MSFTTrans: a multi-task frequency-spatial learning transformer for building extraction from high spatial resolution remote sensing images

Bo Yu\textsuperscript{a,b}, Fang Chen\textsuperscript{a,b,c}, Ning Wang\textsuperscript{a,b,c}, Lu Yang\textsuperscript{e}, Haiping Yang\textsuperscript{f} and Lei Wang\textsuperscript{a,b}

\textsuperscript{a}International Research Center of Big Data for Sustainable Development Goals, Beijing, China; \textsuperscript{b}Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China; \textsuperscript{c}University of Chinese Academy of Sciences, Beijing, China; \textsuperscript{d}State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China; \textsuperscript{e}Pattern Recognition and Intelligence Vision Lab, Beijing University of Posts and Telecommunications, Beijing, China; \textsuperscript{f}College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China

\section*{ABSTRACT}
Building extraction is significant in urban planning, economic evaluation, and driverless technology development. However, automatic building extraction from high spatial resolution remote sensing images has been a challenging task due to the various building shapes and colors, imaging conditions, and complex background objects. Current methods in building extraction are generally based on deep convolution networks, and they mostly use an encoder-decoder architecture, wherein detailed building features and small buildings are easily omitted in continuous convolution operations. Moreover, buildings with blurred boundaries are only completely extracted with difficulty. To meet these challenges, we propose a multi-task architecture of frequency-spatial learning Transformer to extract buildings from high spatial resolution remote sensing images. Different from current architecture, we designed a frequency-spatial learning module in the framework of multi-task to synthesize the multi-scale spatial features and frequency decomposition features of high-resolution image. Spiking convolution is proposed in this study to enhance the frequency features of buildings by mimicking the neural transmission in human brains. In this way, multi-scale building features can be better preserved and distinguished from background objects. Moreover, a masked-attention Transformer is adopted to improve multi-scale building mask prediction accuracy by synthesizing successive pixel-wise up-sampled feature maps. We also propose a strategy to evaluate the practical transferability of the proposed method by mimicking practical application cases through training and evaluating images with different spatial resolutions from different study areas and datasets. Experiments using five public building datasets (WHU-Building Satellite Dataset I, WHU-Building Satellite Dataset II, Massachusetts Buildings Dataset, Inria Aerial Image Dataset, xBD Building Dataset) demonstrate the strong potential applicability of our proposed method for practical application cases. Our method outperforms five recently proposed state-of-the-art semantic segmentation methods with 36.60% accuracy improvement on extracted buildings and approximately 53.55% recall progress in extracting small building instances. The implementation code will be released after the paper is published.

\section*{1. Introduction}
Buildings, as a remarkable indicator of human activities in economic development and urban construction (Shibiao et al. 2018), have been recognized as one of the main basic geological data and has received extensive attention for its significance (Li et al. 2022). Due to advances in earth observation technology, numerous high spatial resolution satellites have been launched and released their product images (Chen et al. 2021a; Jia et al. 2022; Yu, Chen, and Chong 2020). The easy access to numerous remote sensing images makes it possible to update building inventories frequently for urban planning (Guo et al. 2021a), population estimation (Xie, Weng, and Weng 2015), navigation (Xie, Weng, and Weng 2015), and emergency rescue (Blaschke 2010). Optical remote sensing images and images achieved using active techniques, including LiDAR, POLSAR and SAR, are mainly used for building extraction (Cao et al. 2020; Deng et al. 2019; Huang et al. 2019). Compared with optical remote sensing images, LiDAR and SAR images are difficult to achieve for large-scale areas with a high frequency, with its high acquisition cost and geometric distortions (Guo et al. 2021b; Sun et al. 2022). Manually updating building inventories for large-scale areas is tedious and time-consuming.
consuming. Therefore, it is significant to develop automatic algorithms to extract buildings from high spatial resolution images. However, the high spectral and textural similarity between buildings and background objects and the variety in building appearance and shape bring about great challenges in extracting intact buildings from remote sensing images (Zhang et al. 2020). Moreover, due to the inconsistent quality of remote sensing images from different sensors, a given object is likely to take different hues and textures (Schiewe 2002). This produces significant challenges for robust automatic building extraction with high transferability. Consequently, much work has focused on efficient automatic methods of building extraction. These methods can be generally categorized into feature-engineering-based and deep-learning-based methods.

Feature-engineering-based methods generally synthesize geometrical, spectral, textural, contextual, and shadow characteristics to distinguish buildings from background objects in remote sensing images (Shao et al. 2020). Edge detection is a commonly used technique to achieve the geometrical characteristics of buildings and has proved to be effective in extracting rectangular buildings, but less so in extracting buildings with complex shapes (Xiaohuan, Wan, and Wang 2016). Object-based multi-scale image segmentation methods have contributed significantly to building extraction (Janalipour and Mohammadzadeh 2016), but they are likely to trigger the phenomena of “over segmentation” and “under segmentation,” wherein segmented buildings are either fragmented or mixed with background objects (Jiang et al. 2008). Based on spectral and geometrical features, several indexes have been constructed to map buildings (Zhang et al. 2006; Hanqiu 2008), but the corresponding threshold has to be set by trial and error. To dispense with the need of setting thresholds manually, many studies have adopted machine learning methods to map buildings. These methods learn the thresholds of multiple features using training samples. Clustering (Zhu and Guo 2014), random forest (Du, Zhang, and Zhang 2015), and support vector machine (SVM) (Turker and Koc-San 2015) are widely used methods. Building extraction has been significantly improved by machine learning methods (Schlosser et al. 2020) but still heavily relies on hand-crafted features. Moreover, the designed features are manually set according to the building distribution pattern in the training images, which may be invalid for other images (Mohit, Sahay, and Rathore 2018), especially from different satellites and study areas. The limited transferability of machine learning methods makes them less applicable in practical cases, such as emergency rescue.

The advent of deep convolutional neural networks has made it possible to dispense with the feature-engineering process by learning from raw data directly (Zheng et al. 2021). Cutting-edge deep learning methods have contributed significantly to automatic building extraction (Luo, Pengpeng, and Yan 2021), owing to the large quantity and variety of publicly available challenging datasets (Shunping, Wei, and Meng 2019; Mnih 2013). Semantic segmentation is a widely used deep learning framework for building extraction (Wang and Miao 2022), which has the capability of assigning labels to each pixel of interest (Guo et al. 2022), since the propose of Fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015) by enabling the network to free from the requirement of restricted size of input image through pixel-wise segmentation. Numerous studies have explored efficient network structures for automatic building extraction within a semantic segmentation framework. Wu et al. (Guangming et al. 2018) proposed a multi-constraint FCN to extract buildings from aerial images. To reduce information loss in a pooling layer, Shariah et al. (Sherrah 2016) implemented FCN without a pooling layer to retain the detailed ground information from the original image. Yi et al. (Yi et al. 2019) designed DeepResUnet to extract urban buildings by synthesizing U-net (Ronneberger, Fischer, and Brox 2015) and residual learning architecture, achieving an improvement in accuracy of 3.52% compared with U-net. Recently developed semantic segmentation networks in computer vision, including PSPNet (Zhao et al. 2017), DANet (Fu et al. 2019), SegNet (Badrinarayanan, Kendall, and Cipolla 2017), and the series of DeepLab networks (including DeepLabv1, DeepLabv2, DeepLabv3, and DeepLabv3+) (Chen et al. 2014, 2018, 2017, 2018a) have been adopted to extract buildings and achieved remarkable performance improvements (Abdollahi, Pradhan, and Alamri 2020; Baheti et al. 2020; Zhao et al. 2021). However, the extracted buildings mostly have the issue of fragmented building segmentations, especially in datasets with different imaging sensors from the training datasets (Yu et al. 2022), hindering practical applications of the proposed methods.
In order to address the issue of incomplete extracted buildings, a variety of studies have introduced boundary regularization and conditional random fields to finetune the extracted buildings (Guo et al. 2022; He and Jiang 2021; Liao et al. 2021; Zhu et al. 2020). Boundary regulation has been proved to be effective in improving boundary extraction accuracy as a post-processing step (Guo et al. 2022; Yang et al. 2022) It has inspired the use of refinement models as auxiliary tasks to finetune building extraction results, as conducted in MagNet (Huynh et al. 2021), which shows remarkable progress. However, in practical applications, the multi-step framework in building extraction is cumbersome compared with end-to-end semantic segmentation networks. Apart from multi-step framework, there has been numerous research adopting building boundary to assist building extraction. The works of (He and Jiang 2021) adopts edge detection to supplement building extraction and achieved state-of-the-art performance. Joint learning of contour and building structure has been proposed as well to preserve boundaries of extracted buildings (Liao et al. 2021). The boundary-assisted models are sensitive to the performance of edge detection. In terms of the cases where edge detection fails to achieve accurate building boundaries, the performance of the building extraction models will be heavily impacted.

