ITC 3/54
Information Technology and Control
Vol. 54 / No. 3/ 2025
pp. 992-1009
DOI 10.5755/j01.itc.54.3.41147

The Application of Transformer Model in Building Information Modeling

Received 2025/04/11 | Accepted after revision 2025/07/17

HOW TO CITE: Wang, Z. (2025). The Application of Transformer Model in Building Information Modeling. *Information Technology and Control*, 54(3), 992-1009. https://doi.org/10.5755/j01.itc.54.3.41147

The Application of Transformer Model in Building Information Modeling

Zhe Wang
Fujian Chuanzheng Communications College, Fuzhou 350007, China

Corresponding author: forzaiking@126.com

Building information modeling leverages high-resolution satellite imagery and change detection for urban planning and disaster monitoring. This study enhances building information modeling accuracy by integrating coordinate attention with a Transformer hybrid architecture. Local feature extraction by convolutional neural network and global context modeling by Transformer are combined. Feature exchange techniques and a hollow space pyramid pooling module improve multi-scale change detection. Lightweight designs, including depthwise separable convolutions and Ghost modules, reduce computational costs. Experimental results show the model stabilizes after 80 iterations, achieving 95% accuracy and a 1.37% mIoU improvement. With a Kappa value of 0.795 and minimal parameters, the framework enables efficient synthetic aperture radar-based building change detection, suitable for real-time urban monitoring. The raised model can achieve the task of building information modeling, laying the foundation for large-scale automatic recognition and classification of building images.

KEYWORDS: Building information, Transformer model, U-Net network, SAR data, Convolutional network

1. Introduction

As one of the symbols of urban construction, the reconstruction, demolition, growth, and reduction of buildings can largely represent the changes in the city [11]. At the same time, changes in buildings also have significant implications for land resource utilization and disaster damage detection. Therefore, modeling building information and obtaining changes in buildings take a pivotal part in government decision-making and economic development [15, 18, 21]. At present, the building coverage rate in urban areas of China has reached over 80%, but in the process of urbanization, a large number of illegal buildings and demolition problems have also emerged. In addition, in practical applications, the

investigation of urban buildings is still carried out manually and hard to satisfy the requirement of rapid and large-scale change detection. Synthetic Aperture Radar (SAR) data is an effective method for solving the problem of urban building changes. However, due to the presence of coherent speckle noise, the resolution of SAR images decreases and the images become blurry, seriously affecting the change detection effect of SAR images. Convolutional neural networks (CNNs) have emerged as a powerful research approach in recent years [8]. However, due to the lack of description of the overall context, the accuracy of the building model is not high. The Transformer model, as a powerful global feature learning tool, can effectively capture and process global information in remote sensing images (RSIs) through its self attention mechanism (SAM), thereby improving the accuracy and efficiency of building information modeling (BIM). The application of this model solves the problem of insufficient global information capture in traditional CNNs when processing SAR data, showing significant advantages.

BIM is broadly applied within the construction industry, covering multiple stages such as design, construction, and operation. By simulating the real information of buildings through digital information simulation, it achieves the management and optimization of the entire life cycle of buildings. Wahba M et al. used machine learning and an artificial neural network with a multi-layer perceptron architecture to quantify the building vulnerability index. They combined the building information model with environmental flood hazard assessment and classified building samples into five levels. The results showed that this method had high accuracy [14]. Falegari et al. [2] used BIM-life cycle analysis integration technology and passive design strategies to reduce the impact of buildings on the environment. They analyzed appropriate equipment and building materials that matched their respective climates, reducing the energy demand of the model by 30% of the original value [2]. Omaran et al. [9] developed a visualization system to help determine the cost and applicability of different construction projects. The system employed BIM and VR to improve the visualization and processing capabilities of engineering cost prediction. The results showed that the system could reduce processing costs in construction projects [9]. Ismail [5] proposed a building information model based on the physical Internet to make the rigidity and displacement of the design of the beam column connection system meet the required quality and operation requirements. Through statistical literature and data analysis, it improved the improper specifications and defects of concrete and steel members [5].

The Transformer model is a revolutionary neural network (RNN) architecture that significantly improves the processing efficiency and accuracy of natural language processing tasks by introducing SAMs and parallel processing, solving the efficiency problem of traditional RNNs in processing long sequences. Gibril et al. [3] employed the Swin Transformer as the core network to facilitate the capture of extensive semantic data and the extraction of multi-scale features, with the objective of extracting building information from expansive satellite images. The findings denoted that the Transformer-based model was superior to the CNN-based model and had excellent generality [3]. To achieve remote sensing change detection of buildings, Liang et al. [6] employed Transformer-based progressive sampling to direct the model's attention toward the object of interest. Additionally, a proposal was put forth for an adaptive feature-merging module that would fully integrate the features of a CNN with those of a transformer. Following verification, the model denoted superior efficacy contrast to other advanced methods currently in use [6]. Yiming et al. [17] addressed the issue of ViT model lacking shape information enhancement for building objects. They constructed an effective dual path visual segmentation framework based on Transformer model and used multi-shape convolution kernels to perceive and enhance the shape features of buildings. The model achieved significant improvements in three public building datasets [17]. Liu et al. [7] put forth a novel approach to point cloud registration within the context of BIM. This method encoded the process of self-similarity matrix multiplication, integrating geometric insights from the levels of points, lines, and grids to mitigate the impact of normal blurring. The results indicated that the method had good generalization ability among different types of point clouds [7]. The comparison of the method proposed in this paper with the existing literature is shown in Table 1.

Table 1
A comparison of the proposed method with the existing literature.

Research Purpose	Method	Result	Shortcomings	Reference
Quantify building vulnerability index via BIM and environmental flood hazard assessment	Machine learning + multi-layer perceptron ANN	High accuracy in classifying building samples into 5 vulnerability levels	Limited focus on dynamic urban changes; requires extensive manual data labeling	Wahba et al. [14]
Reduce environmental impact of buildings via BIM-life cycle analysis	BIM integration with passive design strategies and climate-appropriate materials/equipment	Reduced energy demand by 30%	Complexity in real-world implementation; lacks scalability for large-scale cities	Falegari et al. [2]
Improve cost and applicability visualization in construction projects	BIM + VR for engineering cost prediction	Reduced processing costs	Limited adaptability to real-time data updates; high computational overhead	Omaran et al. [9]
Enhance beam-column connection design quality using BIM and physical Internet	Statistical analysis of concrete/steel specifications	Improved rigidity and displacement compliance	Narrow scope (specific structural elements); lacks multi-scale feature extraction	Ismail [5]
Extract building information from satellite imagery	Swin Transformer for multi-scale feature extraction	Superior to CNN-based models in generality and accuracy	High computational complexity; unsuitable for edge devices	Gibril et al. [3]
Remote sensing building change detection	Hybrid CNN-Transformer with adaptive feature merging	Outperformed state-of-the-art methods	Limited robustness under SAR noise; no lightweight design	Liang et al. [6]
Enhance building shape features in segmentation	Dual-path Transformer framework + multi-shape convolution kernels	Improved performance on public datasets	Requires high-resolution images; ignores small-area building changes	Yiming et al. [17]
Improve point cloud registration in BIM	Self-similarity matrix encoding + geometric insights from points/lines/grids	Good generalization across point cloud types	Inefficient for real-time applications; lacks SAR data compatibility	Liu et al. [7]
Enhance SAR-based building change detection for urban monitoring	Hybrid CNN-Transformer + CAM, DSC, Ghost modules, improved VSPP	95% accuracy, 76.39% mIoU, 0.7955 Kappa, 50% fewer parameters	/	This paper

