in our study. Besides, we attempted to provide as much accuracy as possible through the model to identify scenes with buildings. As shown in the confusion matrix of the ConvNeXt-base model in Fig. 13a, the accuracy for the class with building reached 100 %, as expected. Additionally, there were apparent artificial traces in the scene for those misclassified images without building (Fig. 13b), such as roads and cultivated land. Careful examination revealed small-sized buildings in some of these images.

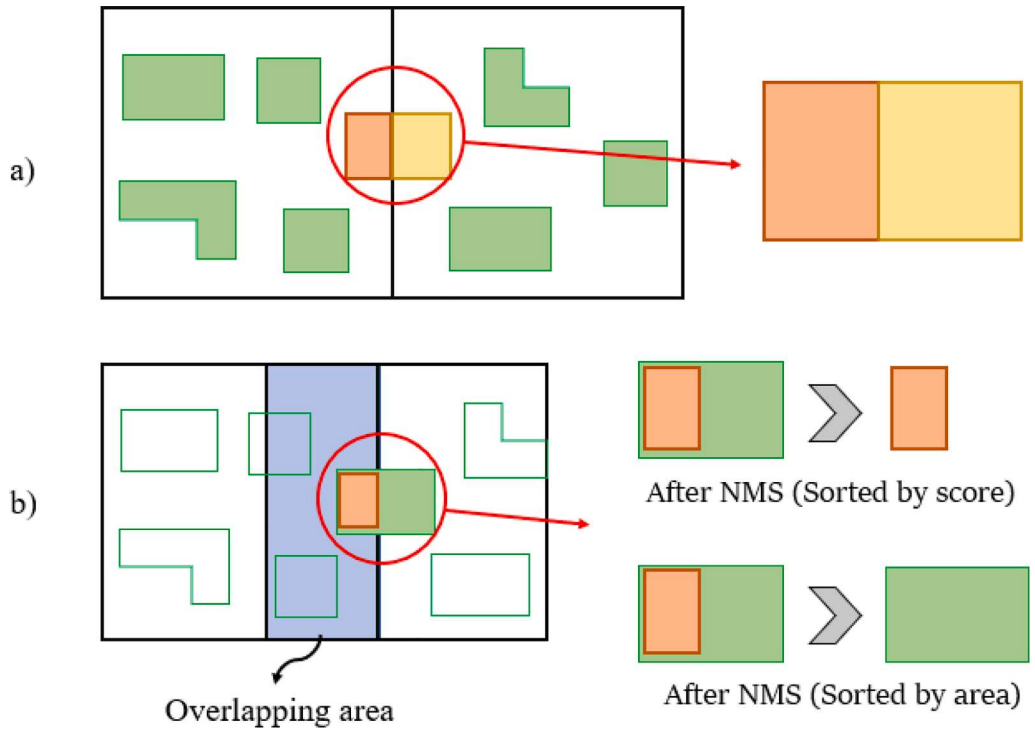

Fig. 11. Effect of different patch cropping strategies on buildings in the edges. (a) No overlap between patches leading to partial segmentation; b) overlap between patches. Complete footprint results can be obtained through area-based non-maximum suppression (NMS), whereas partial footprint result is retained through score-based NMS.

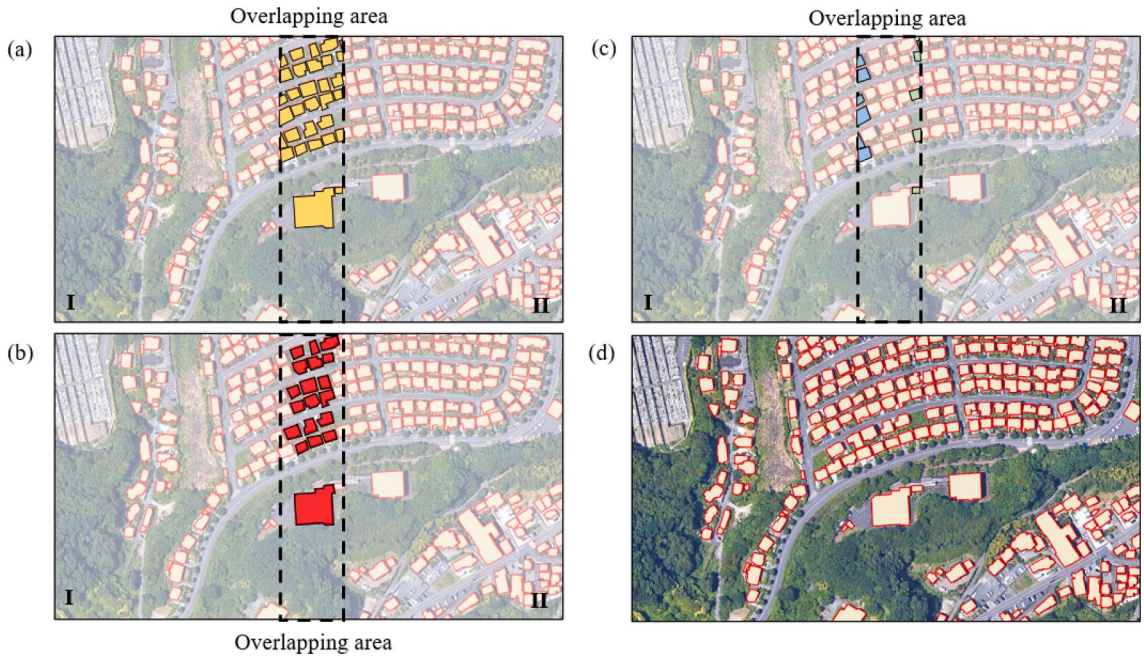

Fig. 12. Results of scene mosaicking. (a) The original prediction result, where yellow masks are the duplicates generated in the overlapping area; (b) complete masks (red) in the center of the overlapping area; (c) partial masks (blue and green) eliminated by NMS, based on the area; (d) the final footprint map after scene mosaicking. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Therefore, the model could effectively avoid misclassification because of the mislabeling of the training dataset.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{2 \times Precison \bullet Recall}{Precision + Recall}$$

Table 2
Setting details for experiments of scene classification, super-resolution, and instance segmentation.

Experiment	Scene classification	Super-resolution	Instance segmentation
Platform	MMclassification	BasicSR	Detectron2
Optimizer	AdamW	Adam	AdamW
Optimizer momentum	$\beta_1,\beta_2 = 0.9, 0.999$	$\beta_1,\beta_2 = 0.9, 0.99$	$\beta_1,\beta_2 = 0.9, 0.999$
Weight decay	0.05	0	0.05
Base learning rate	5e-3	1e-4	1e-4
Batch size	128	16	16
Training iterations	1100	5000	30,000
Learning rate schedule	Cosine decay	Step decay [4000,]	Step decay [24,000, 27,000]
Warmup iterations	220	–	1000
Warmup schedule	Linear	–	Linear
Data augmentation	RandAugment (N = 9, M = 0.5) Mixup (prob = 0.8) Cutmix (prob = 1)	Horizontal clip Random rotation	Random cropping Random flipping Random rotation Resize [400, 600]

Table 3
Result of the scene classification performance of different SOTA backbones.

Model	Precision	Recall	F1	Accuracy
ResNeSt-101	0.987	0.988	0.988	0.988
Swin-T-b	0.949	0.948	0.948	0.948
ConvNeXt-b	**0.992**	**0.993**	**0.993**	**0.993**

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

5.3. Super-resolution

To validate the feasibility of the super-resolution, in this section, we compare the effect of different SR approaches for building extraction. To this end, we trained a Mask R-CNN model with the MPViT backbone for instance segmentation, and selected three state-of-the-art SR approaches. They included the Real-ESRGAN (Wang et al., 2021b), BSRGAN (Zhang et al., 2021b) and SwinIR (Liang et al., 2021). All of them were fine-tuned on the dataset described in section 4.2.2, where the BSRGAN and SwinIR used the image degradation method of BSRGAN to generate HR-LR image pairs (0.6 m to 0.3 m). In addition, bicubic interpolation was used as the baseline for the experiment.

We first selected two images for the qualitative evaluation. Fig. 14 shows the reconstructed HR images with two up-sampling scale factors and the corresponding extracted building footprints from the different SR methods. The first columns are the original LR images and the ground-truth labels. In addition, specific areas within each image were magnified in the red window to reflect the reconstruction details of the different SR methods. The results of bicubic interpolation either failed to remove degradation or added unnatural textures. The deep learning SR model could reconstruct some texture details; however, the images generated by BSRGAN showed contour blurring. In comparison, the Real-ERSGAN and SwinIR models improved visual clarity. This could recover more realistic and natural textures. Moreover, as shown in Fig. 14a, the deep learning SR models perform better at extracting building boundaries than the bicubic interpolation because bicubic interpolation did not yield additional details and possessed certain artefacts. The prediction results in Fig. 14b are almost identical, but the result of the Real-ESRGAN model has a slight advantage in terms of contour detail.

In addition to the subjective qualitative judgments, we applied peak signal to noise ratio (PSNR) and mean structural similarity index (MSSIM) to evaluate the performance of the SR model. PSNR is the ratio of the maximum possible pixel value ($MaxI$) and the pixel-level mean square error (MSE) between HR imagery (f_{ij}) and SR image (f'_{ij}), which are objective indicators to evaluate image quality (Equation (7)). The larger the PSNR value, the better the image quality. However, PSNR does not consider the human eye's visual recognition and perception characteristics. Furthermore, the evaluation results are often different from the subjective human perception. Therefore, MSSIM was used as a supplement. It divides the image into N blocks through the sliding window method and the weighted mean values (μ_x, μ_u), variance values (σ_x, σ_y), and covariance (σ_{xy}) for each window, which measure the similarity of images in terms of brightness, contrast, and structure (Equation (8)). Besides, precision, recall, and F1 values were used to evaluate the object-wise accuracy of the prediction result, which was defined similarly to Equation (6), with a different TF judgment criterion. In building instance segmentation, TP and FP indicate the correct and incorrect building detections, respectively. As illustrated in Fig. 15, we used an IoU threshold of 0.5 to classify the predictions as correct (TP). The test data described in sections 4.2.2 and 3.2 were used for SR and instance segmentation models, respectively.