Most of the published models above are evaluated on the sub-datasets that are subtracted from the same dataset as the training datasets (Wang et al. 2022) or on the datasets with limited variability in imaging sensors or spatial resolutions from that of training datasets (Benjamin et al. 2022). The actual model transferability is difficult to be evaluated, and the capability for practical applications remains to be estimated objectively. Moreover, to deal with the fragmented buildings extracted, and the low transferability of current works, we proposed a multi-task frequency and spatial learning Transformer to extract buildings from datasets with different imaging sensors, different spatial resolutions and different study areas. Multi-task learning framework has been embedded to the proposed network structure to supplement building extraction (Diakogiannis et al. 2020), and delivered considerable performance improvement over single-task framework. Transformer was firstly proposed by (Vaswani et al. 2017) for natural language processing and has recently been used for the semantic segmentation of natural images and medical images, with remarkable efficiency and accuracy (Chen et al. 2021b; Cheng, Schwing, and Kirillov 2021a). It is an encoder-decoder network, organized in residual architecture without convolution and pooling layers. Based on the Transformer decoder proposed in (Cheng et al. 2021b), we propose a Transformer in frequency domain to synthesize the concatenated feature embeddings learned from the spatial and frequency domains. The frequency domain readily distinguishes blurred objects by capturing different frequencies (Qian et al. 2020), but the spatial domain is better at learning multi-scale features. As improvement of the Transformer proposed for semantic segmentation (Cheng et al. 2021), our Transformer is designed by synthesizing learning in the frequency domain and in the spatial domain in a multi-task architecture, which can contribute to improving building extraction from images of different imaging sensors. Moreover, since spikes are known as information transmission neurons of human brain (Ghosh-Dastidar and Adeli 2009), spiking convolution is introduced in this study to strengthen the feature learning capability by simulating the powerful learning ability of human brain. The main contributions of this study are as follows:

1. We propose a multi-task spatial-learning and frequency-based Transformer to automatically extract buildings from high spatial resolution images. The proposed model can capture detailed building features and is sensitive to small buildings via the joint learning of the frequency domain and the spatial domain organized in a multi-task manner.

2. Since improving building extraction performance is our main aim in this study, spiking convolution is introduced in our frequency learning task to enhance the frequency feature capability for automatic building extraction with a low computation cost, which makes it promising for future practical applications.

3. The impact of frequency learning and spatial learning on building extraction is explored, in terms of different evaluation datasets. That provides foundation for model construction of building extraction in different cases.

2. Proposed architecture

2.1 Overview

Typical encoder-decoder network structures are widely used to synthesize multi-scale features from
input image for building extraction. However, continuous encoded features are likely to lose spatial details of small buildings with low spatial resolution after continuous down-sampling. Therefore, we use the frequency features of a high spatial resolution image with multi-scale spatial features in a multi-task framework (shown in Figure 1). We feed the concatenated features to a frequency learning-based masked-attention Transformer decoder after they are enhanced by spiking convolution for multi-scale building extraction.

2.2 Frequency-spatial feature learning component

The frequency-spatial feature learning component (Figure 1) is a multi-task architecture, consisting of two tasks, feature learning in the spatial domain and in the frequency domain. For feature learning in the spatial domain, we used ResNet-101 (Kaiming et al. 2016) as the backbone encoder, as shown in Figure 2.

ResNet-101 is a network of 101 layers organized in a residual manner, wherein the first \( 7 \times 7 \) convolution layer in works of (Kaiming et al. 2016) is replaced by three continuous \( 3 \times 3 \) convolutions, as commonly conducted in semantic segmentation (Cheng et al. 2020). Layers L1, L2, and L3 correspond to the ResNet decoding layer in Figure 1. Different from local feature learning by continuous convolutions in spatial domain, feature learning in frequency domain is conducted by enhancing the frequency spectrum globally. Frequency domain is a domain measuring variation degree of pixel intensities, wherein low frequency indicates gradual variation of pixel intensities (ground objects other than edges), and high frequency indicates sharp variation (ground object boundaries and noises). Therefore, frequency domain is more sensitive and better in maintaining small buildings by learning high frequency features globally, while spatial domain is easy to filter out small buildings by continuous local convolutions.

Figure 1. The proposed MSFTrans, consisting of frequency-spatial feature learning (see Section 2.2) and building prediction (see Section 2.3) components. DCT is short for discrete cosine transform, and \( \text{DCT}^{-1} \) is the inverse operation. \( F^* \) indicates feature maps in different modules. \( M \) is the number of filters that are used to decompose the frequency spectrum, and \( N \) is the number of queries with weights to map input feature maps.

Figure 2. Network structure of ResNet encoder in Figure 1. Convolution with stride of 2 is conducted to encode feature map to half the size at the beginning layer of L1-L4.
For feature learning in the frequency domain, we transformed the input image into the frequency domain $F_{img}$ using discrete cosine transform (DCT) (Ahmed, Natarajan, and Rao 1974) and then decomposed the spectrum of $F_{img}$ into different bands to distinguish buildings from other ground objects in the image (shown in Figure 3). Since different ground objects are likely to have different frequency spectra, we manually designed a filter $f$ to decompose the spectrum of the frequency image $F_{img}$ to $M$ bands. A corresponding learnable filter $f_w$ was added to filter $f$ to better distinguish the frequency band of the buildings from the background objects. The operation of the frequency filter on the frequency image $F_{img}$ was conducted following Equation (1). To squeeze the values of $f_w$ within the range of $-1$ to $1$, a function $\sigma$ was adopted [Equation (2)]. Equation (1) shows that the decomposed band $F_{di}$ ($i = 1, \ldots, M$) is calculated by element-based multiplication between the frequency image $F_{img}$ and the summary of the $i$ th frequency filter $f_i$ and $\sigma(f_w)$.

After partitioning the frequency spectrum of $F_{img}$ to different bands, spiking convolution is conducted to enhance the frequency that is more significant for building extraction. According to the theory of signals and systems, convolution with the spiking function in the spatial domain can be transferred to multiplication with a coefficient (Aboozar et al. 2020; Yu et al. 2022). Therefore, the spiking convoluted feature map $F_{ds}$ can be calculated by multiplication with a scaling factor $k$ in Equation (3). In this study, we rescaled each decomposed feature map $F_{di}$ with a scaling factor $k_i$ ($i = 1, \ldots, M$) to calculate the corresponding feature map $F_{ds}$.

$$F_{di} = F_{img} \cdot [f_i + \sigma(f_w)], i = 1, \ldots, M$$ (1)

$$\sigma(x) = \frac{1 - \exp(-x)}{1 + \exp(-x)}$$ (2)

$$F_{ds} = k \times F_{di}$$ (3)

### 2.3 Building prediction component

Using the concatenated feature maps generated in the frequency-spatial feature learning component, a building prediction component was used to predict multi-scale building masks. The gradually up-sampled feature maps in the ResNet decoder, after enhancement by spiking convolution, were fed to each decoder layer of the Transformer, as shown in Figure 1. The structure of ResNet decoder is the inverse of the network structure of ResNet encoder in Figure 2. There are $L$ Transformers in the component, and each Transformer has three decoder layers. The building prediction component has a total of $3L$ decoder layers, and each decoder layer has the network structure shown in Figure 4. Different from the Mask2former in the work of Cheng et al. (Cheng, Schwing, and Kirillov 2021a), the successive up-sampled feature maps of the ResNet decoder $F_{res}$ are transformed to the frequency domain using DCT and enhanced by spiking convolution before being fed to the Transformer [Equation (4)]. The enhanced frequency features $F_{res-sf}$ are used to aid small building extraction and blurred building boundary enhancement with strong sensitivity of high

![Figure 3. Overview of frequency decomposition filter module with spiking convolution. M frequency filters are conducted on the input image $F_{img}$ and generate M feature maps, $F_{di}$. Spiking convolution is conducted on each feature map $F_{di}$ by scaling with a factor $k_i$, $i = 1, \ldots, M$.](image-url)
spectrum frequency. In this paper, the sizes of upsampled feature maps are set thus: \((\frac{W}{32} \times \frac{H}{32}), (\frac{W}{16} \times \frac{H}{16})\), and \((\frac{W}{8} \times \frac{H}{8})\). \(W\) and \(H\) are the width and height of the input image size, respectively.