In summary, although existing studies have made some progress in BIM through CNN or Transformer, their performance is insufficient in capturing long-distance dependencies, resulting in prominent false detection and missed detection of small-area building changes, which is difficult to meet the needs of high-precision urban monitoring. Moreover, the traditional Transformer model has a large number of parameters and high computational complexity, making it difficult to

run in real time in edge devices. The speckle noise and low-resolution characteristics of SAR images exacerbate the problem of blurred building boundaries. The existing attention mechanisms fail to effectively combine the direction-aware features, resulting in insufficient background noise suppression ability. To this end, the study proposes a lightweight CNN-Transformer hybrid architecture, aiming to solve the above problems and provide an efficient solution for real-time building change detection of SAR data. This study combined the local feature extraction strength of CNN with the global context modeling of Transformer, and then achieved lightweight through Depthwise Separable Convolutions (DSC) and Ghost modules. This enabled real-time city monitoring on edge devices. Finally, an improved Atrous Spatial Pyramid Pooling (ASPP) module was introduced to capture multi-scale building features and ensure robust detection of different building sizes.

The innovation of this paper is mainly reflected in the following three aspects: (1) Hybrid architecture design: For the first time, the coordinate attention mechanism is combined with Transformer-UNET. The background noise of SAR images is suppressed through the Coordinate Attention Module (CAM), and at the same time, the global modeling ability of Transformer is utilized to capture long-distance dependencies; (2) Lightweight optimization: By reconstructing the decoder through DSC and Ghost modules, the computational cost is significantly reduced while maintaining accuracy, providing a technical foundation for real-time monitoring. (3) Multi-scale feature enhancement: Improve the ASPP module to introduce Strip Pooling, enhance the perception ability of narrow and long building structures, and reduce the missed detection rate.

The core contributions of this study include: (1) Proposing a hybrid architecture combining CAM and improved ASPP, which significantly improves the detection accuracy of changes in small-area buildings through the enhancement of direction-aware features and multi-scale information fusion; (2) By introducing DSC and the Ghost module, the model parameter count was reduced to 36% of the original Transformer, and the FLOPs was decreased to one-third, achieving efficient deployment of edge devices.

2. Methods and Materials

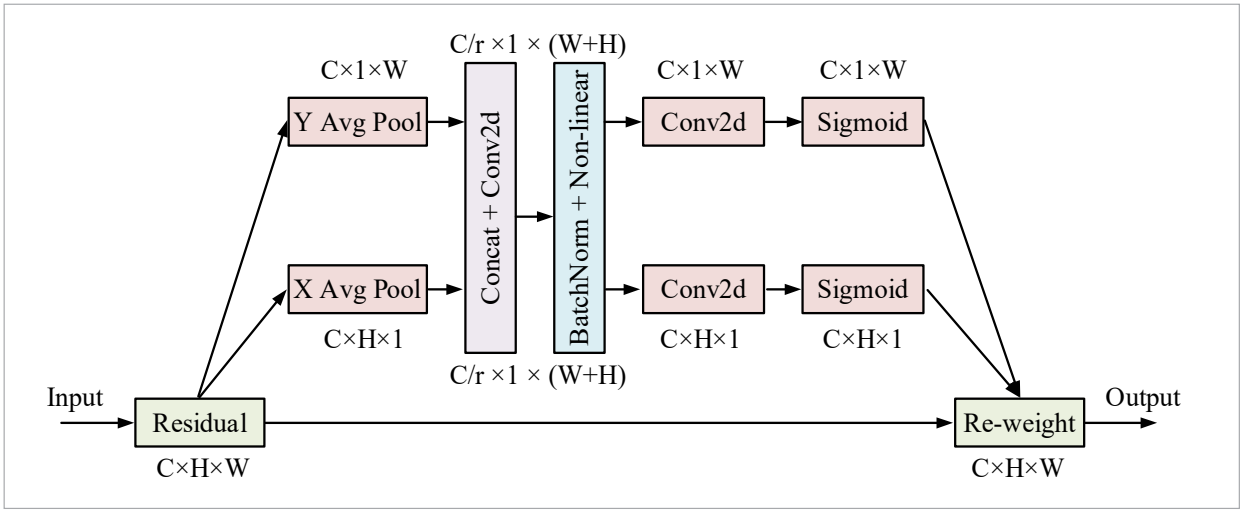
Research aims to design innovative models and modules to adapt to the characteristics of building RSIs, enabling them to extract rich contextual information from generated windows and learn better target representations. Additionally, deep separable convolution modules and Ghost modules are introduced to lessen the computational cost and memory usage of Transformer models.

2.1. BIM Based on Improved Transformer Model

The method based on the combination of U-Net and Transformer has been broadly utilized in image segmentation [13, 19, 1]. The main reason is that a single Transformer model lacks low-level feature details, resulting in low localization accuracy; By using CNN-Transformer hybrid encoding, both overall and local information can be considered. Therefore, the study applied this method to building change detection in SAR images. However, due to the complexity of SAR images and changes in buildings, the original Transformer-UNet model still needs further improvement.

At present, due to the complexity of the variation characteristics information in SAR images and the high background noise, it is difficult to distinguish them from the surrounding environment. The Coordinate Attention Module (CAM) can better eliminate background noise while maintaining valid information, effectively solving this problem [10, 22]. The structure of CAM is denoted in Figure 1, which depicts two principal stages: the embedding of coordinate information and the generation of coordinate attention. At the stage of incorporating coordinate data, the module employs two 1D global pooling operations to integrate features in both the horizontal and vertical dimensions, resulting in the generation of a pair of direction-aware Feature Maps (FMs). In the coordinate attention generation step, these FMs are combined and passed to a shared 1x1 convolutional transformation function. After nonlinear activation, they are divided into two separate tensors. Next, two 1x1 convolutional transformation functions are used to generate attention maps in horizontal and vertical directions. The final output is the product of the input FM and these attention maps, thereby enhancing the model's expressive power.

Figure 1
Structure diagram of CAM.