Table 4 shows a quantitative assessment of how the super-resolution-based building extraction was performed using different SR methods. The best result is shown in bold. We can find that all deep learning SR methods achieved better performance than the baseline model,

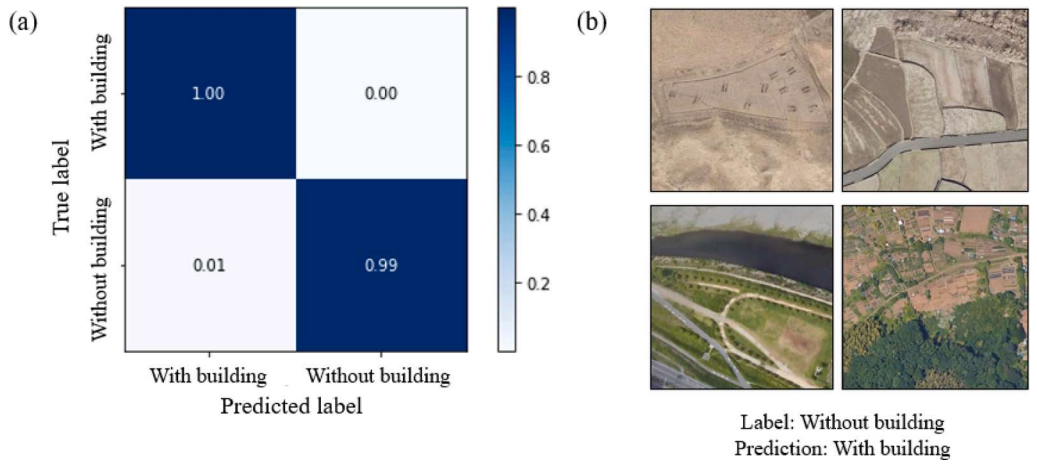

Fig. 13. Scene classification result of ConvNeXt-base model. (a) The normalized confusion matrix of scene classification on the test data; (b) examples of images without buildings that were misclassified as images with buildings.

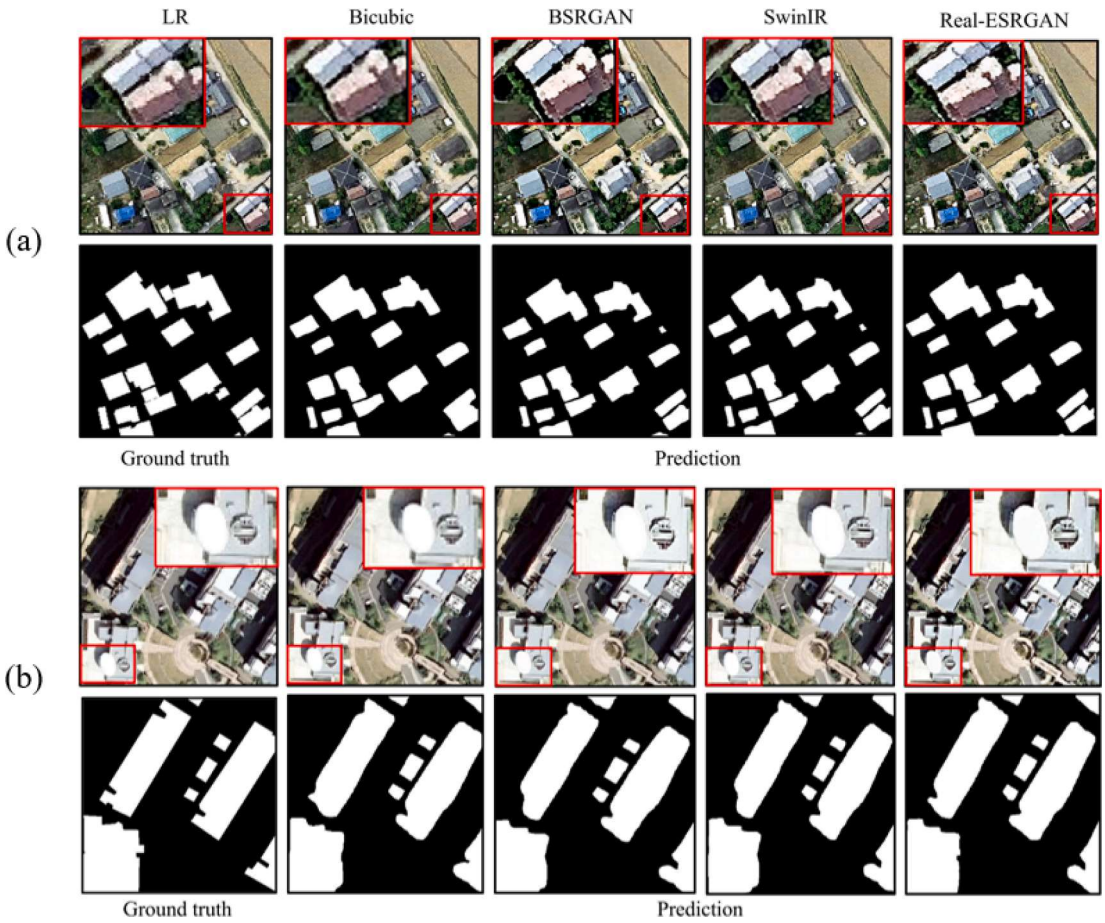

Fig. 14. Qualitative comparison of different super-resolution on GSI images with an upsampling scale factor of two. The reconstructed HR images and the extracted building footprint are the first and second rows of (a) and (b). The first column is the original LR image and ground truth label.

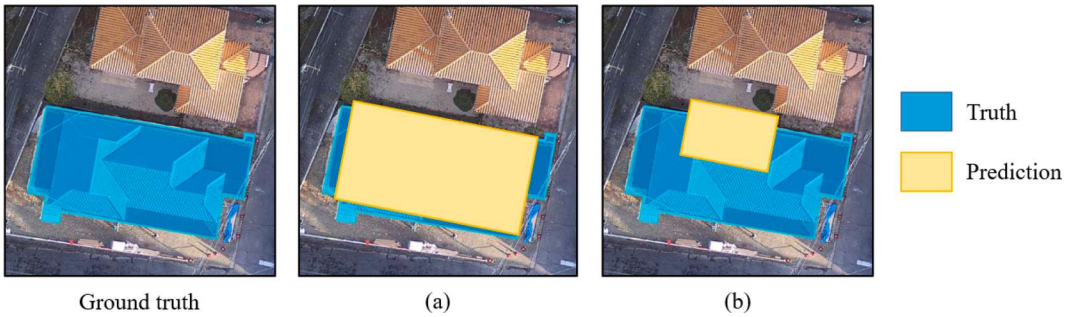

Fig. 15. Schematic representation of the object-wise metric. When the intersection over union (IoU) of the manually labelled ground truth (blue) and the predicted building (yellow) is ≥ 0.5, the prediction is considered TP (a), whereas when it is less than 0.5, the classification is considered FP (b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quantitative performance of different SR models with the upsampling scale factor of two.

Model	Precision	Recall	F1	PSNR	MSSIM
Bicubic	**0.84**	0.60	0.70	26.79	0.860
BSRGAN	**0.84**	0.63	0.72	29.05	0.907
SwinIR	0.83	0.62	0.71	**29.55**	**0.917**
Real-ESRGAN	**0.84**	**0.64**	**0.73**	29.06	0.908

especially for the recall of building extraction results. It suggests that compared to simple interpolation, SR helps to enhance the distinction between background and buildings in reconstructed images. It can fully use internal information in LR images. This result follows the same trend as the qualitative evaluation. Moreover, the SwinIR outperforms other methods in metrics of SR, achieving a PSNR of 29.55 and SSIM of 0.917. However, for building extraction performance, the Real-ESRGAN model performed the best, outperforming the precision, recall and F1 of SwinIR by 0.01, 0.02 and 0.02, respectively. This phenomenon suggests that the performance of the SR model does not dominate the performance of SR-based instance segmentation. In other words, simply appending high performing SR model may not always provide excellent performance.

For example, unlike other models, the images generated by SwinIR, although closer to the original image's visual style, have relatively low contrast. This may not be beneficial for building extraction. The interpretation of this result exceeds our expectations, and we plan to investigate the reasons in future work. However, while it is impossible to assert that the CNN-based SR models perform better based on this result, Transformer-based SR models are still in the exploratory stage. Therefore, we recommend using the latest and most potent CNN-based models for SR-based building extraction.

$$\mathrm{PSNR} = 10log_{10}\left(\frac{MaxI}{\sqrt{MSE}}\right) \tag{7}$$

$$SE = \frac{1}{C \bullet M \bullet N}\sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(f_{ij} - f_{ij}^{'}\right)^2$$

$$\mathrm{MSSIM}(X, Y) = \frac{1}{N}\sum_{k=1}^{N}\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \times \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \times \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{8}$$

5.4. Instance segmentation

In this section, the Mask R-CNN model with an MPViT-base backbone (Mask R-CNN M) was used for building extraction in Hyogo Prefecture to test the framework's effectiveness proposed in this paper. All test images were upsampled 2 × by using the fine-tuned Real-ESRGAN model to enhance the image texture details. Before segmentation, the test images were fed into the scene classification network. The number of test images in Hyogo Prefecture was reduced from 81,438 to 30,000 after filtering, decreasing the workload by>60 %. Finally, the entire inference required a total of 30 h of operating time. Fig. 16a shows the large-scale distribution map of building footprints in Hyogo Prefecture, with city-level and more zoomed-in results in Fig. 16b and c, respectively. The prediction results are the raw output of the model without post-processing. On a large scale, the distribution of the building footprint matches well with the built-up areas on the satellite image. The roads and rivers between buildings are distinguished. Additionally, Fig. 16d demonstrates the prediction results of some typical buildings and