\[
F_{\text{res-sp}} = \text{DCT}(F_{\text{res}})
\]  

(4)

As shown in Figure 4, each decoder layer of the Transformer takes the modules of the masked attention, self-attention, and feed-forward network (FFN) continuously. The self-attention layer is used to explore relationships between different parts of the image, and the masked attention layer is used to explore the relationship between input and segmented image. The gradually upsampled feature maps after enhanced by spiking convolution. \(F_{\text{res-sp}}\) are firstly linearly transformed using two learnable weights \(W_K\) and \(W_V\) to generate matrices \(K\) and \(V\), respectively. The input query features \(F_q\) are linearly transformed by weight \(W_Q\) to generate matrix \(Q\). To enhance the contextual features of multi-scale buildings, the binary mask \(BM\) from the former Transformer is used to focus on the potential regions with buildings. The binary mask \(BM_i\) of the \(i\) th \((i = 1, \cdots, 3L)\) Transformer decoder layer is generated with a threshold of 0.5 on the building probability prediction image of the former Transformer. The threshold 0.5 is commonly adopted to binarize the prediction image with probability in semantic segmentation (Badrinarayanan, Kendall, and Cipolla 2017; Baheti et al. 2020). To keep pace with the values of matrices \(K\), \(V\), and \(Q\), the binary mask is further transformed by an attention transformation \(f_{at}\) [Equation (5)]. Using matrices \(K\), \(V\), and \(Q\) together with the attention transformed binary mask \(BM_i\), the masked attention and the self-attention modules update the query features via Equations (6) and (7), respectively; softmax is the operation calculated as Equation (8). To generate building prediction image in spatial domain, an inverse DCT is conducted to transform the learned frequency features to the spatial domain for FFN to learn the output query features. Two linear transformation layers make up the module of the FFN, with an active Rectified Linear Unit (ReLU) function [Equation (9)]. Normalization is conducted after each module, that is, Masked module, Self-attention module, and FFN module in Figure 4[Equation (10)]. \(S\) is the summary of the features transferred to the normalization operation; \(\text{Mean}\) and \(\text{STD}\) are the mean and standard variance of \(S\), respectively.

\[
f_{at}(BM_i) = \begin{cases} 0, & BM_i(x, y) = 1 \\ -\infty, & BM_i(x, y) = 0 \end{cases}
\]  

(5)

\[
\text{Masked attention} = \text{softmax} (f_{at}(BM_i) + Q \times K^T) \times V
\]  

(6)

\[
\text{Self-attention} = \text{softmax} \left( \frac{Q \times K^T}{\sqrt{dk}} \right) \times V
\]  

(7)

\[
\text{softmax} = \frac{e^x}{\sum_{i=1}^{N} e^x_i}
\]  

(8)

\[
\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2
\]  

(9)

\[
\text{Norm}(s) = \frac{s - \text{Mean}(s)}{\text{STD}(s)}
\]  

(10)
3. Experiments and results

3.1 Dataset preparation

To evaluate the efficiency and transferability of the proposed model, we used five publicly released building datasets to conduct the experiments. The five building datasets comprised the WHU-Building Satellite Dataset I (Shunping, Wei, and Meng 2019), WHU-Building Satellite Dataset II (Shunping, Wei, and Meng 2019), Massachusetts Buildings Dataset (Mnih 2013), Inria Aerial Image Dataset (Maggiori et al. 2017), and the xBD Building Dataset (Gupta et al. 2019). Instead of training and evaluating using only the sub-datasets of one dataset, as is common in much research (Guo et al. 2022), the five datasets were re-organized to better evaluate the transferability of the proposed model (Table 1).

The Inria Aerial Image Dataset, WHU-Building Satellite Dataset II, and the training dataset of the xBD Building Dataset were used to train the model, as they have large quantities of images and varying illumination conditions from area-to-area with different spatial resolutions from different imaging sensors. All the images were cropped to a size of 512 × 512 pixels, and the images of buildings and non-buildings were labeled as such. The xBD Building Dataset was originally used for building damage assessment triggered by natural disasters. The dataset contains 850,736 buildings with four damage levels. It is a good resource, as it provides real images for practical building extraction, especially useful in emergency rescue. The dataset comprises pre-event and post-event images of different disasters. Since our aim is to extract buildings, pre-event images were used in our training data for model construction. To accord with our aim of building extraction, the labels of all the buildings were re-assigned as “1.”

The Massachusetts Buildings Dataset, WHU-Building Satellite Dataset I, the pre-event images of the hold-out dataset, and the test dataset of the xBD Building Dataset are used as four different evaluation cases for building extraction. In accordance with the training dataset, all the images were cropped to a size of 512 × 512 pixels before being fed to the trained model for evaluation. The hold-out dataset and test dataset from the xBD building dataset are evaluation case III and evaluation case IV, respectively, and were used to assess the proposed model. The main reason of adopting hold-out dataset and test dataset of xBD building dataset for evaluation is that the datasets are officially used to evaluate the generalization performance and the extraction accuracy of the submitted model in the xView2 Challenge, which is officially proposed to evaluate the actual capability in practical building extraction of the technology in artificial intelligence. Since our aim is to evaluate the practical applicability of the proposed methods in building extraction, which accords with the principle of xView2 Challenge, the hold-out dataset and the test dataset were used for evaluation.

3.2 Evaluation metric

To statistically evaluate the building extraction performance of different evaluation datasets, we adopted the widely used parameters of precision, recall, F1-score, intersection over union (IOU), and mean intersection over union (mIOU) in this study (Badrinarayanan, Kendall, and Cipolla 2017; Chen et al. 2021b). Precision represents the percentage of building pixels that are correctly extracted. Recall indicates the percentage of ground truth building pixels that are correctly extracted. F1-score is a balanced indicator of precision and recall, and IOU is the ratio of intersected area to union area of extracted buildings and ground truth buildings in the dataset. mIOU is used to assess the general segmentation performance by calculating the mean IOUs of buildings and background objects in the dataset. The evaluation parameters are calculated according

| Dataset                                | Spatial resolution | Imaging sensor          | Study area                  | Training/evaluating the model       |
|----------------------------------------|--------------------|-------------------------|-----------------------------|------------------------------------|
| Inria Aerial Image Dataset             | 0.3 m              | Aerial Image            | 810 km² of multiple cities  | Training                           |
| WHU-Building Satellite Dataset II      | 0.45 m             | Satellite images        | 860 km² of East Asia        | Training                           |
| WHU-Building Satellite Dataset I       | 1 m                | Aerial Image            | 45,361.79 km² of 15 countries | Training                           |
| Massachusetts Buildings Dataset        | 0.3 m to 2.5 m     | QuickBird, WorldView series, IKONOS, ZY-3 | 340 km² of Boston | Evaluating I                       |
| xBD hold-out dataset                  | 0.3 m              | WorldView-3             | 860 km² of East Asia        | Evaluating III                     |
| xBD test dataset                      | 0.3 m              | WorldView-3             |                             | Evaluating IV                      |
to Equations (11)-(15), wherein TP is short of true positive pixels, FP represents false-positive pixels, and FN is false-negative pixels. The indicator frame per second (FPS) is adopted to evaluate the implementation efficiency of the methods.