Firstly, a global average pooling method is employed to extract FMs in the width and height dimensions. Specifically, on an input feature tensor X, the pooling core with a lateral coordinate of size $(H,1)$ is applied to encode the characteristics of each channel. Therefore, the height of the cth channel is h, and the output can be represented as Equation (1).

$$y_c^h(h)=\frac{1}{W}\sum_{0\le i\le W}x_c(h,i). \tag{1}$$

In Equation (1), W is the width of the feature graph, that is, the dimension of the feature tensor in the horizontal direction. c is the channel index, which represents a particular channel or channel in the feature map. h is a height index that represents a particular row in the feature graph. w is the width index that represents a particular column in the feature graph. y_c^h represents the average characteristic value of channel c at height h, where $x_c(h,i)$ is the value of the characteristic graph of channel c with height h and width i. Similarly, the outputs of c channels with w width can be represented by Equation (2).

$$y_c^w(w)=\frac{1}{H}\sum_{0\le j\le H}x_c(j,w). \tag{2}$$

In Equation (2), y_c^w represents the average characteristic value of the c channel in width w, $x_c(j,w)$ is the numerical value of the characteristic map of the cth channel with a height j and width w, and H is the height of the characteristic map. This method combines features from two directions of an image to form a pair of directional FMs. This process is distinct from channel attention, which transforms feature tensors into a unified feature vector via 2D global pooling. Coordinate attention represents a decomposition of channel attention into two 1D feature encoding processes, which aggregate features along two spatial directions. In this manner, long-range dependencies can be captured along one spatial direction while precise positional information is maintained along another. Subsequently, the generated FMs are encoded into a pair of direction-aware and position-sensitive attention maps, which can be complementarily applied to the inputting FM to promote the representation of the object of interest.

By using Equations (1)-(2), accurate location information of the global perception field can be obtained effectively. After information embedding transformation, the aggregated FMs generated by Equations (1)-(2) are transformed using a 1x1 convolution transformation function F {1} to gain the intermediate FM f of the horizontally and vertically encoded spatial information, as shown in Equation (3).

$$f=\delta\left(F_1\left(\left[y^h,y^w\right]\right)\right). \tag{3}$$

In Equation (3), [ ] means the concatenation operation at the spatial scale, δ represents the nonlinear activation function, $f \epsilon R^{C/r\times(H+W)}$ represents the horizontal and vertical spatial information encoding, and r represents the reduction rate of module size. Using the transformation functions F_h and F_w, the f^h and f^w obtained by decomposing f are transformed into tensors X with equal channel numbers, resulting in Equation (4).

$$\begin{cases} g^h = \delta\left(F_h\left(f^h\right)\right) \\ g^w = \delta\left(F_w\left(f^w\right)\right) \end{cases}. \tag{4}$$

The final output of CAM can be written as Equation (5).

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j). \tag{5}$$

Compared to existing attention modules, CAM is a technique that enables the capture of information between multiple channels simultaneously. Additionally, it allows for the capture of directional and positional information of interest, thereby strengthening the model's ability to precisely identify and localize objects of interest. Using dilated convolution can solve the problem of detail loss caused by resolution reduction (pooling or stride convolution), while also expanding the receptive field. Assuming that in the two-dimensional scenario, the corresponding outputs at each point i are represented by q, and the weights ω of the features are used to represent them. Then formula (6) is used to represent the convolution calculation of the input feature layer p.

$$q_i = \sum_k p_{[z+l\times k]} \times \omega_k. \tag{6}$$

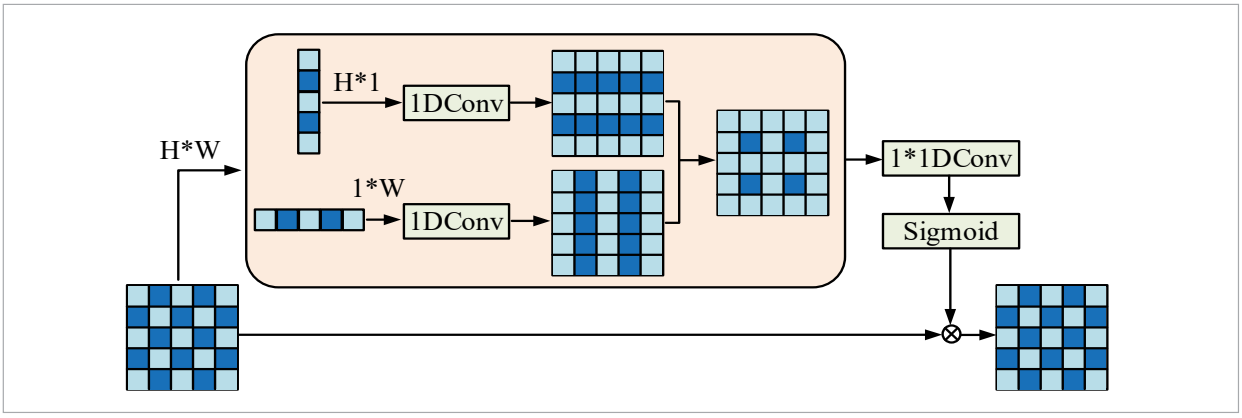
In Equation (6), k denotes the size of the convolution core, and l means the dilation rate. In the case of l=1, it is the standard convolution. The Void Space Pyramid Pool (VSPP) module aims to further extract multi-scale information. The design inspiration for this module comes from spatial pyramid pooling, but by introducing dilated convolutions at different rates, VSPP can capture a wider range of contextual information. The improved VSPP structure proposed by the research is shown in Figure 2. When an object is in the shape of a long strip (such as a newly built building), it may contain interference from other unrelated areas. In Strip Pooling (SP), the rectangular sampling window can reduce the acquisition of irrelevant information, thereby alleviating the impact of these problems to some extent.

In response to the small differences in the process of BIM, the study adopts SP for feature extraction. This method is different from the two-dimensional pooling method, as it calculates the average of all eigenvalues of rows and columns. In a spatial range of $1\times N$ or $N\times 1$, the output $z^h \epsilon R^H$ after the horizontal SP can be expressed using Equation (7).

$$z_i^h = \frac{1}{W} \sum_{0\le j\le W} x_{i,j} \tag{7}$$

Similarly, the outputting $z^w \epsilon R^W$ after vertical SP is indicated as Equation (8).

Figure 2
Improved VSPP structure.



$$z_i^w = \frac{1}{H} \sum_{0 \le i \le H} x_{i,j} \,. \tag{8}$$

The SP module is constructed based on the strip pool operation, and its structure is shown in Figure 3. It considers a long and narrow kernel, namely 1×N or N×1. Based on SP, further research is conducted on spatial pooling architecture design. By introducing a novel SP, the backbone network can efficiently simulate long-distance dependencies. The introduction of a novel building block can facilitate more accurate detection of small and narrow buildings, as well as enhance the efficacy of building change detection.

The structure of the optimized Transformer-UNet model is shown in Figure 4. To solve the problem of unclear edge detection caused by excessive background noise in the detection of building changes in SAR images, this paper integrates the CA module into the encoder and decoder of the Transformer-UNet model. In the encoder path, the CA module is integrated into the CNN part that combines CNN

Figure 3
SP structure.