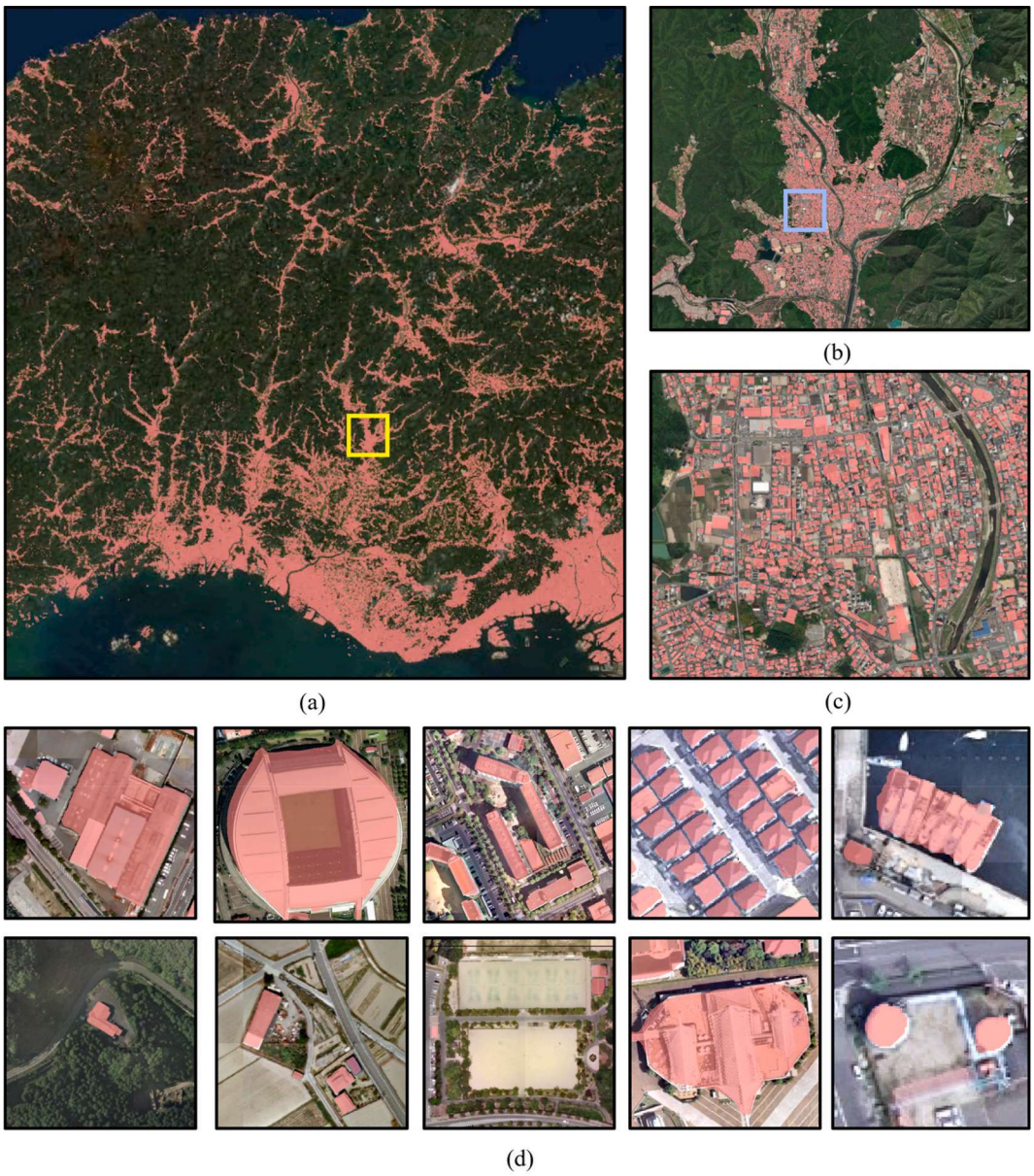

Fig. 16. Predicted distribution map of the building footprint. (a) Hyogo Prefecture. (b) One of the cities in Hyogo Prefecture is the yellow box in (a). (c) Red masks show the extracted building footprint over the suburban area, the blue box in (b). (d) Examples of prediction results of some typical buildings and ground objects. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

objects, including intricately shaped buildings and detached chalets in open areas. It can be seen that the model can properly extract the building footprints with different shapes, sizes and densities, but ships and LPG tanks are also mix-extracted due to a lack of training data. It should be noted that no training data from the test area was used in the model training. Despite that, the trained model still provides satisfactory extraction results.

To quantitatively evaluate the performance of the applied model, we compared the results of different instance segmentation methods on the multi-scene test dataset introduced in section 3.2.2. We choose the Frame Filed Learning approach with a Res101-Unet backbone (FFL mask) (Girard et al., 2021) and Deep snake (Peng et al., 2020) to represent the semantic-segmentation and contour-based approaches. A Mask R-CNN model with a ResNeXt-101 backbone (Mask R-CNN R), the best-performing benchmark model in the Detectron2 model zoo, was also applied as the baseline (Xie et al., 2017). For a fair comparison, all the models were transfer-learned on the same training set with the same learning rate. Figs. 17-20 show the predicted results of four models in non-built-up, rural, suburban, and urban areas. Specific regions from each area were selected separately for magnification to clearly and visually interpret the experimental results. In each figure are the original image, the ground truth, and the predictions of each model (left to right). Additionally, the quantitative evaluation results of the four models in three areas are shown in Table 5; the best result is shown in bold.

Fig. 17 shows a non-built-up area located in the mountains. The area is sparsely built and has some man-made and natural objects that can be confused with buildings, such as piles of building materials and trees. As shown in Fig. 17b, the FFL mask and Deepsnake model mistakenly extract more non-building objects than the Mask R-CNN model, and the MPViT backbone showed a better performance than the ResNeXt backbone. Besides, Fig. 17c illustrates two small cabins in an open field, which occupies only a dozen pixels but is still successfully extracted by all models. In summary, the Mask R-CNN M model can correctly identify the locations of buildings in a natural environment.

Fig. 18a shows the prediction result of the rural area, which consists mainly of small residential buildings with low density. The non-building areas include a large amount of agricultural land and some industrial facilities. Generally, the accuracy of building extraction increased for decreasing building density in an area with similar architectural styles. This is contrary to our experimental results, in which the accuracy in rural areas was lower than that in suburban areas. This can be attributed to the complex roof structure in rural areas because of additional construction. Also, other features that are easily confused with buildings additionally increase the difficulty of detection. However, the Mask R-CNN M model effectively avoids this misidentification, as shown in Fig. 18b, where ponds are misclassified as buildings by other models (yellow circles).

The suburban area shown in Fig. 19a has predominantly residential buildings of small and moderate sizes. Their alignment is consistent on a local scale. Due to the considerably pronounced outline of each building, facilitating easy extraction, the overall results are better. However, in Fig. 19b, the Mask R-CNN R model misses some small buildings; the Deep snake model identifies adjacent buildings as continuous built-up areas, and the FFL mask model results in some small speckled fragments—misidentifying the open space on the left as buildings. In addition to the single houses in the suburban area, there are also mid-rise flats located further apart. As shown in Fig. 19c, the FFL mask and Mask R-CNN R model tend to over-extract the surrounding structures of the building. In comparison, the Mask R-CNN M model has the best performance across different kinds of residential buildings.

As shown in Fig. 20a, urban areas are characterized by complex building types and high densities. The building size and spacing are small for densely built-up residential areas, making distinguishing a bigger challenge. This effect can be seen in the enlarged results, especially for buildings with similar color and shape, resulting in multiple buildings being identified as one (Fig. 20b). In addition, complex ground conditions and rapidly changing building dimensions in localized areas further affect urban recall. Fig. 20c displays some representative high-rise buildings and associated details. The shadows and side parts of a high-rise building can seriously affect the accuracy of building extraction based on the angle and timing of the shot. Although all models avoid mis-extracting shadows and find the location of buildings, the precise outline of a high-rise cannot be accurately provided due to the side or surrounding structures, as shown in the red circles of Fig. 20c.

In addition to qualitative analysis, the Mask R-CNN M model can achieve a better performance than the others in precision, recall, and F1, in approximately all areas. Especially in non-built-up areas, the accuracy

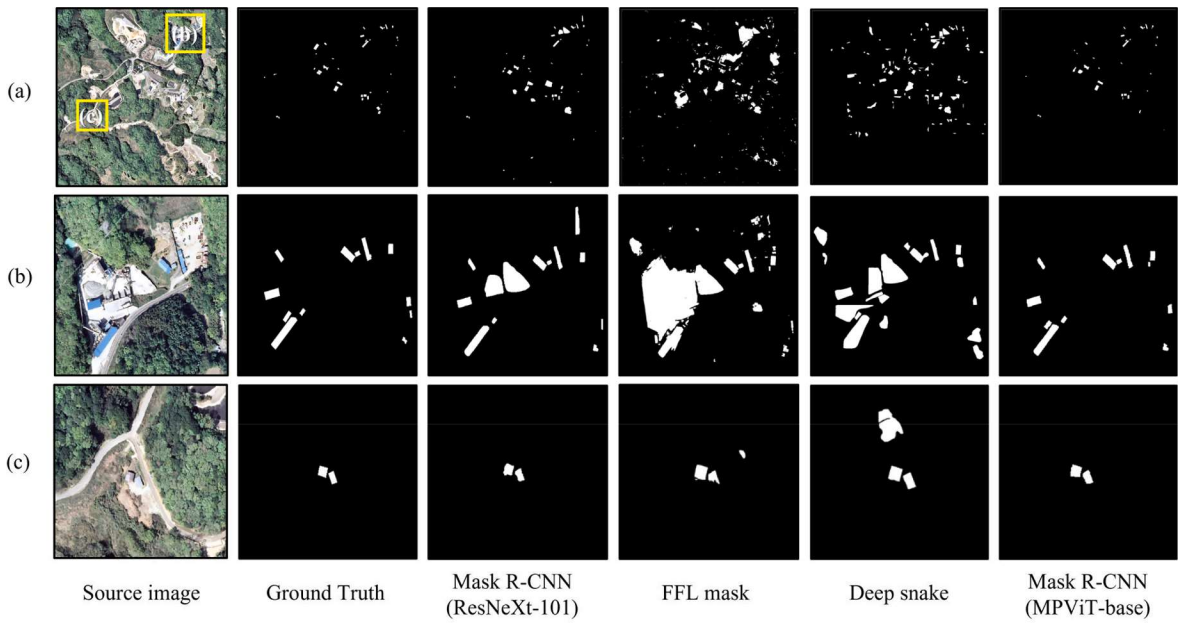

Fig. 17. Prediction results of different models and corresponding source image and ground truth in the non-built-up area. (b) and (c) show the magnified results of the yellow boxes in (a), representing the construction site and cabins in the wild. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