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (11)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (12)
\]

\[
F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (13)
\]

\[
\text{mIOU} = \frac{1}{2} \sum_{i=0}^{1} \frac{TP_i}{TP_i + FP_i + FN_i} \quad (15)
\]

\[
\text{IOU} = \frac{TP}{TP + FP + FN} \quad (14)
\]

3.3 Implementation details

The experiments were conducted on the Pytorch platform in the Ubuntu 14.04 environment. Four GeForce RTX™ 3090 GPUs with a memory storage of 24 GB each were used simultaneously to train the proposed model. The model’s convergence was optimal for a batch size of four and an epoch of 11 k. The number of filters in Figure 3 decomposing frequency image \( F_{img} \) is set three after multiple trials, and the frequency spectrum band range of filters \( f \) is set thus: 0–1/50, 1/50–1/10, 1/10–1. In terms of the number of Transformer layers in building prediction module in Figure 1, the number of queries \( N \) is set 100 and the number of Transformer layers \( L \) is set 3. The input query features \( F_q \) in Figure 4 are initialized by zero when training the model. Our proposed frequency learning branch can be generalized to other networks as a learning module. For data augmentation, we adopted the widely used strategies of random cropping, color jittering, horizontal flipping, and scale jittering between 0.5 and 2.0. The Adam learning strategy (Liu et al. 2021) with a poly-learning rate schedule (Chen et al. 2018) was adopted for both the backbone network and Transformer training. The initial learning rate of our proposed model was set to 0.0001, with a weight decay of 0.05, and multiplied by 0.1 through optimization. Our model was trained using the pre-trained model released by Mask2Former (Cheng et al. 2021). As seen from Equations (16–18), the training loss of our proposed MSFTrans \( L_M \) is a combination of binary cross entropy loss \( L_{CE} \) and dice loss \( L_{DL} \), which have been widely used for semantic segmentation (Cheng, Schwing, and Kirillov 2021a; Yi et al. 2019). \( p_{gt}(i) \) is the ground truth label of the pixel indexed by \( i \), and \( p_{result}(i) \) is the segmented result of the pixel indexed by \( i \).

\[
L_M = 0.5 \times L_{CE} + 0.5 \times L_{DL} \quad (16)
\]

\[
L_{CE} = - \sum_{i=0}^{W \times H} \left[ p_{gt}(i) \log(p_{result}(i)) + (1 - p_{gt}(i)) \log(1 - p_{result}(i)) \right] \quad (17)
\]

\[
L_{DL} = 1 - \frac{2 \times \sum_{i=0}^{W \times H} p_{gt}(i)p_{result}(i)}{\sum_{i=0}^{W \times H} p_{gt}^2(i) + \sum_{i=0}^{W \times H} p_{result}^2(i)} \quad (18)
\]

4. Comparisons with the state-of-the-art methods

We compared five state-of-the-art end-to-end methods with our method: DeepLabv3+ (Chen et al. 2018b), HRNet (Sun et al. 2019), PSPNet (Yu, Yang, and Chen 2018; Zhao et al. 2017), BiSeNet (Fan et al. 2021), and SegNet (Badrinarayanan, Kendall, and Cipolla 2017), all of which have achieved state-of-the-art performance in building extraction. For a fair comparison, the selected methods were implemented with publicly released code and the training methods in this study.

The evaluation statistics of our proposed network MSFTrans and the five comparison methods on each of the four evaluation datasets (Massachusetts Buildings Dataset, WHU-Building Satellite Dataset I, the hold-out dataset, and the test dataset of the xBD building dataset) are listed in Tables 2–5, respectively. Our proposed MSFTrans produced a remarkable result for F1, IOU, and mIOU compared with the five methods. For evaluation dataset I (Table 1), MSFTrans shows a 15.16% improvement in F1, 12.67% improvement in IOU, and an 11.76% improvement in mIOU. For evaluation dataset II (Table 1), MSFTrans shows a 17.09% improvement in F1, 15.86% improvement in IOU, and an 8.96% improvement in mIOU. Higher improvements are seen in evaluation datasets III and IV (Table 1): 29.61% in F1, 36.60% in IOU, and 19.66%
in mIOU for both datasets using MSFTrans. The significant improvements seen in our proposed MSFTrans method, when compared with the five end-to-end networks, demonstrate its advanced capability in building extraction for practical scenarios.

The performances of building extraction in evaluation datasets I and II are generally lower than those of the other two datasets. One possible reason for this is the large quantity of small building instances in the dataset with low spatial resolution compared with that of the training datasets. Buildings in images with lower spatial resolution are likely to have a smaller size, which generates challenges in evaluating the generalization of the proposed model for small building extraction. Faced with smaller buildings, our proposed MSFTrans still delivers the best performance, with an improvement in recall of 53.55%. This validates the high efficiency of our proposed frequency-spatial based Transformer in maintaining and capturing multi-scale building features by learning global frequency spectrum feature in frequency domain and local multi-scale feature in spatial domain. Compared with local convolutions in spatial domain as conducted in other methods, where small buildings are easy to be filtered out, our frequency-spatial based Transformer is more effective in extracting buildings, regardless of the variation in building size. The implementation efficiency of our proposed method is lower than the other methods with the structure of multi-task and time-consuming frequency-spatial domain transformation. It can be recognized as acceptable with an FPS of 10.24, about 2 frames/second smaller than HRNet. Notably, the evaluation statistics in this paper are generally lower than those of published results (Guo et al. 2022) because of the different evaluation strategy. Our aim is to objectively evaluate the transferability and the practical applicability of the proposed model

| Table 2. Evaluation statistics of evaluation dataset I (%) |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Precision | Recall | F1 | IOU | mIOU | FPS |
| DeepLabv3+ | 15.01 | 34.91 | 20.99 | 11.73 | 48.00 | 45.53 |
| HRNet | 20.16 | 45.73 | 27.98 | 16.27 | 50.86 | 12.72 |
| PSPNet | **43.35** | 33.15 | 37.57 | 23.13 | 51.13 | 53.33 |
| SegNet | 31.46 | 20.87 | 25.09 | 14.35 | 43.79 | 119.14 |
| BiSeNet | 20.00 | 44.46 | 27.58 | 16.00 | 50.64 | 87.21 |
| MSFTrans | 40.83 | **74.42** | **52.73** | **35.80** | **62.62** | **10.51** |

| Table 3. Evaluation statistics of evaluation dataset II (%) |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Precision | Recall | F1 | IOU | mIOU | FPS |
| DeepLabv3 | **71.76** | 27.87 | 40.15 | 25.12 | 55.73 | 44.07 |
| HRNet | 40.16 | 40.77 | 40.46 | 25.36 | 58.98 | 12.16 |
| PSPNet | 64.09 | 26.35 | 37.34 | 22.96 | 54.65 | 53.32 |
| SegNet | 52.75 | 18.06 | 26.90 | 15.54 | 48.73 | 122.97 |
| BiSeNet | 69.85 | 25.11 | 36.94 | 22.66 | 53.73 | 89.21 |
| MSFTrans | 59.05 | **57.7** | **58.36** | **41.21** | **67.94** | **10.24** |

| Table 4. Evaluation statistics of evaluation dataset III (%) |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Precision | Recall | F1 | IOU | mIOU | FPS |
| DeepLabv3 | 32.57 | 78.82 | 46.09 | 29.95 | 62.68 | 44.35 |
| HRNet | 43.98 | 56.60 | 49.50 | 32.89 | 63.72 | 13.34 |
| PSPNet | 57.99 | 23.91 | 33.85 | 20.38 | 53.24 | 51.08 |
| SegNet | 38.04 | 13.06 | 19.44 | 10.77 | 45.84 | 125.19 |
| BiSeNet | 45.92 | 73.14 | 56.41 | 39.29 | 67.49 | 91.21 |
| MSFTrans | **85.53** | **88.07** | **86.78** | **76.65** | **87.51** | **10.36** |

| Table 5. Evaluation statistics of evaluation dataset IV (%) |
|---------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Precision | Recall | F1 | IOU | mIOU | FPS |
| DeepLabv3+ | 33.69 | 81.13 | 47.61 | 31.24 | 63.30 | 44.68 |
| HRNet | 45.19 | 59.00 | 51.17 | 34.39 | 64.47 | 12.89 |
| PSPNet | 58.32 | 25.05 | 35.04 | 21.25 | 53.74 | 51.68 |
| SegNet | 37.98 | 13.57 | 19.99 | 11.11 | 45.99 | 123.01 |
| BiSeNet | 46.50 | 74.08 | 57.13 | 39.99 | 67.79 | 88.65 |
| MSFTrans | **84.85** | **88.73** | **86.74** | **76.59** | **87.45** | **10.38** |
in building extraction by mimicking practical cases. The evaluation images are different from the training dataset in imaging sensors, spatial resolutions and study areas. With different spatial distribution patterns of ground objects in different sizes, the evaluation statistics are not as high as those in published works (Chen et al. 2022) that evaluate on images with similar spatial distribution patterns as the training dataset.

In spite of statistical evaluations above, a visual comparison of the methods on the four evaluation datasets is conducted. Five random images with the ground truth labeling images of each evaluation dataset were selected. The extracted results of the different methods from different evaluation dataset are shown in Figures 5–8, respectively. In accordance with the evaluated statistical comparisons, the building extraction performance of our proposed MSFTrans is remarkably better than that of the other methods, especially for the small building instances in Figure 5. BiSeNet is less effective when being transferred to extract small buildings [Figure 5(b)], but better in extracting large buildings. DeepLabv3+, HRNet, and PSPNet are slightly better than BiSeNet at distinguishing small building instances, but less sensitive to large buildings [Figures 5(c–e)]. As shown in Figure 5(f), SegNet can extract building areas roughly, but without a detailed boundary distribution pattern. Compared with the five methods, our proposed MSFTrans method is easily transferred to extract both small and large building instances with fine boundary details [Figure 5(g)]. The issue of small building omission may be overcome by enlarging the variety of small building instances, specifically when training the building extraction model.