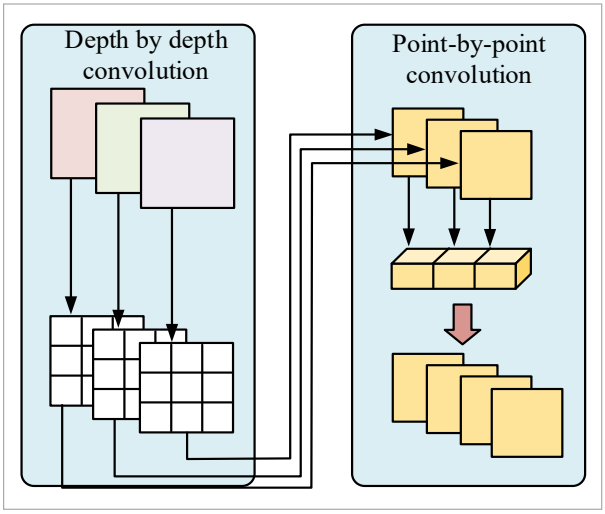
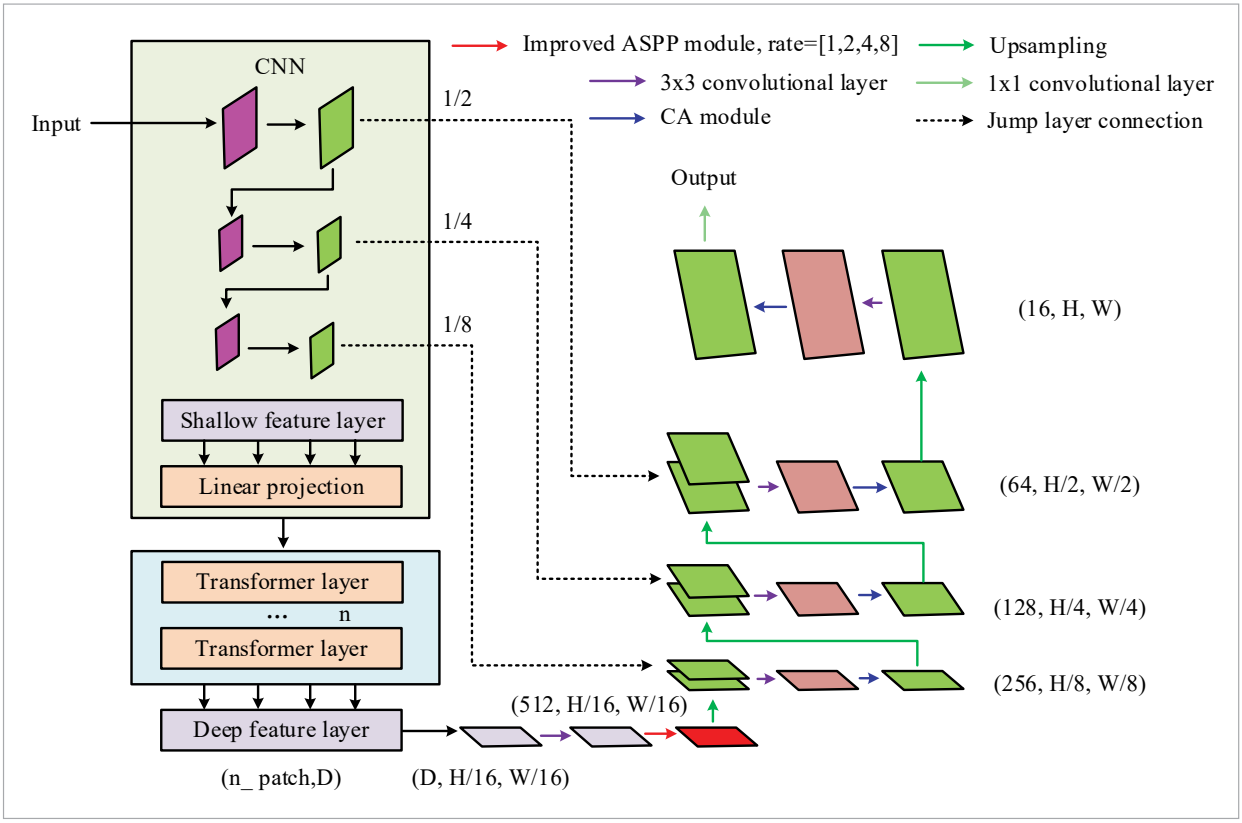


Figure 4
Improving the Transformer model framework.



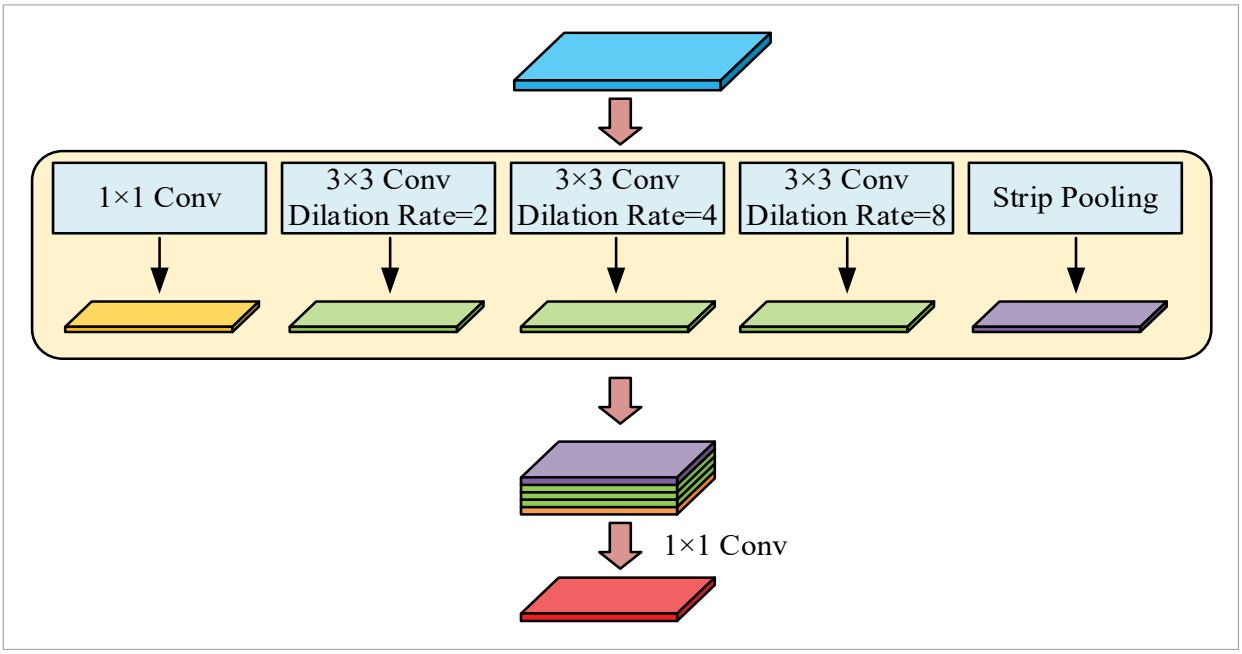
and Transformer; In the decoder path, the CA is placed after all the convolutional layers. The introduction of the CA module enables the model to pay more attention to the regions of interest and obtain broader information through effective positioning in the pixel coordinate system, thereby distinguishing the background and the foreground more effectively and achieving a better semantic segmentation effect. Furthermore, due to the lightweight design of the CA module, it does not impose excessive computational burdens. In order to obtain the multi-scale information of the target, improve the accuracy of detecting changes in small-area buildings, and reduce missed and false detecsions, this paper also introduces the improved ASPP module between the encoder and decoder of Transformer-UNet, so as to obtain the multi-scale information of the target before upsampling.