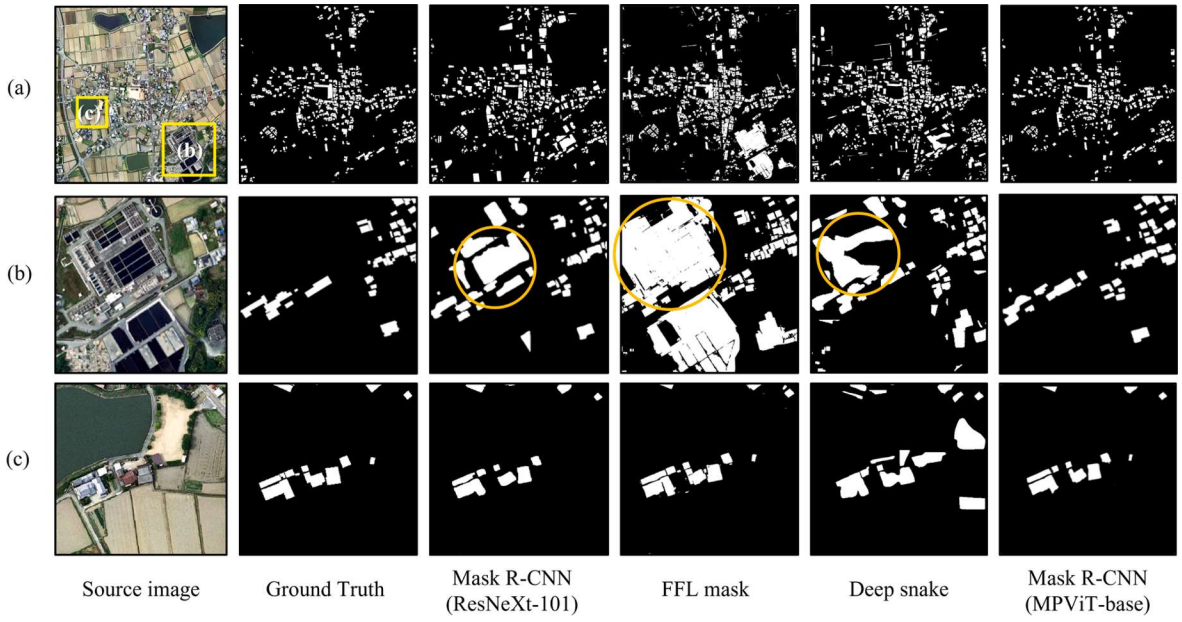

Fig. 18. Prediction results of different models and corresponding source image and ground truth in the rural area. (b) and (c) show the magnified results of the yellow boxes in (a), representing the complicated industrial facilities and compact buildings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

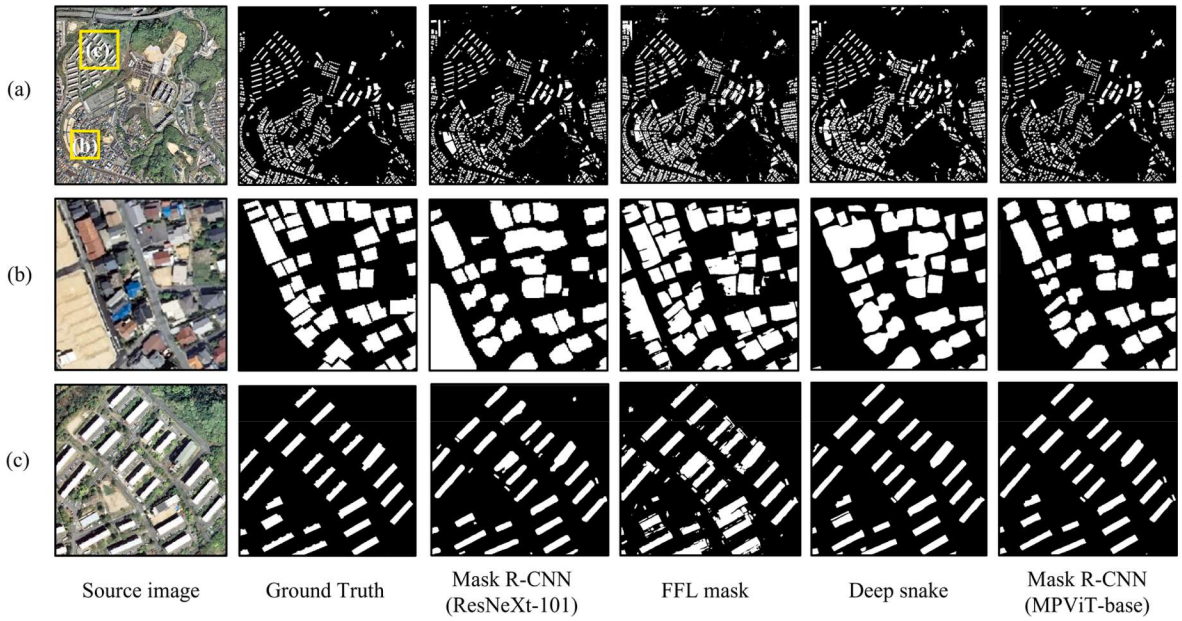

Fig. 19. Prediction results of different models and corresponding source image and ground truth in the suburban area. (b) and (c) show the magnified results of the yellow boxes in (a), representing the small- and moderate-sized residential buildings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the FFM mask and Deepsnake model is lower due to the scarcity of buildings and a large number of misidentified. In contrast, the Mask R-CNN M model maintains the same level of performance as in other areas. In urban areas, the FFL mask model has the highest recall. The frame, filed as additional supplementary information, facilitates the division of adjacent buildings, especially for areas with a high density. However, the semantic-segmentation-based method is trained at the pixel level. That leads to a tendency to over-extract objects similar to the building in color, resulting in a lower precision. Instead, end-to-end instance segmentation first detects the region of interest and then segments the building from the background, helping avoid speckle-like errors in semantic segmentation. On the other hand, the Deep snake model showed the worst performance. Apart from a high false positive rate, difficulties existed in segmenting adjacent buildings. This is because Deep snake utilizes only limited information from the boundary vertices. The inability to fully use features within the building dramatically increases the difficulty of extraction, resulting in the output being limited to simple polygons. Finally, the Mask R-CNN M model showed improved precision, recall and F1 values of 0.05, 0.04 and 0.05 over the baseline. It showed that the multi-scale feature fusion and self-attention mechanisms of the applied MPViT backbone could facilitate better model fitting and improve performance. However, assessing the feasibility of

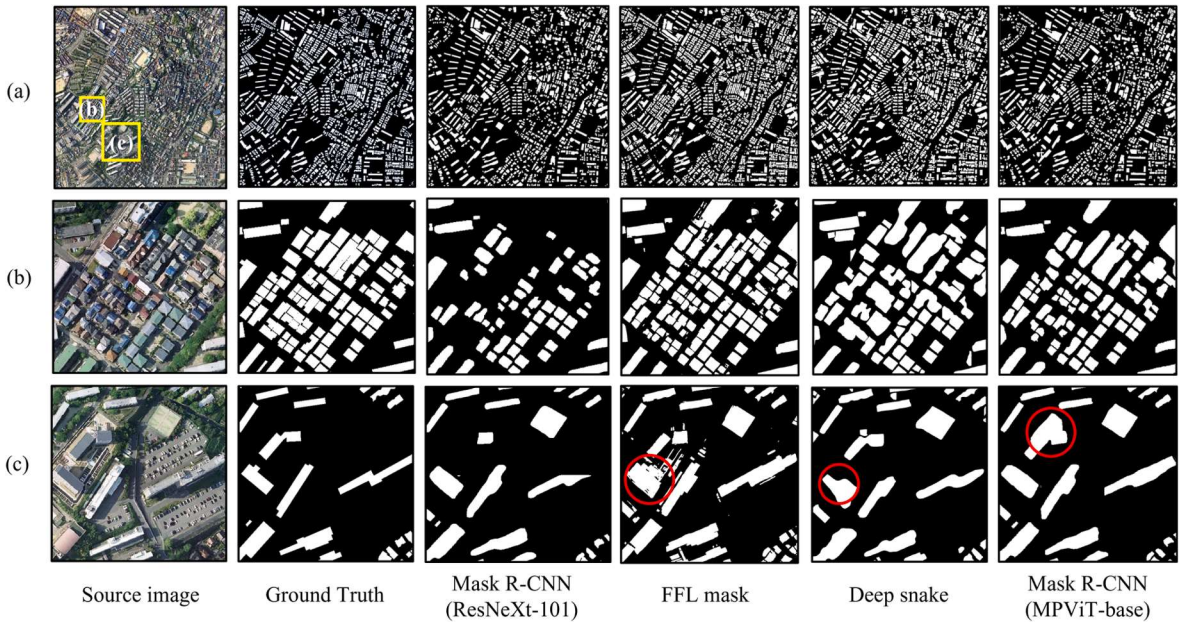

Fig. 20. Prediction results of different models and corresponding source image and ground truth in the urban area. (b) and (c) show the magnified results of the yellow boxes in (a), representing dense-built residential buildings and high-rise office buildings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 5
Object-wise metric of different models of different areas in Hyogo Prefecture.

Area	Model	Precision	Recall	F1
Non-built-up	Mask R-CNN (ResNeXt-101)	0.51	0.61	0.56
	FFL mask	0.21	0.46	0.29
	Deep snake	0.18	0.34	0.24
	Mask R-CNN (MPViT-base)	**0.77**	**0.66**	**0.71**
Rural	Mask R-CNN (ResNeXt-101)	0.77	0.62	0.69
	FFL mask	0.68	0.61	0.64
	Deep snake	0.58	0.56	0.57
	Mask R-CNN (MPViT-base)	**0.79**	**0.63**	**0.70**
Suburban	Mask R-CNN (ResNeXt-101)	0.82	0.69	0.75
	FFL mask	0.81	0.70	0.75
	Deep snake	0.66	0.62	0.64
	Mask R-CNN (MPViT-base)	**0.90**	**0.74**	**0.81**
Urban	Mask R-CNN (ResNeXt-101)	0.80	0.51	0.63
	FFL mask	0.77	**0.60**	**0.67**
	Deep snake	0.65	0.48	0.55
	Mask R-CNN (MPViT-base)	**0.84**	0.57	**0.68**
Average	Mask R-CNN (ResNeXt-101)	0.79	0.60	0.68
	FFL mask	0.75	**0.65**	0.70
	Deep snake	0.62	0.54	0.58
	Mask R-CNN (MPViT-base)	**0.84**	0.64	**0.73**

SR-based instance segmentation methods for open-source satellite image datasets is challenging, owing to the disparate resolutions and color in the test images. In summary, the Mask R-CNN M model achieves a maximum average precision and F1 value of 0.84 and 0.73, and a recall of 0.64, very close to the FFL mask (0.65), which is a specialized approach for building extraction. It shows exciting potential in building extraction from open-sourced images.