For evaluation dataset II, whose images have various spatial resolutions, MSFTrans is able to extract most buildings correctly and neatly, in spite of the building omission in the background composed of bare land with similar spectral intensity of buildings in Figure 6(g). As shown in Figure 6(b), BiSeNet is sensitive to large buildings, but introduces many falsely extracted buildings, which are actually vegetation. The building extraction results of DeepLabv3+ are better than those of HRNet, PSPNet, and SegNet, with less omitted building instances and clearer building boundaries; however, background objects near the building boundaries are mis-extracted as buildings, as shown in Figure 6(c–f). In different scenarios, MSFTrans can better distinguish buildings from background objects [Figure 6(g)].

For datasets III and IV, most buildings are well extracted by MSFTrans, with intact boundaries, despite the cloud disturbance in Figure 7. For buildings with blurred boundaries in Figures 7 and 8, all the five methods in comparison fail to

Figure 5. Visual comparison of extracted building instances using the different methods of evaluation dataset I: (a) original images; (b) results from BiSeNet; (c) results from DeepLabv3+; (d) results from HRNet; (e) results from PSPNet; (f) results from SegNet; (g) results from MSFTrans, TP indicates true positive pixels, FN and FP represent false-negative and false-positive pixels, respectively.
capture detailed building boundaries and suffer from building omission [Figures 7(b–f) and 8(b–f)]. Moreover, there are many bare ground pixels misclassified as buildings by SegNet and PSPNet [Figure 7(e–f)]. The results indicate that the frequency-spatial learning mechanism in MSFTrans better captures the features of building instances from background objects in different circumstances [Figures 7(g) and 8]. In images without building instances [Figure 7(a)], the bare ground is easily committed as buildings by each of the methods. To confront such issue, the network structure may be modified to focus more on textural features in the future study.

Figure 6. Visual comparison of extracted building instances using the different methods of evaluation dataset II: (a) original images; (b) results from BiSeNet; (c) results from DeepLabv3+; (d) results from HRNet; (e) results from PSPNet; (f) results from SegNet; (g) results from MSFTrans, TP indicates true positive pixels, FN and FP represent false-negative and false-positive pixels, respectively.

Figure 7. Visual comparison of extracted building instances using the different methods of evaluation dataset III: (a) original images; (b) results from BiSeNet; (c) results from DeepLabv3+; (d) results from HRNet; (e) results from PSPNet; (f) results from SegNet; (g) results from MSFTrans. TP indicates true positive pixels, FN and FP represent false-negative and false-positive pixels, respectively.
5. Ablation studies and discussions

5.1 Comparison of frequency learning with spatial learning

Using the multi-task framework of joint spatial-frequency feature learning, the features can be better captured for building extraction. In order to explore the effectiveness of each task branch in feature learning for building extraction, we conducted a comparison of the building extraction performance between single-task learning model and our proposed MSFTrans. The single-task learning model is constructed by, respectively, removing the frequency feature learning branch (MSFTrans-FL) and spatial feature learning branch (MSFTrans-SL) (shown in Figure 1). MSFTrans-FL and MSFTrans-SL are trained and evaluated with the same strategy. The evaluation statistics of MSFTrans-FL and MSFTrans-SL for each evaluation dataset were calculated and listed in Table 6. When transferred to evaluation datasets I and II, whose spatial resolution varies considerably from that of the training dataset, the impact of spatial learning branch in MSFTrans-SL is more significant than that of the frequency domain in MSFTrans-FL with higher mIOU decrease by 3.02% and 1.25% compared with MSFTrans. The impact of spatial learning and frequency learning branches are similar with a mIOU decrease of 2% when images have higher spatial resolutions, as cases of datasets III and IV. It indicates that the multi-scale features captured by spatial learning branch are more significant for building extraction from images with different spatial resolutions.

![Figure 8. Visual comparison of extracted building instances using the different methods of evaluation dataset IV.](image)

**Table 6.** Evaluation statistics of MSFTrans-FL and MSFTrans-SL for each evaluation dataset (%).

| Method     | Precision | Recall | F1   | IOU  | mIOU |
|------------|-----------|--------|------|------|------|
| I          |           |        |      |      |      |
| MSFTrans-FL| 44.57     | 58.32  | 50.52| 33.8 | 60.57|
| MSFTrans-SL| 40.98     | 58.85  | 48.31| 31.85| 59.6 |
| MSFTrans   | 40.83     | 74.42  | 52.73| 35.8 | 62.62|
| II         |           |        |      |      |      |
| MSFTrans-FL| 65.63     | 51.6   | 57.77| 40.62| 67.26|
| MSFTrans-SL| 58.46     | 54.21  | 56.25| 39.14| 66.69|
| MSFTrans   | 59.05     | 57.7   | 58.36| 41.21| 67.94|
| III        |           |        |      |      |      |
| MSFTrans-FL| 82.31     | 85.64  | 83.94| 72.33| 85.18|
| MSFTrans-SL| 84.94     | 83.83  | 84.38| 72.98| 85.51|
| MSFTrans   | 85.53     | 88.07  | 86.78| 76.65| 87.51|
| IV         |           |        |      |      |      |
| MSFTrans-FL| 81.75     | 86.14  | 83.88| 72.24| 85.11|
| MSFTrans-SL| 83.87     | 84.74  | 84.3 | 72.87| 85.42|
| MSFTrans   | 84.85     | 88.73  | 86.74| 76.59| 87.45|
Frequency learning branch in MSFTrans has negative impact on building extraction precision with 3.74% increase for datasets of I and II after being removed in MSFTrans-FL. That indicates frequency features are slightly sensitive to the change of spatial resolution in falsely extracting background objects as buildings. However, frequency learning branch improves recall remarkably with similar progress of spatial learning branch of 16.1%. That demonstrates the significant role of frequency features and multi-scale spatial features of our proposed MSFTrans in extracting buildings from images with different spatial resolutions, which can strengthen the transferability of our proposed MSFTrans. In terms of datasets III and IV, whose spatial resolution is similar with that of training datasets, frequency learning branch gets higher progress in improving precision than spatial learning branch, while vice versa in improving recall. Both learning branches are equally significant in enhancing the reliability and transferability of proposed MSFTrans for building extraction in different study areas from different imaging sensors.

5.2 Impact of frequency decomposition learning module

The frequency decomposition module was used in our proposed MSFTrans method to enhance the frequency spectrum band of buildings from high-resolution images. In order to evaluate the effectiveness of the frequency decomposition learning module, we modified the proposed network by removing the frequency decomposition module (MSFTrans-FD), as demonstrated in Figure 9. Moreover, to seek for the optimum number of filters in the frequency decomposition module, we conducted two more experiments, in which we set M as 1 and 5. The same training and evaluation strategy was adopted on the modified network. The corresponding evaluation statistics for each dataset are shown in Table 7.

It is worth noting that the frequency decomposition module significantly enhances the proposed model in capturing the features of small buildings, with at least 13.82% decrease in precision of MSFTrans-FD in datasets I and II (Table 7), whose image spatial resolution is comparatively lower and the building instances are generally smaller than the other evaluation datasets. Moreover, the frequency decomposition module improves mIoU most remarkably by 7.16% in dataset II, whose image spatial resolution varies considerably. It has shown strong transferability in extracting buildings by decomposing the frequency spectrum to different bands. As can be seen from Table 7, the decomposition of the frequency spectrum to three filters (as set in MSFTrans) achieves the best performance in building extraction, with the largest F1, IOU, and mIoU for each of the datasets.

5.3 Impact of spiking convolution

Spiking convolution in this study is used to enhance the frequency feature by multiplication with different factors for building extraction. The role of spiking convolution is evaluated by removing it (MSFTrans-SC) from our proposed network and the corresponding evaluation statistics for each evaluation dataset are demonstrated in Table 8. The impact of spiking
convolution is significant in enhancing the proposed method in capturing building instances, especially in dataset I with mIOU improvement of 3.26%. Spiking convolution is generally effective in enhancing the feature capturability of building instances in images with different spatial resolutions. It is worth to mention that spiking convolution has a negative impact on the precision of building extraction by falsely recognizing background objects as buildings in images with different spatial resolution from the training dataset (dataset I and II), but it improves recall remarkably with the high sensitivity to intensity change in frequency. In terms of dataset III and IV, spiking convolution enhances the model transferability in both precision and recall, making the proposed MSFTrans highly potential for practical applications.