2.2. The Lightweighting of Improved Transformer Models

The traditional Transformer method has disadvantages such as large computational complexity, slow detection speed, and high requirements for instrument configuration. Therefore, research is being conducted on the lightweight design of existing Transformer models, aiming to maintain high detection accuracy while minimizing model parameters and computational complexity. DSCs are a kind of separable convolution that can handle both spatial and depth dimensions. This approach enables for the extraction of additional features and the reduction of more parameters [16, 4]. The diagram of the DSC network is presented in Figure 5.

It assumes that there exists a sample set with the shape $D_1 \times D_2 \times D_3$ for depth wise separable convolution. First is to use the same amount of convolution kernels as the previous layer's channels as the first convolution. In deep learning, filters are the fundamental components of CNNs applied to extract local features from input data. When there is only one filter with a fixed size, it will slide along the width and height directions of the input data when processing it, and apply the same operation to each position. The result of this process is to generate an FM with the same number of channels as the inputting data, as each channel's data is independently filtered to generate a new FM. This processing method preserves all channel information of the input data, so the quantity of outputting FM is analogous to the quantity of channels in the inputting layer. This algorithm uses convolutional kernels to

Figure 5
Diagram of DSC network structure



weight and merge the features obtained in the previous step, forming a new FM. On this basis, a feature mapping method based on CNNs is proposed. The time complexity T of DSC can be calculated using equation (9).

$$T = D_1 \times D_2 \times D_3 \times B_1 \times B_2 \times B_3 \times C_{in} + D1 \times D2 \times D3 \times C_{in} \times C_{out} \quad (9)$$

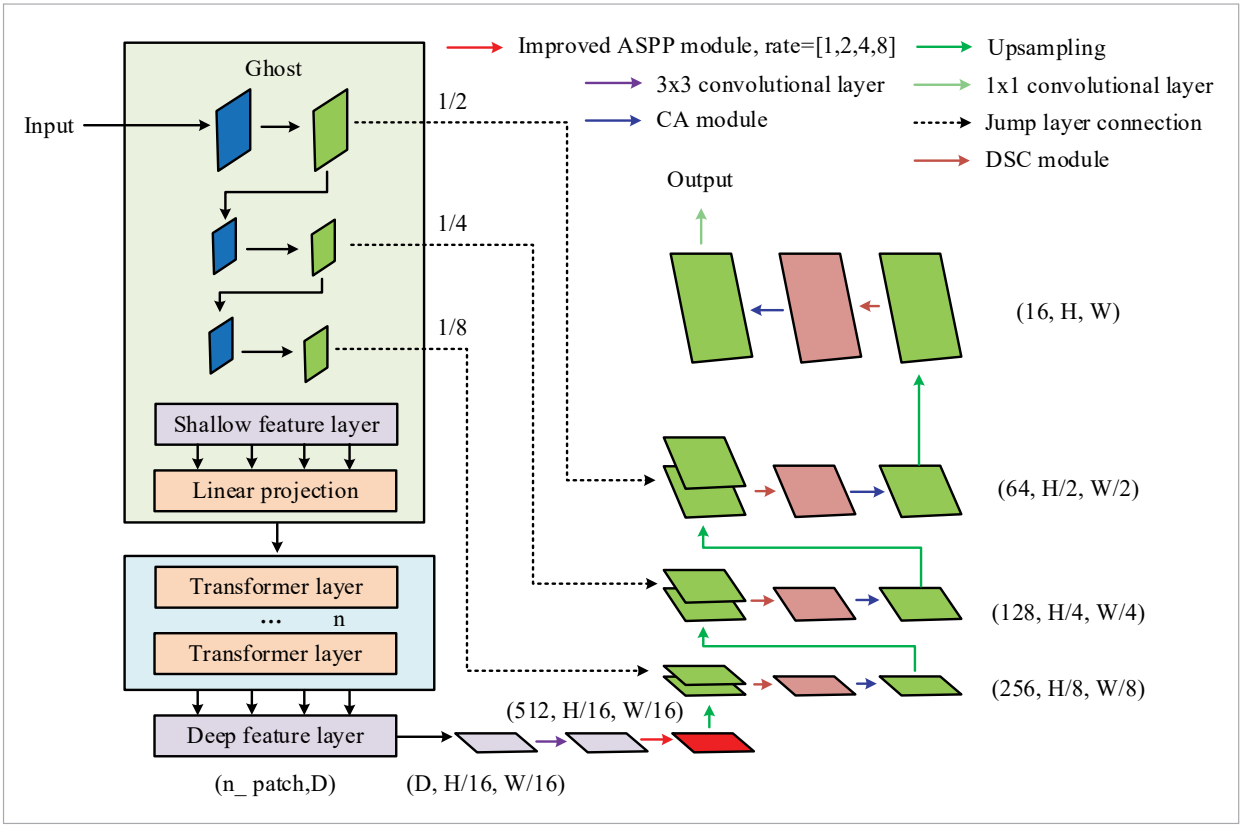
In Equation (10), C_{in} and C_{out} represent the amount of channels in the upper and lower layers, respectively. The time complexity T' of conventional convolution can be calculated using Equation (10).

$$T' = D_1 \times D_2 \times D_3 \times B_1 \times B_2 \times B_3 \times C_{in} \times C_{out} \quad (10)$$

The traditional feature extraction method of CNNs uses convolution functions as the operating units to perform operations on each input channel. However, stacking in multi-layer convolutions generates a large number of redundant features, which can be used for training neural networks but incur significant overhead. Ghost is a structure that generates massive FMs with minimal operations and is a method of compressing models. This method employs a shortcut to establish a connection between the inputting and outputting of two Ghost modules. The second Ghost module does not require ReLU, while the rest is implemented through a batch of normalization (BN) and ReLU.

Figure 6 shows the improved transformer model structure. To achieve the goal of model lightweighting, the study adopts DSC and Ghost model to improve detection efficiency while ensuring the accuracy of change detection. Firstly, the original convolution in the Transformer-UNet decoder is replaced, reducing its parameters, speeding up inference, and improving computational efficiency. The improved Transformer

Figure 6
The framework of the lightweight improved Transformer model.



model utilizes the implicit positional representation capability of convolution (with zero padding) to perform conditional positional encoding on inputs of any size. Then, the hierarchical pyramid structure is executed by gradually reducing the number of tokens and replacing class tokens with an average pool. Therefore, with the support of multi-level features, Transformer can easily handle object detection and image segmentation tasks.