6. Discussion

6.1. Mask R-CNN backbone

The experimental results in Section 5.4 conclude that using the pre-trained MPViT model to replace the ResNeXt backbone of Mask R-CNN can provide better performance in building extraction. This shows the potential of the Transformer applied to remote sensing images. Moreover, it is more accurate and robust than other models in extracting buildings with variable density, texture, and type across different areas. Therefore, it can be concluded that using MPViT in the structure of extraction models can provide the same or more satisfactory results as SOTA methods.

However, the Mask R-CNN network is limited by its inability to classify each scale roof feature because of its CNN-based feature extraction network. First, the path between the high bottom features in its feature extraction network is excessively long, which can easily cause the loss of feature information transfer and the ineffective utilization of the bottom features. Second, the extracted feature map only carries its higher-level information features, which do not completely utilize the feature information at each scale, resulting in lower detection accuracy. This results in the ineffective utilization of the extracted full-scale features of the building, even with a deep extraction structure (ResNeXt-101), which reduces the model's accuracy (Wang et al., 2022).

The enhanced Mask R-CNN model employed in this study was effective because it extracts features at multi-scales and combines global and local features. First, a multi-scale feature extraction structure benefits the perceptual field by gathering complete target information and ensuring the integrity of the medium and big building footprint. Second, the increased feature fusion may enhance and decrease noise in the obtained feature information. The proposed network structure can better use multi-scale features and capture spatial interactions between targets, enabling the symbiotic detection of multi-scale structures. The quantitative and qualitative analysis of the experimental results indicates that these considerations contribute to the enhanced potential of the model for feature extraction. Thirdly, apart from determining the pixel class from the global features, the extracted multi-scale features evaluate the feature information in the neighbourhood on each scale. This enhances the prediction accuracy of nearby pixels. The result is

extracted footprints with straighter edge lines and more pronounced angles, where the boundary errors are no longer negligible at lower resolutions. In addition, the open-source dataset used in this study was multi-sourced and exhibited a relatively lower resolution (>0.6 m) than the HR remote sensing images used in related studies (Wagner et al., 2020). The LR increases the difficulty for the model to learn building features with irregular shapes and sizes. These characteristics of the open-source dataset necessitate learning global features, in which the MPViT backbone performs better than CNN.

Moreover, according to the inference time shown in Table 6, the MPViT model was extended by 30 %, 20 %, 14 %, and 8 %, higher than ResNeXt-101 for non-built-up, rural, suburban, and urban areas, respectively. As the MPViT-b model (16.4G FLOPs) requires higher computational performance than ResNeXt-101 (~7.8 GFLOPs), feature extraction and fusion through the MPViT-b model consumed more computational time. However, although the performance is improved, we can assume that the model does not excessively increase the model complexity and reduce operational efficiency. The performance improvement is acceptable and meaningful in practical applications.

Although our study shows the advantages of using MPViT to extract buildings from satellite images, the results are significantly inadequate for practical applications, particularly for the generalization ability exhibited by different test areas and the ability to segment individual buildings in dense building areas. While considering global features facilitates the identification of the complete building and the reduction of false extractions, it also increases the risk of acquiring additional noise from open-source images. Such noise will accumulate layer by layer to increase feature depth, resulting in reduced recall because of the omission of some buildings. This issue limits the application scenarios of the open-source dataset. However, this limitation is partially related to the quality of the test dataset. Overall, there is still substantial scope for improving the accuracy of building extraction from LR open-source satellite images.

6.2. *Effect of color normalization and image super-resolution on building extraction*

Most solutions to the challenges created by the resolution discrepancies and color difference in the open-source satellite images dataset are associated with data augmentation, transfer-learning-based methods, and image-preprocessing-based methods. The data augmentation method typically includes certain transformations, such as color modifications and affine transformations, to increase the diversity of the training dataset. However, it cannot compensate for the high-frequency information contained in LR images (Yu et al., 2017). Transfer learning approaches can quickly retrain models using knowledge learned from upstream tasks. However, once the target region or the data source of the test image changes, sufficient new training data ought to be prepared again, limiting migration learning efficiency in practical applications (Pires de Lima and Marfurt, 2019). In step two of the proposed framework in this paper, color normalization based on reference images can mitigate the color differences between training and test images. Moreover, the SR method based on deep learning can enhance images' resolution and texture details. As the experimental results showed the potential of the color normalization and image super-resolution method in achieving building extraction using open-sourced images, the impact of color normalization and SR can be further analyzed. Therefore, we performed ablation experiments. We used the Mask R-CNN with MPViT backbone model trained in Section 5.4 to assess different images: original LR image, color-normalized LR image and color-normalized upsampled image. Considering the difference in resolution between the test (0.6 ~ 1 m) and training (0.3 m) images and exploring the impact caused by these differences, we used two different scaling factors: (1) enlarging the test image by 2×, with a resolution of 0.3 m, and (2) down-sampling the original image to 1.2 m followed by 4x SR, with a resolution of 0.3 m. The image of the suburban area (0.6 m resolution) was selected for the test, and the Real-ERSGAN was used for SR. The training setting and degradation approach for the 4x SR model is the same as that for the 2x model. Also, a GE image with a resolution of 0.3 m was used for comparison. Due to the different shooting angles and capture times between GSI and GE images, the ground truth of the GE image was adjusted accordingly. Table 7 supplies a quantitative evaluation of the building extraction by different image strategies. According to the results, the color-normalized image with 2x SR applied in Section 5.3 showed better performance than other strategies and was very close to the GE image. It demonstrated the advantage of the color normalization and image super-resolution method adopted in this study. Firstly, the recall of the toned image improved by 0.02, suggesting that color normalization not only improves the visual appearance but benefits the building extraction too. Besides, the precision and recall of images after 2x SR improved by 0.05 and 0.04, respectively. This shows that SR could facilitate the complete utilization of the internal information of LR remote sensing images. Furthermore, the precision and recall of the processed GSI image were only 0.01 lower than the GE image, the image source of which is the same as the training data. It showed that the proposed color-normalization and image super-resolution could mitigate the image differences between the test and training sets. However, the accuracy with 4x SR is severely degraded. This is because the Real-ESRGAN model is trained with synthetic LR images as input for the edges and details of SR images. However, the image lacks enough detail with a 1.2 m resolution to support subsequent reconstructing. Besides, the noise and blur of LR images are equally amplified after SR, increasing the difficulty of extracting useful information by the instance segmentation model.

Furthermore, to demonstrate the effect of the method on the building extraction of a wide area, the number of buildings and the total footprint area of Hyogo Prefecture are aggregated in a 250 m grid. A total of 1,726,006 (29.12 km^2) out of the 3,301,488 buildings (32.46 km^2) are extracted, whereas 1,677,000 buildings in 27.68 km^2 are extracted without color normalization and image super-resolution. The number of buildings and footprint area increased by 3.0 % and 5.0 %, respectively. Subsequently, we calculated the symmetric mean absolute percentage error (SMAPE) of the predicted and actual values for comparison. The SMAPE is an accuracy measure based on relative errors, defined by Equation (9), where A_t is the actual value and P_t is the predicted value. The actual values were created based on the urban planning survey data of Hyogo Prefecture obtained from the Japan Fundamental Geospatial Data Download Service[2]. As shown in Fig. 21, irrespective of the SMAPE

Table 6
Inference time of two backbones in different areas.

	Backbone	Non-built-up	Rural	Suburban	Urban
Inference time (s)	ResNeXt-101	18.7	37.5	45.8	68.0
	MPViT-base	26.5	47.3	53.3	74.0

Table 7
Object-wise metric of prediction results in a selected suburban area under different image color normalization and image super-resolution strategies.

Image	Resolution (m)	Precision	Recall	F1
Original LR	0.6	0.85	0.68	0.76
Color-normalized LR	0.6	0.85	0.70	0.77
Color-normalized SR 2x	0.3	**0.90**	**0.74**	**0.81**
Color-normalized SR 4x	0.3	0.83	0.61	0.70
GE HR	0.3	0.91	0.75	0.82

[2] URL: https://fgd.gsi.go.jp/download/menu.php (Access on 03/25/2022).