5.4 Impact of building size

In order to explore the impact of building size on the extraction performance of our proposed method, we categorized the building instances of each evaluation dataset into four classes according to the number of pixels in each building instance. Detailed size categorization thresholds for each evaluation dataset are shown in Table 9. The thresholds are manually set according to the number distribution of building instances for each size category. Since evaluation dataset I is concentrated with small building instances, the size thresholds of the building instances are manually set, with the number of pixels ranging from 50 to 500. The thresholds in categorizing buildings of evaluation datasets II–IV are set with the same values (Table 9). The evaluation statistics of extracted building instances using our proposed MSFTrans method for each size category of each evaluation dataset are calculated and shown in the chart of Figure 10.

Clearly, the extraction performance gets better increasing size of building instances for each of the evaluation datasets. For dataset I, the precision, F1, IOU, and mIOU increase significantly when extracting buildings of size category 4 (pixel size

---

**Table 7.** Evaluation statistics of different frequency decomposition filters for each evaluation dataset (%), MSFTrans is represented by MSFTrans-3FD to coincide with other methods.

| Method | Precision | Recall | F1    | IOU  | mIOU |
|--------|-----------|--------|-------|------|------|
| I      | MSFTrans-FD | 27.01  | 68.78 | 38.79 | 24.06 | 56   |
|        | MSFTrans-1FD | 33.38  | 67.38 | 45.01 | 29.05 | 58.62|
|        | MSFTrans-3FD | 40.83  | 74.42 | 52.73 | 35.8  | 62.62|
|        | MSFTrans-5FD | 33.75  | 62.31 | 43.78 | 28.73 | 57.82|
| II     | MSFTrans-FD | 31.29  | 65.46 | 42.34 | 26.86 | 60.78|
|        | MSFTrans-1FD | 37.61  | 55.59 | 44.86 | 28.92 | 61.57|
|        | MSFTrans-3FD | 59.05  | 57.7  | 58.36 | 41.21 | 67.94|
|        | MSFTrans-5FD | 55.04  | 55.89 | 55.46 | 38.37 | 66.39|
| III    | MSFTrans-FD | 76.33  | 91.24 | 83.12 | 71.12 | 84.6 |
|        | MSFTrans-1FD | 81.33  | 88.46 | 84.74 | 73.53 | 85.85|
|        | MSFTrans-3FD | 85.53  | 88.07 | 86.78 | 76.65 | 87.51|
|        | MSFTrans-5FD | 85.98  | 84.64 | 85.3  | 74.37 | 86.26|
| IV     | MSFTrans-FD | 75.68  | 91.89 | 82.99 | 70.94 | 84.47|
|        | MSFTrans-1FD | 80.66  | 89    | 84.62 | 73.35 | 85.73|
|        | MSFTrans-3FD | 84.85  | 88.73 | 86.74 | 76.59 | 87.45|
|        | MSFTrans-5FD | 84.87  | 85.28 | 85.07 | 74.03 | 86.05|

**Table 8.** Evaluation statistics of spiking convolution for each evaluation dataset (%).

| Method | Precision | Recall | F1    | IOU  | mIOU |
|--------|-----------|--------|-------|------|------|
| I      | MSFTrans-SC | 44.96  | 53.6  | 48.89 | 32.36 | 59.36|
|        | MSFTrans  | 40.83  | 74.42 | 52.73 | 35.8  | 62.62|
|        | MSFTrans-SC | 63.47  | 53.1  | 58.64 | 41.49 | 67.9 |
|        | MSFTrans  | 59.05  | 57.7  | 58.36 | 41.21 | 67.94|
|        | MSFTrans-SC | 83.14  | 87.04 | 85.04 | 73.98 | 86.08|
|        | MSFTrans  | 85.53  | 88.07 | 86.78 | 76.65 | 87.51|
|        | MSFTrans-SC | 82.46  | 87.64 | 84.97 | 73.87 | 85.99|
|        | MSFTrans  | 84.85  | 88.73 | 86.74 | 76.59 | 87.45|

**Table 9.** Size categorization thresholds of evaluation datasets.

| Evaluation Dataset | 1 | 2 | 3 | 4 |
|--------------------|---|---|---|---|
| I                  | <50 | ≥50 < 100 | ≥100 < 500 | ≥500 |
| II–IV              | <500 | ≥500 < 1000 | ≥1000 < 2000 | ≥2000 |
≥500). For datasets II–IV, the evaluation statistics of building instances with size category 1 (pixel size <500) are significantly lower than those of the other size categories. The precision of building extraction, especially for small buildings, is a key factor that needs improvement to better lend itself to practical scenarios. Generally, with the evaluation strategy proposed in this study in mimicking practical application cases with remote sensing images of different study areas from different imaging sensors, our proposed MSFTrans method can be directly applied to extract buildings from different remote sensing images with different spatial resolutions given the satisfactory evaluation performances.

6. Conclusions

Inconsistent spectral characteristics within large building instances, omission of small buildings, blurred building boundaries, and complex background objects are the main impediments to large-scale practical application. To surmount this problem, we proposed a multi-task frequency-spatial learning Transformer for building extraction from high spatial resolution images. Frequency decomposition features from high-resolution images were used to supplement multi-scale feature maps to preserve detailed ground object information and distinguish blurred buildings by capturing different frequencies. Spiking convolution was proposed to enhance the frequency characteristics of various buildings. Extracted frequency-spatial features were fed to a masked-attention Transformer by synthesizing continuous up-sampled feature maps for accurate multi-scale building mask prediction.

Experiments on five publicly released building datasets demonstrated the high transferability of our proposed MSFTrans method in extracting buildings from different high spatial resolution images. Using the five state-of-the-art methods as a baseline, our proposed method performed remarkably. When transferred to extract buildings from dataset concentrated with small buildings, MSFTrans obtained approximately 53.55% progress in recall. Moreover, 29.61% F1, 36.60% IOU, and 19.66% mIOU were achieved by the MSFTrans method when applied to extract buildings from images with high spatial resolution (0.3 m); and 15.16% F1, 12.67% IOU, and 11.76% mIOU were achieved when applied to extract buildings from images with low spatial resolution (1 m). The significant improvements found when using our proposed MSFTrans method demonstrated the advanced capability of the spatial-frequency learning mechanism in building extraction in practical scenarios. The size of building instance significantly impacted extraction performance. The accuracy of

Figure 10. Evaluation statistics for extracted building instances with different size categories for each evaluation dataset: (a) evaluation dataset I; (b) evaluation dataset II; (c) evaluation dataset III; (d) evaluation dataset IV.
large building extractions was higher than that of small ones. Improving building instance using a number of pixels smaller than 500 should considerably improve general building extraction performance in future studies.

Based on the intact and multi-scale buildings extracted using our proposed MSFTTrans method, in future work, we will focus on improving the extraction performance of small building instances by enlarging the variety of training samples and creating more effective network structures to synthesize textural characteristics.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19030101), the Youth Innovation Promotion Association, CAS (2022122), the China-ASEAN Big Earth Data Platform and Applications (CADA, guikeAA20302022), and the High-Resolution Earth Observation System Project (20-Y30F10-9001-20/22).

**Authors’ contributions**

Bo Yu proposed the network and wrote the manuscript. Lei Wang designed the experiment. Fang Chen, Haiping Yang and Ning Wang polished the manuscript.

**Code availability**

The implementation code can be referred to in https://github.com/yubozuzu123/MSFTTrans.

**Data availability statement**

The data that support the findings of this study are openly available in https://lear.inrialpes.fr/~jegou/data.php, http://gpcv.whu.edu.cn/data/, and https://www.Kaggle.com/bradj98/massachusetts-buildings-dataset.

**References**

Abdollahi, A., B. Pradhan, and A. M. Alamri. 2020. “An Ensemble Architecture of Deep Convolutional Segnet and Unet Networks for Building Semantic Segmentation from high-resolution Aerial Images.” *Geocarto International* 37 (12): 3355–3370. doi:10.1080/15481603.2020.1856199.

Aboozar, T., A. Belatreche, L. Yuhua, G. Cosma, L. P. Maguire, and T. M. McGinnity. 2020. “A Review of Learning in Biologically Plausible Spiking Neural Networks.” *Neural Networks* 122: 253–272. doi:10.1016/j.neunet.2019.09.036.