Before performing SAR image change detection, the original SAR image needs to be preprocessed first. This includes steps such as radiometric, atmospheric, and geometric corrections, and filtering. Radiometric correction and atmospheric correction can reduce radiation noise and atmospheric interference in SAR images. Geometric correction can eliminate geometric distortions in SAR images, enabling accurate comparison of SAR images at different time points. Filtering can reduce noise and clutter in SAR images and enhance the visibility of changing signals. The mathematical expression for SAR image logarithmic ratio differential image is denoted in Equation (11).

$$DI = \log\left(\frac{X_1 + \text{eps}}{X_2 + \text{eps}}\right). \tag{11}$$

In Equation (11), X_1 and X_2 represent two different regional images, and eps is a very small decimal. In addition, to enhance the difference between the changed and non-changed regions, the nonlinear transformation of the differential image is studied, expressed as Equation (12), where DI' is the enhanced differential image.

$$DI' = \frac{1}{1 + e^{-DI}}. \tag{12}$$

To avoid imbalanced samples of different types, which may lead to the network tending towards unchanged categories, the study used joint loss function, dice loss L_D, and cross-entropy loss function (LCE) L_{CE} to supervise the model. The joint loss L is denoted as Equation (13).

$$L = L_D + L_{CE}. \tag{13}$$

3. Results

To prove the efficacy of the raised method, multiple evaluation criteria were used for quantitative assessment, including Mean Intersection over Union (MIoU), F1 value, accuracy, etc. These assessment criteria collectively constitute a comprehensive assessment of the model's efficacy, reflecting the accuracy, efficiency, and recognition ability of the model for different categories from different perspectives.

3.1. Performance Analysis of Model Checking

The operating system used in this experiment was Ubuntu 18.04.5 LTS, and the CPU was Intel(R) Core(TM) i9-11900K @ 3 50GHz, 32GB of memory, GeForce RTX 3090 GPU. Deep learning framework was Pytorch 1.7.0+Jul10. The number of iterations was 200, and the batch size was set to 8. The initial learning rate was set to 0.001, and with each repetition, the learning rate would decrease to 90% of its original level. When the number of iterations was 100, the learning rate became 0.00001. The experiment used Adam optimization algorithm to optimize all training models. The experiment utilized the Data Archiving and Distribution Center of the Alaska Satellite Facility (Alaska Satellite Facility Distributed Active Archive Center). Acquisition of SAR images from Sentinel-1 satellite data of ASF DAAC (ASF DAAC: https://search.asf.alaska.edu/#/?zoom=3.495&center=130.110, 20.186), covers the October 1, 2022 to June 1, 2023 China fuzhou two periods of data. The high-resolution (10m) and dual temporal (8-month interval) characteristics of this dataset enable precise capture of building contour dynamics; SAR imaging penetrates cloud layers and light limitations, suitable for monitoring in areas with multiple clouds and rain; Covering dense urban areas, suburbs, and mixed terrain, simulating real and complex environments. This dataset has been optimized through preprocessing and diversified scene coverage, effectively verifying the performance of the model in noise suppression, small target detection, and multi-scale feature extraction, ensuring the universality and practical value of the research conclusions.

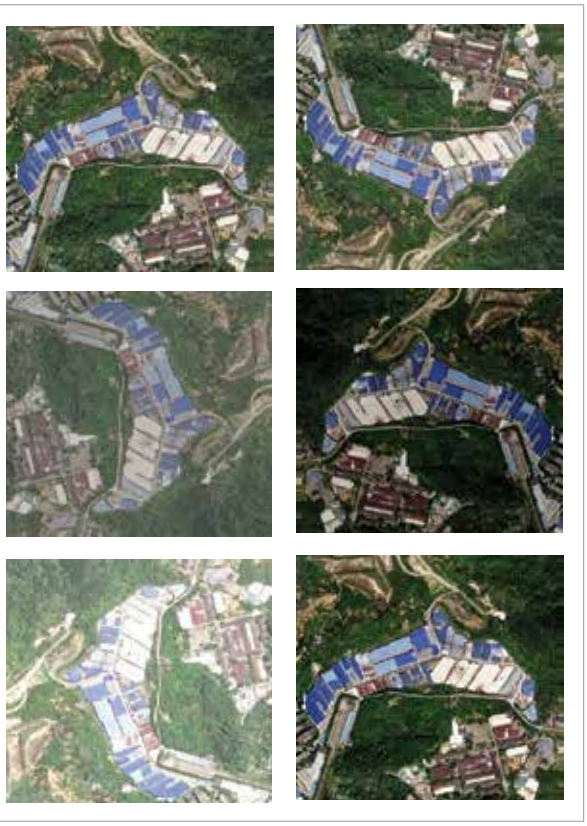
The raw data is preprocessed to generate 4682 sample blocks of 512 × 512 pixels, which are divided into a training set (4214) and a testing set (468) in a 9:1 ratio. Some of the samples are shown in Figure

7. The data covers various types of building changes such as new construction, demolition, and renovation, and includes dense urban, suburban, and mixed terrain scenarios to ensure the model's generalization ability. To improve model reliability and experimental reproducibility, this study refined the preprocessing process. The preprocessing steps include: (1) radiometric correction to eliminate sensor response bias; (2) Geometric correction (combined with SRTM 30m DEM), controlling registration error<1 pixel; (3) Lee filter (7 × 7 window) suppresses speckle noise; (4) Logarithmic ratio differential enhances contrast in the changing region; (5) Sample annotation (integrating OpenStreetMap and manual validation) ensures label accuracy>95%.

By training the loss function curve of the experimental model, it can be determined whether the settings of parameters such as learning rate of the experimental model are reasonable. The loss values and accuracy changes of the experimental model during training are indicated in Figure 8. Through the estimation of the first-order moment and the second-order moment, Adam can dynamically adjust the update step size of each parameter. Especially in the later stage of training, it can achieve more fine parameter fine-tuning, thereby making the loss curve tend to be smooth. The initial learning rate is set at 0.001. It decays at a 90% ratio in each round of iteration and drops to 0.00001 after 100 rounds. This exponential attenuation method avoids the optimization fluctuation caused by the sudden change of the learning rate, enabling the model to converge rapidly in the early stage of training and approach the optimal solution slowly in the later stage, further ensuring the stability of the loss curve. From Figure 8(a), the loss decreased sharply in the first 20 epochs, stabilized after 80 epochs, and converged by 100 epochs. From the change curve of the loss value, during the iteration, as the