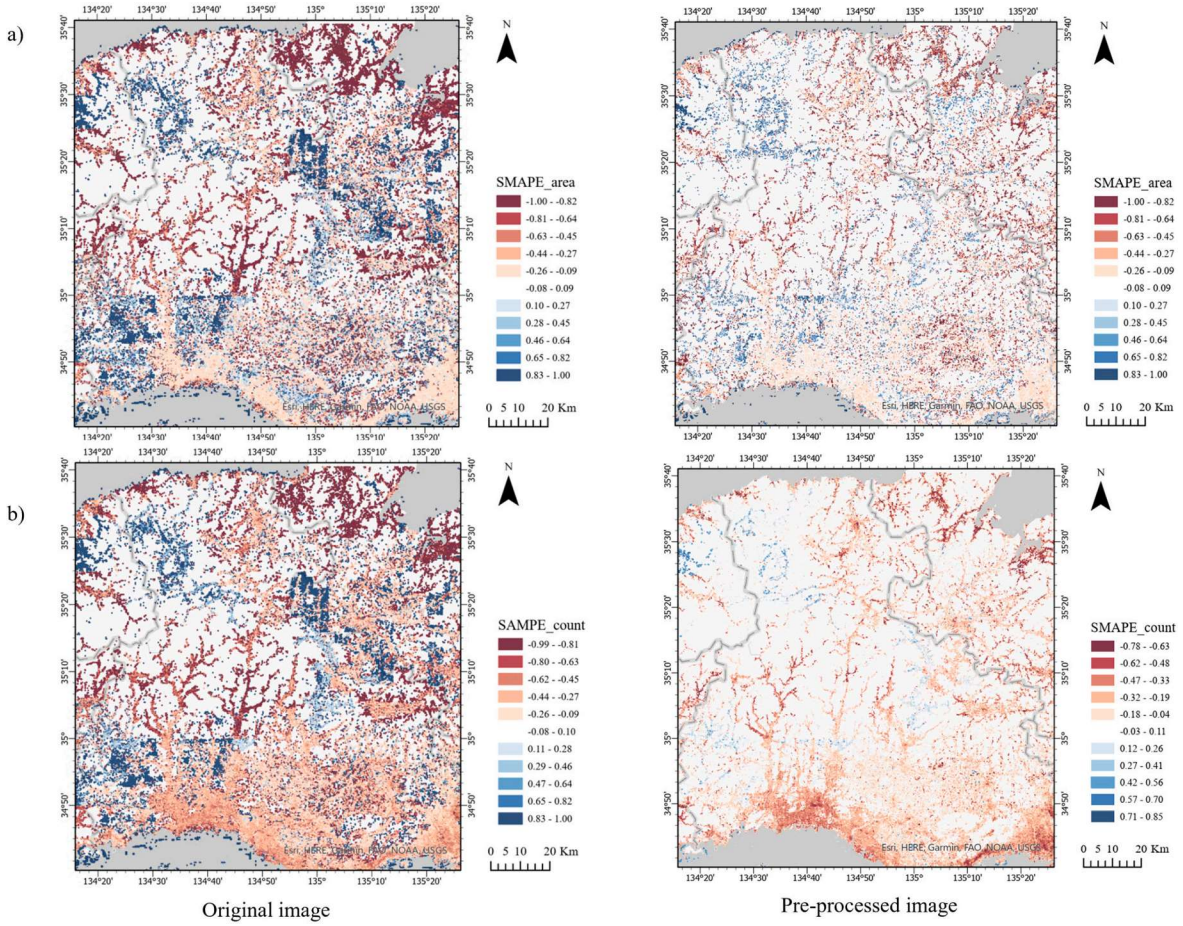

Fig. 21. Spatial distribution of SMAPE values calculated based on (a) footprint area and (b) the number of buildings in 250 m grid unit; the lighter the color, the smaller the error. The left figure is the result of original GSI image; that on the right is processed through SR and color normalization.

values calculated based on either metric, the image processed through SR and color normalization (right column) was significantly better than the result of the original images (left column). In particular, the phenomenon of over-detection is well mitigated (blue areas), demonstrating that the proposed method can be applied to large-scale building extraction.

Fig. 22. Effect on building extraction when the reference image used for color normalization fails to match the original image. The reference image from GE with a resolution of 3 m was acquired in September 2020, while the source image was acquired in 2012.

$$\mathrm{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|P_t - A_t|}{|P_t| + |A_t|} \quad (9)$$

In addition, we would like to discuss the issue regarding the color-normalization method, where the reference image did not match the details of the original one. Problems such as feature coordinate deviations and land cover changes are inevitable due to different camera views and shooting times. That can lead to significant variances in the color of a building roof, affecting building extraction. To test the problem's effect, we applied a reference image with a long interval from the test image for color-normalization and attempted building extraction. As shown in Fig. 22, the reference image was taken eight years apart from the original one, and part of the woods have been cut down. Due to vast differences in the ground objects' color, the forest and the surrounding building in the toned images are tinted with earthy ochre. However, there is no significant change in the prediction result before and after color normalization, with only minor differences in outline detail. This is due to the use of data augmentation in model training, which allows the model to recognize buildings with a larger range of variation in hue and brightness, improving the robustness of the model. Additionally, using reference images with lower resolution can mitigate the problem of image misalignment due to the angle of the shot (usually within a few pixels). However, satellite images taken at a similar time are still recommended to be used as reference maps to avoid significant changes in ground objects.

Studying building extraction based on open-source datasets is difficult because of the differences in color space, resolution, and image acquisition methods among multiple source datasets. However, the results of this study showed that the proposed color normalization and image super-resolution approach could improve the performance of building extraction using open-source satellite images, and satisfactory accuracy can be achieved with HR training data. Moreover, while previous investigations focused on the impact of the SR technique on semantic segmentation (Guo et al., 2019; Xu et al., 2021b; Zhang et al., 2021c), we extended the SR method to instance segmentation, which has essential academic and practical significance. This study can provide a reference for creating a high-quality building footprint map using an open-source remote sensing dataset.

6.3. Discussion on aggregated results for a wide area

In the previous section, we showed the aggregation results for large-scale building extraction. As shown in the right column in Fig. 21a, in terms of footprint areas (equivalent to pixel-wise level), SMAPE is controlled to a low range for most of the test area. In contrast, the number of buildings, which is a stricter indicator than the footprint area, underperforming in coastal areas at the bottom part of Fig. 21b, indicates that the accuracy of object-wise level still requires improvement, particularly for densely built-up areas in cities.

To further analyze the accuracy of large-scale building extraction, we performed regression analysis on the aggregated data of the number of buildings with the actual value according to land use patterns. Land-use patterns influence buildings' type, density, and distribution even in areas with similar architectural styles. We classified land use in Hyogo Prefecture into five typical cases: rural areas, low-rise building areas, dense low-rise building areas, high-rise building areas, and industrial and public facility areas. The definitions of each land use type are explained in Table 8. The example satellite images of the different land-use types are shown in Fig. 24, based on the land-use subdivision mesh of Hyogo Prefecture (Version 1.3), obtained from the Japan National Land Information Download Service[3].

As shown in Fig. 23, the regression results exhibited a coefficient of determination of 0.82 and a coefficient of regression of 0.45, which indicates that the predicted value was approximately 0.5 times that of the actual value. Although the point distribution was dispersed, the four cases can be grouped as enclosed within the red circles. Case 1 is located on both sides of the theoretical regression curve ($y = x$), where the absolute value of SMAPE is less than 0.1. Case 2 is close to the vertical axis, where the actual value = 0 and the predicted value is high. Case 3 is in the region close to the horizontal axis, where the actual value $\neq 0$ and the predicted value = 0. The fourth case is in the lower right region, where approximately the predicted value is $<$ one-third of the actual value..

Fig. 25 shows the proportion of the different cases. Case 1, where the prediction result is good, has the most significant number of cases. Its distribution is roughly consistent with rural areas and partially scattered in the low-rise building area. It is due to the low number and density of buildings in these two areas, where the buildings are less difficult to extract. In contrast, the distribution of Cases 2–4 is less numerous and scattered, and it is hard to find the pattern from the distribution map alone. To determine the possible causes of these three exceptional cases, we counted the number of grids for different land-use categories in Cases 2–4. As the total number of grids varied for different land-use categories, the statistics were normalized for easy comparison, as shown in Fig. 25. First, for Case 2, pixels with no buildings were classified as with buildings, primarily in rural areas and industrial and public facility areas, with minor occurrences in other regions. Although some scenes without buildings were filtered through scene classification, images without buildings were retained because of the relatively conservative classification principle. These areas contain objects similar to buildings, such as vegetable sheds and warehouses. For Case 3, approximately no buildings were extracted in the area with buildings, where the distribution trend is roughly the same as in Case 2. We examined the test image of Case 3 and the corresponding ground truth and identified three primary causes for this discrepancy. First, the model could not distinguish the buildings because of the image resolution ($>$1 m), despite the SR enhancement of the image details. This primarily occurred in rural areas with a low number of buildings. Second, the blurring of the toned image because of the cloud cover in the GE images that were used as a reference influenced the color normalization process. The third is the time difference between the capture time of satellite images and the ground truth data survey time. Although some buildings were annotated in the ground truth, the image showed that they have either been demolished or have not yet been constructed. Finally, Case 4′s frequency increased with increasing building complexity and density, particularly for dense low-rise and high-rise building areas. This was primarily because of the difficulty of finely dividing adjacent buildings with decreasing building spacing. It makes the predicted adjacent buildings grouped as one object. It implies that the prediction result has a high degree of adhesion. Secondly, it is caused by the object-wise judgment criteria based on IoU. Most of the model's misclassifications occur mainly near building boundaries. The larger the building size, the smaller the proportion of misclassified pixels and the higher the IoU. The relatively lower resolution of the open-sourced images makes the buildings take up fewer

Table 8
Definition of the specific land-use categories.

Land-use categories	Definition
Rural area	The land that is used for agricultural purposes.
Low-rise buildings area	The land where residential buildings of three stories or less are distributed in clusters.
Dense low-rise building area	The land with a high density of three stories or fewer residential buildings.
High-rise building area	Residential and urban areas, where buildings are densely built, comprise commercial and business buildings and 4 stories or more condominiums.
Industrial and public facility area	The land is used for industrial and public utilities such as factories, docks, hospitals, and schools.

[3] URL: https://nlftp.mlit.go.jp/ksj/index.html#chiiki (Access on 03/25/2022).

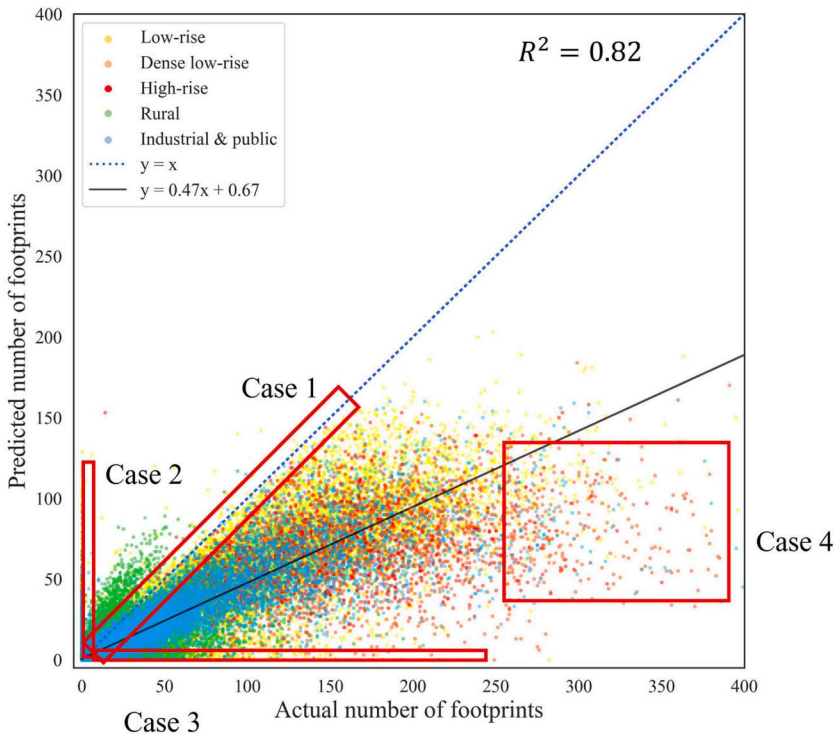

Fig. 23. Regression results for 250 m grid aggregated data of the number of buildings of Hyogo Prefecture. Dots with different colors represent different land-use categories. Red boxes mark Cases 1–4. The horizontal axis plots the actual values, and the vertical axis plots the predicted values. The blue line is the theoretical curve ($y = x$), and the black line is the actual regression curve ($y = 0.47x + 0.67$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