Ahmed, N., T. Natarajan, and K. R. Rao. 1974. “Discrete Cosine Transform.” *IEEE Transactions on Computers* 100 (1): 90–93. doi:10.1109/T-C.1974.223784.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Baheti, B., S. Innani, S. Gajre, and S. Talbar. 2020. “Semantic Scene Segmentation in Unstructured Environment with Modified DeepLabV3+.” *Pattern Recognition Letters* 138: 223–229. doi:10.1016/j.patrec.2020.07.029.

Benjamin, S., M. Laverdiere, H. Lexie Yang, and A. Rose. 2022. “Iterative self-organizing SCene-LEVEL Sampling (ISOSCELES) for large-scale Building Extraction.” *GIScience & Remote Sensing* 59 (1): 1–16. doi:10.1080/15481603.2021.2006433.

Blaschke, T. 2010. “Object Based Image Analysis for Remote Sensing.” *ISPRS Journal of Photogrammetry and Remote Sensing* 65 (1): 2–16. doi:10.1016/j.isprsjprs.2009.06.004.

Cao, S., Q. Weng, M. Du, B. Li, R. Zhong, and Y. Mo. 2020. “Multi-scale three-dimensional Detection of Urban Buildings Using Aerial LiDAR Data.” *GIScience & Remote Sensing* 57 (8): 1125–1143. doi:10.1080/15481603.2020.1847453.

Cheng, B., M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen. 2020. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation. Paper presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 13-19 June 2020.

Cheng, B., I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. 2021b. “Masked-attention Mask Transformer for Universal Image Segmentation.” *arXiv e-prints.* doi:10.48550/arXiv.2112.01527.

Cheng, B., A. Schwing, and A. Kirillov. 2021a. “Per-pixel Classification Is Not All You Need for Semantic Segmentation.” *arXiv* 2107.06278. doi:10.48550/arXiv.2107.06278.

Chen, J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. 2021b. Transunet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv preprint arXiv:2102.04306* Medical image analysis 75.10.1016/j.medima.2021.102287.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2014. “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.” *arXiv* 1412.7062. doi:10.48550/arXiv.1412.7062.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–848. doi:10.1109/TPAMI.2017.2699184.
Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam. 2017. “Rethinking Atrous Convolution for Semantic Image Segmentation.” arXiv:1706.05587. doi:10.48550/arXiv.1706.05587.

Chen, F., N. Wang, B. Yu, and L. Wang. 2022. “Res2-Unet, a New Deep Architecture for Building Detection from High Spatial Resolution Images.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15: 1494–1501. doi:10.1109/JSTARS.2022.3146430.

Chen, F., M. Zhang, H. Guo, S. Allen, J. S. Kargel, U. K. Haritassha, and S. C. Watson. 2021a. “Annual 30 m Dataset for Glacial Lakes in High Mountain Asia from 2008 to 2017.” Earth System Science Data 13 (2): 741–766. doi:10.5194/essd-13-741-2021.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018a. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Paper presented at the European Conference on Computer Vision, Munich, Germany, 06 October 2018.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018b. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. Paper presented at the Proceedings of the European conference on computer vision (ECCV), September 8-14, 2018, Munich, Germany.

Deng, L., Y.-N. Yan, Y. He, Z.-H. Mao, and J. Yu. 2019. “An Improved Building Detection Approach Using L-band POLSAR two-dimensional time-frequency Decomposition over Oriented built-up Areas.” GIScience & Remote Sensing 56 (1): 1–21. doi:10.1080/15481603.2018.1484409.

Diakogiannis, F. I., F. Waldner, P. Caccetta, and C. Wu. 2020. “ResUNet-A: A Deep Learning Framework for Semantic Segmentation of Remotely Sensed Data.” ISPRS Journal of Photogrammetry and Remote Sensing 162: 94–114. doi:10.1016/j.isprsjprs.2020.01.013.

Du, S., F. Zhang, and X. Zhang. 2015. “Semantic Classification of Urban Buildings Combining VHR Image and GIS Data: An Improved Random Forest Approach.” ISPRS Journal of Photogrammetry and Remote Sensing 105: 107–119. doi:10.1016/j.isprsjprs.2015.03.011.

Fan, M., S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei. 2021. Rethinking BiSeNet for Real-time Semantic Segmentation. Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20-25 June 2021.

Fu, J., J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. 2019. Dual Attention Network for Scene Segmentation. Paper presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15-20 June 2019.

Ghosh-Dastidar, S., and H. Adeli. 2009. “Spiking Neural Networks.” International Journal of Neural Systems 19 (4): 295–308. doi:10.1142/S0129065709002002.

Guangming, W., X. Shao, Z. Guo, Q. Chen, W. Yuan, X. Shi, X. Yongwei, and R. Shibasaki. 2018. “Automatic Building Segmentation of Aerial Imagery Using Multi-Constraint Fully Convolutional Networks.” Remote Sensing 10 (3). doi:10.3390/rs10030407.

Guo, H., B. Du, L. Zhang, and X. Su. 2022. “A coarse-to-fine Boundary Refinement Network for Building Footprint Extraction from Remote Sensing Imagery.” ISPRS Journal of Photogrammetry and Remote Sensing 183: 240–252. doi:10.1016/j.isprsjprs.2021.11.005.

Guo, H., Q. Shi, B. Du, L. Zhang, D. Wang, and H. Ding. 2021a. “Scene-Driven Multitask Parallel Attention Network for Building Extraction in High-Resolution Remote Sensing Images.” IEEE Transactions on Geoscience and Remote Sensing 59 (5): 4287–4306. doi:10.1109/TGRS.2020.3014312.

Guo, H., Q. Shi, A. Marinoni, B. Du, and L. Zhang. 2021b. “Deep Building Footprint Update Network: A semi-supervised Method for Updating Existing Building Footprint from bi-temporal Remote Sensing Images.” Remote Sensing of Environment 264: 112589. doi:10.1016/j.rse.2021.112589.

Gupta, R., R. Hosfelt, S. Sajeev, N. Patel, B. Goodman, J. Doshi, E. Heim, H. Choset, and M. Gaston. 2019. Xbd: A Dataset for Assessing Building Damage from Satellite Imagery. arXiv preprint arXiv:1911.09296.

Hanqui, X. 2008. “A New Index for Delineating Built-up Land Features in Satellite Imagery.” International Journal of Remote Sensing 29 (14): 4269–4276. doi:10.1080/01431160802039957.

He, S., and W. Jiang. 2021. “Boundary-Assisted Learning for Building Extraction from Optical Remote Sensing Imagery.” Remote Sensing 13 (4): 760. doi:10.3390/rs13040760.

Huang, J., X. Zhang, Q. Xin, Y. Sun, and P. Zhang. 2019. “Automatic Building Extraction from high-resolution Aerial Images and LiDAR Data Using Gated Residual Refinement Network.” ISPRS Journal of Photogrammetry and Remote Sensing 151 (MAY): 91–105. doi:10.1016/j.isprsjprs.2019.02.019.

Huynh, C., A. T. Tran, K. Luu, and M. Hoai. 2021. Progressive Semantic Segmentation. Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20-25 June 2021.

Janalipour, M., and A. Mohammadzadeh. 2016. “Building Damage Detection Using Object-Based Image Analysis and ANFIS from High-Resolution Image (Case Study: BAM Earthquake, Iran).” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (5): 1937–1945. doi:10.1109/JSTARS.2015.2458582.

Jia, H., F. Chen, C. Zhang, J. Dong, E. Du, and L. Wang. 2022. “High Emissions Could Increase the Future Risk of maize Drought in China by 60–70 %.” Science of the Total Environment 852: 158474. doi:10.1016/j.scitotenv.2022.158474.

Jiang, N., J. Zhang, H. Li, and X. G. Lin. 2008. Semi-automatic Building Extraction from High Resolution Imagery Based on Segmentation. Paper presented at the 2008 International Workshop on Earth Observation and Remote Sensing Applications, Beijing, China, 30 June-2 June-2 July 2008.

Kaiming, H., X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. Paper presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27-30 June 2016.
Liao, C., H. Han, L. Haifeng, G. Xuming, M. Chen, L. Chuangnong, and Q. Zhu. 2021. “Joint Learning of Contour and Structure for Boundary-Preserved Building Extraction.” Remote Sensing 13 (6): 1049. doi:10.3390/rs13061049.