Figure 7

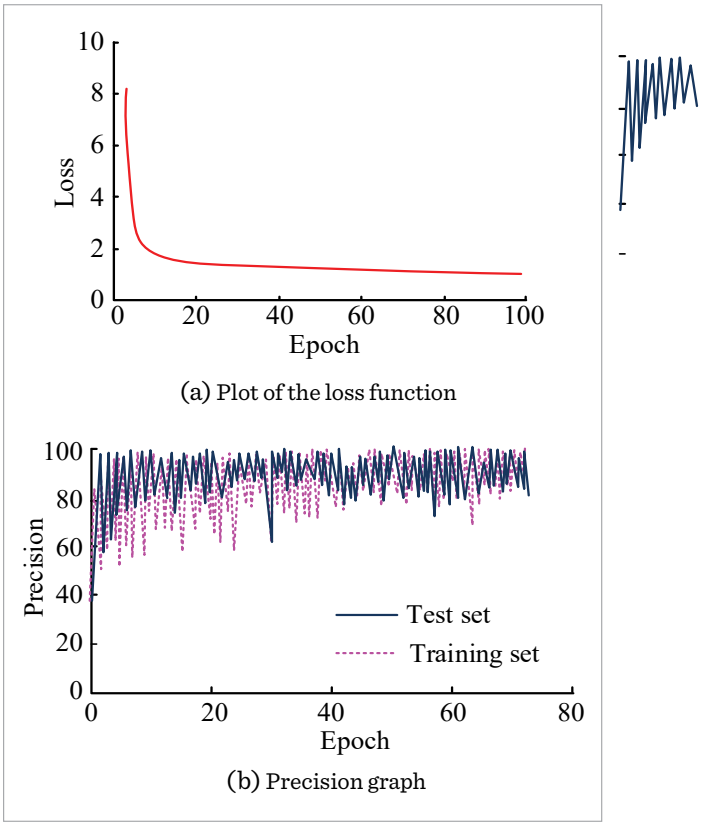
Sample data example (Image source: ASF DAAC: https://search.asf.alaska.edu/#/?zoom=3.495¢er=130.110,20.186)



Note: Picture shows the SAR images of the regions with latitudes and longitude [118°57′36″ E, 25°50′39″N].

Figure 8

The loss value and precision of the model change during training.



training time prolonged, the value of the loss gradually decreased. However, the downward trend indicated that the model was in an ideal learning stage. If the value of the loss decreased to a certain extent and gradually stabilized, it meant that the training of the model has ended and certain results have been achieved. The epoch of this model was 100 times, and when the loss value reached 80, the loss value gradually stabilized and began to converge. From Figure 8(b), accuracy rose rapidly to 90% within 20 epochs, gradually approaching 95% at 100 epochs. This indicates that the improved Transformer model can recognize different types of buildings.

Under the weather interference of cloud cover and brightness changes, the model often experiences

Figure 9

Comparison results of IoU values of different categories.

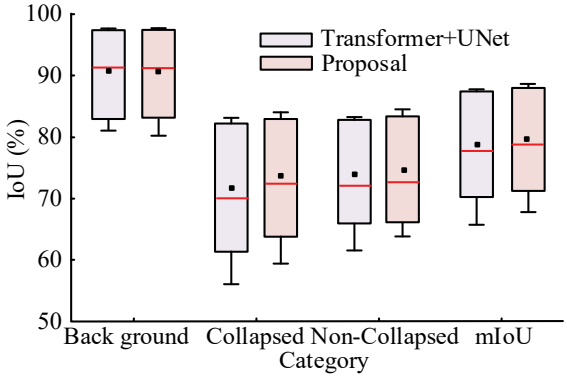
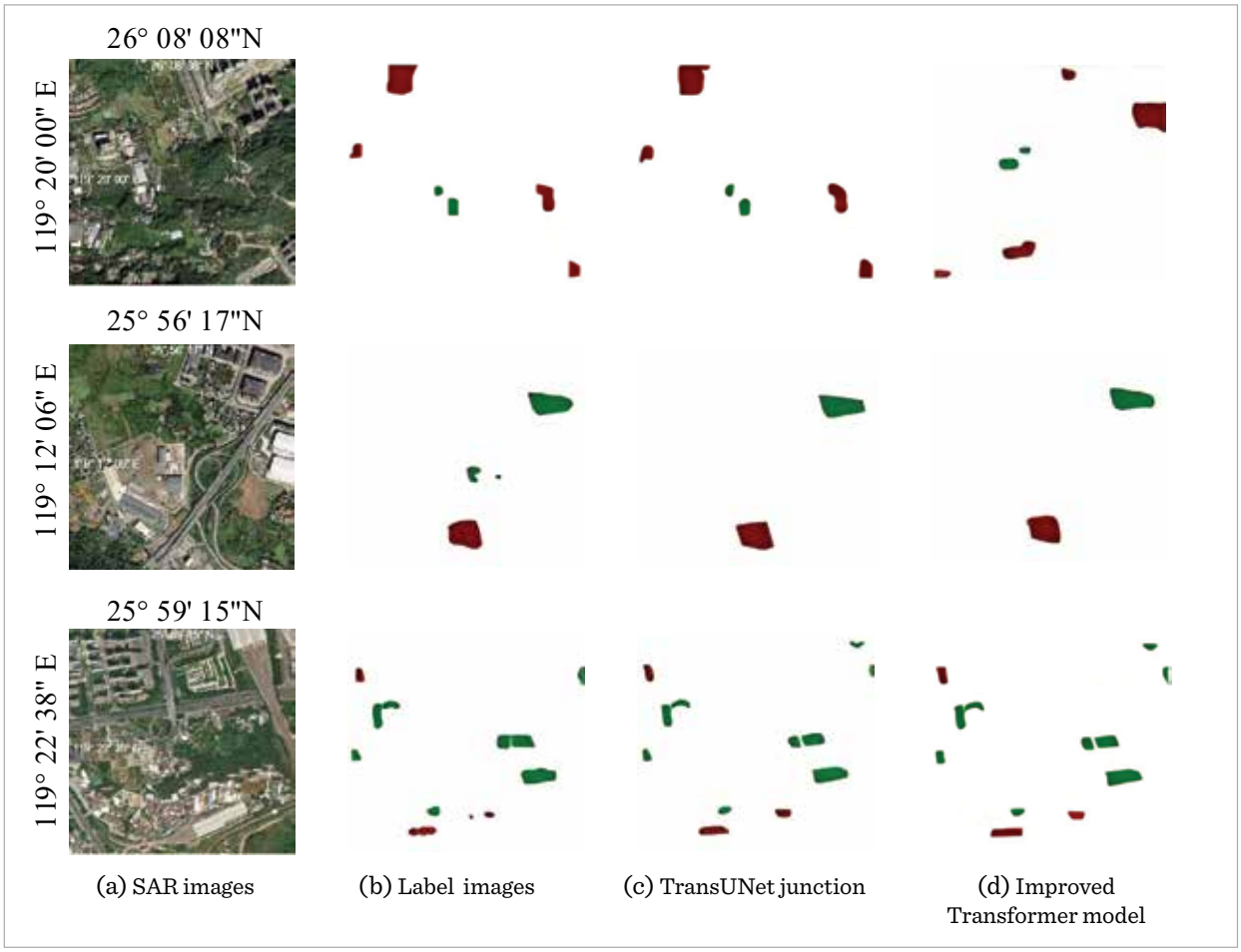


Figure 10

Comparative experimental visualization results (Picture (a) Source from: ASF DAAC: https://search.asf.alaska.edu/#/?zoom=3.495¢er=130.110,20.186)



Note: Picture (a) shows the SAR images of the regions with latitudes and longitude [119°20 ′00 ″E, 26°08′ 08″ N], [119°12 ′06 ″E, 25°56′ 17″ N], and [119°22 ′38 ″E, 25°59′ 15″ N].

incorrect recognition of collapsed building backgrounds and incomplete recognition of non-collapsed buildings. In Figure 9, the improved Transformer model raised in the study improved the IoU accuracy of each category, with lower volatility than the Transformer-UNet model. At the same time, the maximum, median, and mIoU accuracy obtained on the test set were higher than those of the Transformer-UNet model, indicating that the robustness and stability of the proposed model were superior.