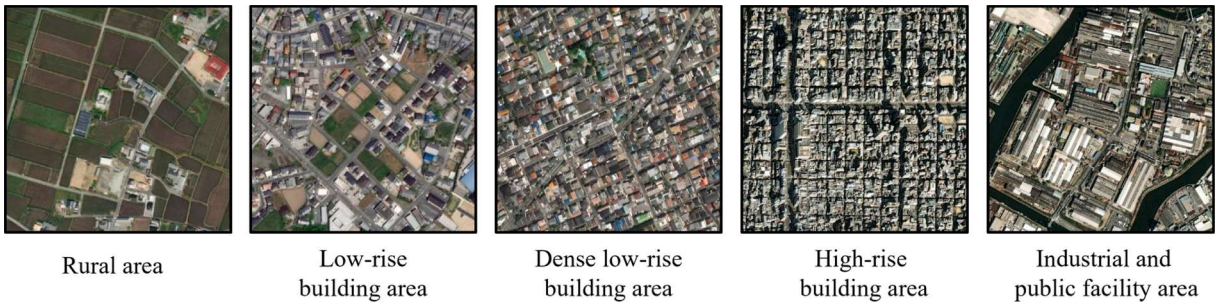

Fig. 24. Spatial distribution of (a) land-use categories and (b) Cases 1–4, and (c) examples of specific land-use categories in Hyogo Prefecture.

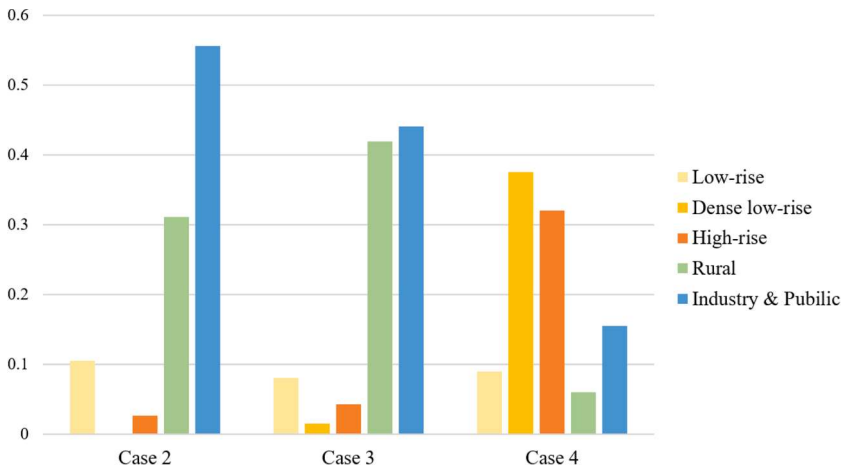

Fig. 25. Proportion of different land-use categories in Cases 2–4, corresponding to the case that pixels with no buildings were classified as buildings, approximately no buildings were extracted in the area with buildings, and the predicted value < one-third of the actual value in built-up areas.

pixels, making prediction more difficult.

6.4. Limitation

Although these results showed the method's potential in achieving building extraction using LR images, further investigation of the mechanism and limitations of the proposed method is necessary. Based on the relationship between the resolutions of the training data (0.3 m) and the GSI image used (mostly 0.6 m), we trained a 2 × SR model and assessed its effect on the performance of building extraction. Beyond that, the potential of using a 4x SR model is only briefly discussed. In practical applications, the resolution of open-source remote sensing data is variable (including the GSI dataset used in this paper). Therefore, the SR model should be investigated with different magnification scales, and the effect on images with different resolutions must be evaluated. Besides, we intend to apply the synthesized open-sourced HR images as training data via super-resolution to test the generalization abilities of the model. Second, to solve the issue of low recall, the instance-segmentation model needs further enhancement in terms of the ability to segment adjacent buildings, particularly for the high-density built-up urban areas and complex additions in rural areas. The influence of the side part of the high-rise in the non-orthographic image also needs to be eliminated. Finally, we selected a prefecture-sized area with a relatively uniform architectural style as the study area. To test whether the framework developed in this study applies to different styles of buildings in other countries or regions, updating and supplementing the training set is a challenge. It is difficult to obtain massive high-precision remote sensing images and create annotations. Therefore, in the future, it is of great practical importance to study unsupervised learning-based building extraction, which helps enhance the model's generalization ability and reduces the cost of practical applications.

7. Conclusion

We proposed a framework for large-scale building extraction using an SR-based instance segmentation algorithm suitable for open-sourced remote sensing datasets. The framework comprised four primary steps: scene classification, color normalization and image super-resolution, building instance segmentation, and scene mosaicking. For open-source datasets with relatively lower resolution (>0.6 m) and color variation, color normalization based on a reference map and the fine-tuned Real-ESRGAN SR model could improve the image quality visually, enhancing color consistency and textural details and contributes to the performance of building extraction. Furthermore, the scene classification model based on the ConvNeXt-base network could effectively distinguish between images with and without buildings to reduce the computational load of building extraction. To demonstrate the validity of the proposed framework, the proposed method was used to extract a wide range of buildings footprints in Hyogo Prefecture (19,187 km^2). We extracted 1,726,006 (29.12 km^2) of the 3,301,488 buildings (32.46 km^2), where the number of buildings and footprint area increased by 3.0 % and 5.0 %, respectively, before color normalization and image super-resolution. Besides, the experimental results showed that the improved Mask R-CNN model based on the MPViT backbone achieved F1 scores of 0.71, 0.70, 0.81, and 0.67 for non-built-up, rural, suburban, and urban areas, respectively, which indicate better performance than the ResNext-101 backbone and other mainstream instance segmentation approach. Furthermore, by using a reference image with a lower resolution (3 m) and data enhancement during model training, the effects of the ground object difference between the reference and original images can be avoided. Thus, we demonstrated the potential of the proposed approach in demonstrating the use of widely available low-resolution open-source data to obtain acceptable building extraction results. This method is practical and particularly useful when additional datasets of HR remote sensing images are unavailable, which may broaden the application of open-source data and benefit related downstream tasks. In the future, we aim to develop SR models with different magnification scales and improve the model performance for dividing adjacent buildings. Furthermore, building extraction based on unsupervised learning will be studied to increase the scope of applicability of the proposed method.

Declaration of Competing Interest

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

Acknowledgement

We are grateful to Dr Hiroyuki Miyazaki, Center for Spatial Information Science (CSIS), University of Tokyo, for providing parts of the training dataset and code used. Besides, we would like to thank Editage (www.editage.com) for English language editing.

References

Acuna, D., Ling, H., Kar, A., Fidler, S., 2018. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 859–868.

Belgiu, M., Drăguţ, L., 2014. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. ISPRS J. Photogramm. Remote Sens. 96, 67–75.

Carvalho, O.L.F.d., de Carvalho Junior, O.A., Albuquerque, A.O.d., Bem, P.P.d., Silva, C. R., Ferreira, P.H.G., Moura, R.d.S.d., Gomes, R.A.T., Guimaraes, R.F., Borges, D.L., 2020. Instance segmentation for large, multi-channel remote sensing imagery using mask-RCNN and a mosaicking approach. Remote Sensing 13, 39.

Castrejon, L., Kundu, K., Urtasun, R., Fidler, S., 2017. Annotating object instances with a polygon-rnn. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5230–5238.

Chen, M., Wu, J., Liu, L., Zhao, W., Tian, F., Shen, Q., Zhao, B., Du, R., 2021. DR-Net: An improved network for building extraction from high resolution remote sensing image. Remote Sens. (Basel) 13, 294.

Chen, S., Han, Z., Dai, E., Jia, X., Liu, Z., Xing, L., Zou, X., Xu, C., Liu, J., Tian, Q., 2020. Unsupervised image super-resolution with an indirect supervised path. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 468–469.

Cheng, G., Xie, X., Han, J., Guo, L., Xia, G.-S., 2020. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 3735–3756.

Contributors, M., 2020. OpenMMLab's Image Classification Toolbox and Benchmark. \url{https://github.com/open-mmlab/mmclassification}.

Cresson, R., Saint-Geours, N., 2015. Natural color satellite image mosaicking using quadratic programming in decorrelated color space. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8, 4151–4162.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703.

Cui, H., Zhang, G., Wang, T., Li, X., Qi, J., 2021. Combined Model Color-Correction Method Utilizing External Low-Frequency Reference Signals for Large-Scale Optical Satellite Image Mosaics. IEEE Trans. Geosci. Remote. Sens. 59, 4993–5007.

Das, A., Chandran, S., 2021. Transfer learning with res2net for remote sensing scene classification. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, pp. 796–801.

de Bem, P.P., de Carvalho Júnior, O.A., de Carvalho, O.L.F., Gomes, R.A.T., Fontes Guimarães, R., 2020. Performance analysis of deep convolutional autoencoders with different patch sizes for change detection from burnt areas. Remote Sens. (Basel) 12, 2576.

Dong, L., Shan, J., 2013. A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. ISPRS J. Photogramm. Remote Sens. 84, 85–99.

Feurer, M., Hutter, F., 2019. Hyperparameter optimization. Automated machine learning. Springer, Cham, pp. 3–33.

Gatys, L.A., Ecker, A.S., Bethge, M., 2015. A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576.

Girard, N., Smirnov, D., Solomon, J., Tarabalka, Y., 2021. Polygonal building extraction by frame field learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5891–5900.

Glasner, D., Bagon, S., Irani, M., 2009. Super-resolution from a single image, 2009 IEEE 12th international conference on computer vision. IEEE 349–356.

Guo, Z., Wu, G., Song, X., Yuan, W., Chen, Q., Zhang, H., Shi, X., Xu, M., Xu, Y., Shibasaki, R., 2019. Super-resolution integrated building semantic segmentation for multi-source remote sensing imagery. IEEE Access 7, 99381–99397.

Gupta, R., Shah, M., 2021. In: Rescuenet: Joint building segmentation and damage assessment from satellite imagery, pp. 4405–4411.