Li, J., X. Huang, L. Tu, T. Zhang, and L. Wang. 2022. “A Review of Building Detection from Very High Resolution Optical Remote Sensing Images.” GIScience & Remote Sensing 59 (1): 1199–1225. doi:10.1080/15481603.2022.2101727.

Liu, Z., Z. Shen, L. Shichao, K. Helwegen, D. Huang, and K.-T. Cheng. 2021. How Do Adam and Training Strategies Help Bnns Optimization. Paper presented at the International Conference on Machine Learning, Vienna, Austria.

Long, J., E. Shelhamer, and T. Darrell. 2015. Fully Convolutional Networks for Semantic Segmentation. Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7-12 June 2015.

Luo, L., L. Pengpeng, and X. Yan. 2021. “Deep Learning-Based Building Extraction from Remote Sensing Images: A Comprehensive Review.” Energies 14 (23): 7982. doi:10.3390/en14237982.

Maggiori, E., Y. Tarabalka, G. Charpiat, and P. Alliez. 2017. Can Semantic Labeling Methods Generalize to Any City? The Inria Aerial Image Labeling Benchmark. Paper presented at the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23-28 July 2017.

Mnih, V. 2013. Machine Learning for Aerial Image Labeling. University of Toronto (Canada) ProQuest Dissertations Publishing.

Mohit, S., S. K. Sahay, and H. Rathore. 2018. Comparison of Deep Learning and the Classical Machine Learning Algorithm for the Malware Detection. 2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), June 27-29, 2018, Busan, Korea (South).

Qian, Y., G. Yin, L. Sheng, Z. Chen, and J. Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining frequency-aware Clues. Paper presented at the European Conference on Computer Vision, Glasgow.

Ronneberger, O., P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. Paper presented at the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Cham, 2015.

Schiewe, J. 2002. “Segmentation of high-resolution Remotely Sensed data-contents, applications and Problems.” International Archives of Photogrammetry & Remote Sensing 34.

Schlosser, A. D., G. Szabó, L. Bertalan, Z. Varga, P. Enyedi, and S. Szabó. 2020. “Building Extraction Using Orthophotos and Dense Point Cloud Derived from Visual Band Aerial Imagery Based on Machine Learning and Segmentation.” Remote Sensing 12 (15): 2397. doi:10.3390/rs12152397.

Shao, Z., P. Tang, Z. Wang, N. Saleem, S. Yam, and C. Sommai. 2020. “BRRNet: A Fully Convolutional Neural Network for Automatic Building Extraction from High-Resolution Remote Sensing Images.” Remote Sensing 12 (6): 1050. doi:10.3390/rs12061050.

Sherrah, J. 2016. “Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery.” arXiv:1606.02585. doi:10.1109/TGRS.2018.2850972.

Shibiao, X., X. Pan, E. Li, W. Baoyuan, B. Shuhui, W. Dong, S. Xiang, and X. Zhang. 2018. “Automatic Building Rooftop Extraction from Aerial Images via Hierarchical RGB-D Priors.” IEEE Transactions on Geoscience and Remote Sensing 56 (12): 7369–7387. doi:10.1109/TGRS.2018.2850972.

Shunping, J., S. Wei, and L. Meng. 2019. “Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set.” IEEE Transactions on Geoscience and Remote Sensing 57 (1): 574–586. doi:10.1109/TGRS.2018.2858817.

Sun, Y., Y. Hua, L. Mou, and X. Xiang Zhu. 2022. “CG-Net: Conditional GIS-Aware Network for Individual Building Segmentation in VHR SAR Images.” IEEE Transactions on Geoscience and Remote Sensing 60: 1–15. doi:10.1109/TGRS.2020.3043089.

Sun, K., B. Xiao, D. Liu, and J. Wang. 2019. “Deep High-Resolution Representation Learning for Human Pose Estimation.” arXiv e-prints: arXiv:1902. 09212. doi:10.48550/arXiv.1902.09212.

Turker, M., and D. Koc-San. 2015. “Building Extraction from high-resolution Optical Spaceborne Images Using the Integration of Support Vector Machine (SVM) Classification, Hough Transformation and Perceptual Grouping.” International Journal of Applied Earth Observation and Geoinformation 34: 58–69. doi:10.1016/j.jag.2014.06.016.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention Is All You Need. Paper presented at the arXiv. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a854aa-Abstract.html

Wang, H., and F. Miao. 2022. “Building Extraction from Remote Sensing Images Using Deep Residual U-Net.” European Journal of Remote Sensing 55 (1): 71–85. doi:10.1080/22797524.2021.2018944.

Wang, Z., X. Nan, B. Wang, Y. Liu, and S. Zhang. 2022. “Urban Building Extraction from high-resolution Remote Sensing Imagery Based on multi-scale Recurrent Conditional Generative Adversarial Network.” GIScience & Remote Sensing 59 (1): 861–884. doi:10.1080/15481603.2022.2076382.

Xiaohuan, Y., X. Wan, and C. Wang. 2016. Building Boundaries Extraction from Points Cloud Using an Image Edge Detection Method. Paper presented at the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 10-15 July 2016.

Xie, Y., A. Weng, and Q. Weng. 2015. “Population Estimation of Urban Residential Communities Using Remotely Sensed Morphologic Data.” IEEE Geoscience and Remote Sensing Letters 12 (5): 1111–1115. doi:10.1109/LGRS.2014.2385597.

Yang, H., X. Meixia, Y. Chen, W. Wei, and W. Dong. 2022. “A Postprocessing Method Based on Regions and Boundaries Using Convolutional Neural Networks and
A New Dataset for Building Extraction.“ Remote Sensing 14 (3): 647. doi:10.3390/rs14030647.

Yi, Y., Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao. 2019. “Semantic Segmentation of Urban Buildings from VHR Remote Sensing Imagery Using a Deep Convolutional Neural Network.” Remote Sensing 11 (15): 1774. doi:10.3390/rs11151774.

Yu, B., F. Chen, and X. Chong. 2020. “Landslide Detection Based on contour-based Deep Learning Framework in case of National Scale of Nepal in 2015.” Computers & Geosciences 135: 104388. doi:10.1016/j.cageo.2019.104388.

Yu, B., L. Yang, and F. Chen. 2018. “Semantic Segmentation for High Spatial Resolution Remote Sensing Images Based on Convolution Neural Network and Pyramid Pooling Module.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11 (9): 3252–3261. doi:10.1109/JSTARS.2018.2860989.

Yu, B., A. Yang, F. Chen, N. Wang, and L. Wang. 2022. “SNNFD, Spiking Neural Segmentation Network in Frequency Domain Using High Spatial Resolution Images for Building Extraction.” International Journal of Applied Earth Observation and Geoinformation 112: 102930. doi:10.1016/j.jag.2022.102930.

Zhang, Z., W. Guo, L. Mingjie, and Y. Wenxian. 2020. “GIS-Supervised Building Extraction With Label Noise-Adaptive Fully Convolutional Neural Network.” IEEE Geoscience and Remote Sensing Letters 17 (12): 2135–2139. doi:10.1109/LGRS.2019.2963065.

Zhang, L., H. Xin, B. Huang, and P. Li. 2006. “A Pixel Shape Index Coupled with Spectral Information for Classification of High Spatial Resolution Remotely Sensed Imagery.” IEEE Transactions on Geoscience and Remote Sensing 44 (10): 2950–2961. doi:10.1109/TGRS.2006.876704.

Zhao, H., J. Shi, Q. Xiaojuan, X. Wang, and J. Jia. 2017. Pyramid Scene Parsing Network. Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21-26 July 2017.

Zhao, D., H. Zhao, R. Guan, and Y. Chen. 2021. “Efficient Building Extraction for High Spatial Resolution Images Based on Dual Attention Network.” International Journal of Computers, Communications and Control 16 (4). doi:10.15837/ijccc.2021.4.4245.

Zheng, Z., M. Ailong, L. Zhang, and Y. Zhong. 2021. “Deep Multisensor Learning for missing-modality all-weather Mapping.” ISPRS Journal of Photogrammetry and Remote Sensing 174: 254–264. doi:10.1016/j.isprsjprs.2020.12.009.

Zhu, X., and D. Guo. 2014. “Mapping Large Spatial Flow Data with Hierarchical Clustering.” Transactions in GIS 18.

Zhu, Q., L. Zhen, Y. Zhang, and Q. Guan. 2020. “Building Extraction from High Spatial Resolution Remote Sensing Images via Multiscale-Aware and Segmentation-Prior Conditional Random Fields.” Remote Sensing 12 (23): 3983. doi:10.3390/rs12233983.