Gupta, V., Sadana, R., Moudgil, S., 2019. Image style transfer using convolutional neural networks based on transfer learning. Int. J. Comput. Syst. Eng. 5, 53–60.

Haklay, M., Weber, P., 2008. Openstreetmap: User-generated street maps. IEEE Pervasive Comput. 7, 12–18.
Hamaguchi, R., Hikosaka, S., 2018. Building detection from satellite imagery using ensemble of size-specific detectors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 187–191.
Hao, C., Li, Z., Haibin, A., Biao, X., Zhonghui, W., 2017. Large collection satellite images color normalization algorithm based on tone reference map. Acta Geodaetica Et Cartographica Sinica 46, 1986.
Haut, J.M., Paoletti, M.E., Fernández-Beltran, R., Plaza, J., Plaza, A., Li, J., 2019. Remote sensing single-image superresolution based on a deep compendium model. IEEE Geosci. Remote Sens. Lett. 16, 1432–1436.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.
Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B., 2020. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE Trans. Geosci. Remote Sens. 59, 4340–4354.
Huang, X., Wang, C., Li, Z., 2019a. High-resolution population grid in the CONUS using microsoft building footprints: A feasibility study. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Geospatial Humanities, pp. 1–9.
Huang, Y., Jin, Y., 2022. Aerial Imagery-Based Building Footprint Detection with an Integrated Deep Learning Framework: Applications for Fine Scale Wildland-Urban Interface Mapping. Remote Sens. (Basel) 14, 3622.
Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019b. Mask scoring r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6409–6418.
Ji, S., Wei, S., Lu, M., 2018. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Trans. Geosci. Remote Sens. 57, 574–586.
Keys, R., 1981. Cubic convolution interpolation for digital image processing. IEEE Trans. Acoust. Speech Signal Process. 29, 1153–1160.
Khan, M.A., Akram, T., Zhang, Y.-D., Sharif, M., 2021. Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. Pattern Recogn. Lett. 143, 58–66.
Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., Cho, K., 2019. Augmentation for small object detection. arXiv preprint arXiv:1902.07296.
Latha, T.P., Sundari, K.N., Cherukuri, S., Prasad, M., 2019. Remote Sensing UAV/Drone technology as a tool for urban development measures in APCRDA. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 42, 525–529.
Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 2017. Photo-realistic single image super-resolution using a generative adversarial network, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681-4690.
Lee, Y., Kim, J., Willette, J., Hwang, S.J., 2021. MPViT: Multi-Path Vision Transformer for Dense Prediction. arXiv preprint arXiv:2112.11010.
Li, Z., Wegner, J.D., Lucchi, A., 2019. Topological map extraction from overhead images, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1715-1724.
Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. Swinir: Image restoration using swin transformer, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1833-1844.
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. European conference on computer vision. Springer 740–755.
Liu, Y., Liu, J., Ning, X., Li, J., 2022. MS-CNN: multiscale recognition of building rooftops from high spatial resolution remote sensing imagery. Int. J. Remote Sens. 43, 270–298.
Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., De Nadai, M., 2021a. Efficient training of visual transformers with small-size datasets. arXiv preprint arXiv:2106.03746.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows, Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012-10022.
Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022b. A ConvNet for the 2020s. arXiv preprint arXiv:2201.03545.
Long, Y., Xia, G.-S., Li, S., Yang, W., Yang, M.Y., Zhu, X.X., Zhang, L., Li, D., 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 4205–4230.
Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.
Ma, A., Wan, Y., Zhong, Y., Wang, J., Zhang, L., 2021. SceneNet: Remote sensing scene classification deep learning network using multi-objective neural evolution architecture search. ISPRS J. Photogramm. Remote Sens. 172, 171–188.
Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). IEEE 3226–3229.
Mnih, V., 2013. Machine learning for aerial image labeling. University of Toronto (Canada).
Neubeck, A., Van Gool, L., 2006. Efficient non-maximum suppression, 18th International Conference on Pattern Recognition (ICPR'06). IEEE, pp. 850-855.
O'Shea, K., Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
Paisitkriangkrai, S., Sherrah, J., Janney, P., Hengel, V.-D., 2015. Effective semantic pixel labelling with convolutional networks and conditional random fields, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 36-43.
Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X., 2020. Deep snake for real-time instance segmentation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8533-8542.
Pires de Lima, R., Marfurt, K., 2019. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. Remote Sens. (Basel) 12, 86.
Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.
Sango, K., 2009. Development of Electronic National Land Basic Maps (orthoimages). Map 47, 15–20.
Shao, Z., Tang, P., Wang, Z., Saleem, N., Yam, S., Sommai, C., 2020. BRRNet: A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images. Remote Sens. (Basel) 12, 1050.
Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1874-1883.
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. arXiv preprint arXiv:2107.12283.
Suzumura, T., Sugiki, A., Takizawa, H., Imakura, A., Nakamura, H., Taura, K., Kudoh, T., Hanawa, T., Sekiya, Y., Kobayashi, H., 2022. mdx: A Cloud Platform for Supporting Data Science and Cross-Disciplinary Research Collaborations. arXiv preprint arXiv: 2203.14188.
Tang, X., Ma, Q., Zhang, X., Liu, F., Ma, J., Jiao, L., 2021. Attention consistent network for remote sensing scene classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2030–2045.
Tong, X.-Y., Xia, G.-S., Lu, Q., Shen, H., Li, S., You, S., Zhang, L., 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 237, 111322.
Touzani, S., Granderson, J., 2021. Open data and deep semantic segmentation for automated extraction of building footprints. Remote Sens. (Basel) 13, 2578.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.
Wagner, F.H., Dalagnol, R., Tarabalka, Y., Segantine, T.Y., Thomé, R., Hirye, M., 2020. U-net-id, an instance segmentation model for building extraction from satellite images—Case study in the Joanopolis City. Brazil. Remote Sensing 12, 1544.
Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021a. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 568–578.
Wang, X., Xie, L., Dong, C., Shan, Y., 2021b. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1905–1914.
Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C., 2018. Esrgan: Enhanced super-resolution generative adversarial networks. Proceedings of the European conference on computer vision (ECCV) workshops.
Wang, Y., Li, S., Teng, F., Lin, Y., Wang, M., Cai, H., 2022. Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province. China. Remote Sensing 14, 265.
Wei, S., Ji, S., 2021. Graph convolutional networks for the automated production of building vector maps from aerial images. IEEE Trans. Geosci. Remote Sens. 60, 1–11.
Wei, S., Zhang, T., Ji, S., 2021. A concentric loop convolutional neural network for manual delineation level building boundary segmentation from remote sensing images. IEEE Trans. Geosci. Remote Sens.
Wen, Q., Jiang, K., Wang, W., Liu, Q., Guo, Q., Li, L., Wang, P., 2019. Automatic building extraction from Google Earth images under complex backgrounds based on deep instance segmentation network. Sensors 19, 333.
Wu, T., Hu, Y., Peng, L., Chen, R., 2019. Improved anchor-free instance segmentation for building extraction from high-resolution remote sensing images. Remote Sens. (Basel) 12, 2910.
Xiao, Q., Liu, B., Li, Z., Ni, W., Yang, Z., Li, L., 2021. Progressive data augmentation method for remote sensing ship image classification based on imaging simulation system and neural style transfer. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 9176–9186.
Xie, R., Xia, M., Yao, J., Li, L., 2018. Guided color consistency optimization for image mosaicking. ISPRS J. Photogramm. Remote Sens. 135, 43–59.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.
Xu, L., Liu, Y., Yang, P., Chen, H., Zhang, H., Wang, D., Zhang, X., 2021a. HA U-Net: Improved Model for Building Extraction From High Resolution Remote Sensing Imagery. IEEE Access 9, 101972–101984.
Xu, P., Tang, H., Ge, J., Feng, L., 2021b. ESPC_NASUnet: An End-to-End Super-Resolution Semantic Segmentation Network for Mapping Buildings From Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 5421–5435.
Xu, W., Xu, Y., Chang, T., Tu, Z., 2021c. Co-scale conv-attentional image transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9981–9990.
Xu, Y., Luo, W., Hu, A., Xie, Z., Xie, X., Tao, L., 2022. TE-SAGAN: An Improved Generative Adversarial Network for Remote Sensing Super-Resolution Images. Remote Sens. (Basel) 14, 2425.
Xue, L., Li, Z., Qingdong, W., Haibin, A., 2020. Multi-temporal remote sensing imagery semantic segmentation color consistency adversarial network. Acta Geodaetica et Cartographica Sinica 49, 1473.

Yan, H., Li, Z., Li, W., Wang, C., Wu, M., Zhang, C., 2021. ConTNet: Why not use convolution and transformer at the same time? arXiv preprint arXiv:2104.13497.
Yang, N., Tang, H., 2020. GeoBoost: An incremental deep learning approach toward global mapping of buildings from VHR remote sensing images. Remote Sens. (Basel) 12, 1794.
Yi, Y., Zhang, Z., Zhang, W., Zhang, C., Li, W., Zhao, T., 2019. Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network. Remote Sens. (Basel) 11, 1774.
Yu, X., Wu, X., Luo, C., Ren, P., 2017. Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. GIScience & Remote Sensing 54, 741–758.
Yu, Y., Zhang, K., Yang, L., Zhang, D., 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput. Electron. Agric. 163, 104846.
Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.
Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6023–6032.
Zhai, Y., Chen, S., 2020. A Seismic Hazard Prediction System for Urban Buildings Based on Time-History Analysis. Mathematical Problems in Engineering 2020.
Zhang, G., Lu, X., Tan, J., Li, J., Zhang, Z., Li, Q., Hu, X., 2021a. Refinemask: Towards high-quality instance segmentation with fine-grained features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6861–6869.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.
Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., 2022. Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746.
Zhang, K., Liang, J., Van Gool, L., Timofte, R., 2021b. Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4791–4800.
Zhang, L., Dong, R., Yuan, S., Li, W., Zheng, J., Fu, H., 2021c. Making low-resolution satellite images reborn: a deep learning approach for super-resolution building extraction. Remote Sens. (Basel) 13, 2872.
Zhang, T., Tang, H., Ding, Y., Li, P., Ji, C., Xu, P., 2021d. FSRSS-Net: High-resolution mapping of buildings from middle-resolution satellite images using a super-resolution semantic segmentation network. Remote Sens. (Basel) 13, 2290.