Figure 10 shows the visualization results of the comparative experiments of some samples. Figure 10(a) shows the original SAR image. By comparing Figure 10(b) to Figure 10(d), it can be found that the improved Transformer model can effectively handle the blurrness problem of building boundaries, greatly promoting the missed detection and false detection problems of buildings in a small area, making it more consistent with the labeled graph.

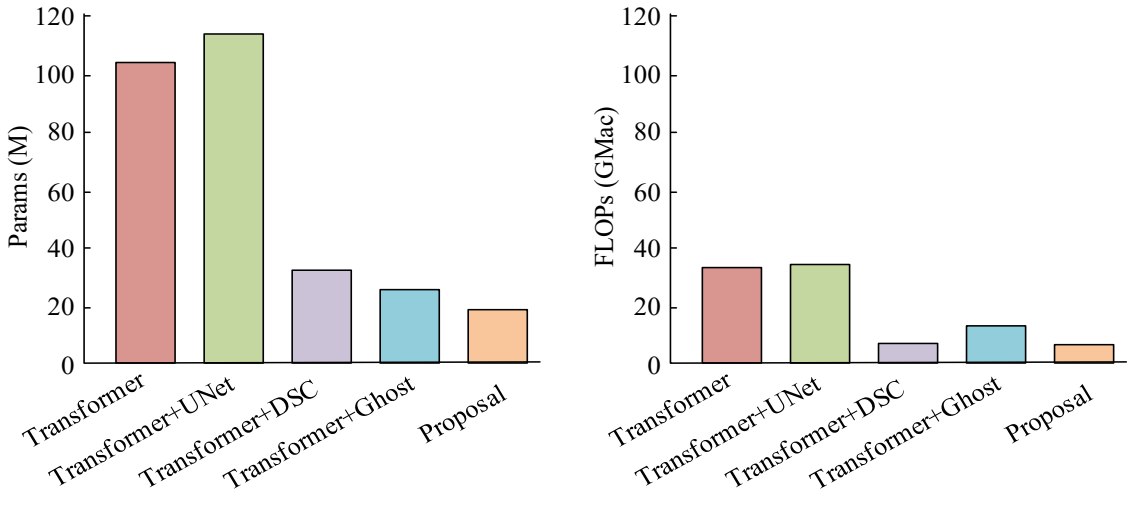
3.2. Ablation Experiment and Analysis of Lightweight Effect

Table 1 presents the comparison results of the improved model with the baseline algorithm Transformer-UNet, as well as the method Swin Transformer-UNet proposed in reference [12] and the method Dual-Path CNN-Transformer proposed in reference [20] on 10 performance indicators. The model in this paper has achieved significant advantages in core indicators such as Mean Intersection over Union (mIoU), F1 Score, and Accuracy. Specifically, the mIoU and F1 scores increased by 0.57% and 0.93% respectively compared with Swin Transformer-UNet, confirming the model's precise segmentation ability for the boundaries of architectural change areas. Meanwhile, its Recall Rate and False Negative Rate (FNR) are superior to all comparison methods. Especially in the detection of small area changes, the missed detection rate is significantly reduced. The Specificity of 99.92% and the False Positive Rate (FPR) of 0.08% reflect the strong reliability of the test results. Among the comparison methods, the Dual-Path CNN-Transformer achieves 91.20% and 99.91% respectively in terms of accuracy and specificity by extracting dual-path features of local details and global semantics. However, its missed detection rate for road ancillary facilities is as high as 19.20%, exposing its limitations in complex scenarios. The model in this paper achieves breakthroughs through three aspects of improvement: introducing the class activation map CAM to suppress the background noise of SAR images, adopting the lightweight structure of depth-separable convolution DSC combined with the Ghost module to reduce the computational complexity, and optimizing the pyramid pooling ASPP module in the hollow space to enhance the multi-scale feature fusion ability. Experiments prove that this scheme surpasses the existing methods in terms of accuracy, computational efficiency and generalization, provid-

Table 1
Performance comparison between the improved model and the latest method.

Model	Transformer-UNet	Swin Transformer-UNet	Dual-Path CNN-Trans	This study
mIoU (%)	74.11	75.82	75.05	76.39
F1 (%)	84.1	85.3	84.95	86.23
Accuracy (%)	99.15	99.28	99.22	99.36
Kappa	0.7662	0.781	0.7748	0.7955
Precision (%)	89.23	90.45	91.2	92.15
Recall (%)	82.47	83.65	80.8	85.71
Specificity (%)	99.87	99.89	99.91	99.92
Dice (%)	83.74	85.1	84.5	87.34
FPR (%)	0.13	0.11	0.09	0.08
FNR (%)	17.53	16.35	19.2	14.29

Figure 11
Analysis of lightweight effect.



(a) Comparison of the number of parameters of each network model

(b) Comparison of computation amount of each network model

ing a better technical path for the detection of architectural changes in SAR images.

Research was conducted to optimize model parameters by incorporating DSC and Ghost modules. By analyzing the impact of DSC, Ghost, and other parameters on model parameterization, the differences in the number of parameters and computation time were compared. Figure 11 shows the comparison results of different network parameter numbers and computational complexity. From the figure, replacing CNN with DSC could reduce the number of parameters in the Transformer model by half and the number of FLOPs by one-third of the original. Replacing CNN with Ghost significantly reduced the number of parameters in the Transformer model. The improved Transformer model proposed by the research, although incorporating CAM and improved ASPP modules, improved parameters and FLOPs contrast to the original Transformer model, but still had the least total amount of parameters and the least computational complexity. The parameter count and FLOPs of the improved Transformer model after lightweighting were much lower than those of the original Transformer model.

The study combined other classic change detection models to qualitatively and quantitatively analyze SAR data, to validate the utilization value of the raised method in SAR images. The study selected several networks as comparative experiments, including FCN, UNet, UNet++, ResUnet, and Transformer-UNet. The first four algorithms were based on the classic CNN, while the latter one was implemented within the Transformer framework. The quantitative outcomes of various network modes are indicated in Table 2. From the table, there were significant differences in the objective accuracy index of each method's experimental results. UNet and UNet++ adopted a skip connection method, utilizing low-level features for spatial information fusion, which resulted in slightly inferior segmentation performance compared to Transformer-UNet. After multiple convolution processes, the residual data in ResUNet images could effectively extract spatial information from the images, while containing less information about small-scale buildings. This method adopted the traditional CNN method, which combines the transformer with a standard convolutional network in order, resulting in a 0.98% improvement in the segmentation accuracy of the fused image compared to UNet. At the same time, it also had

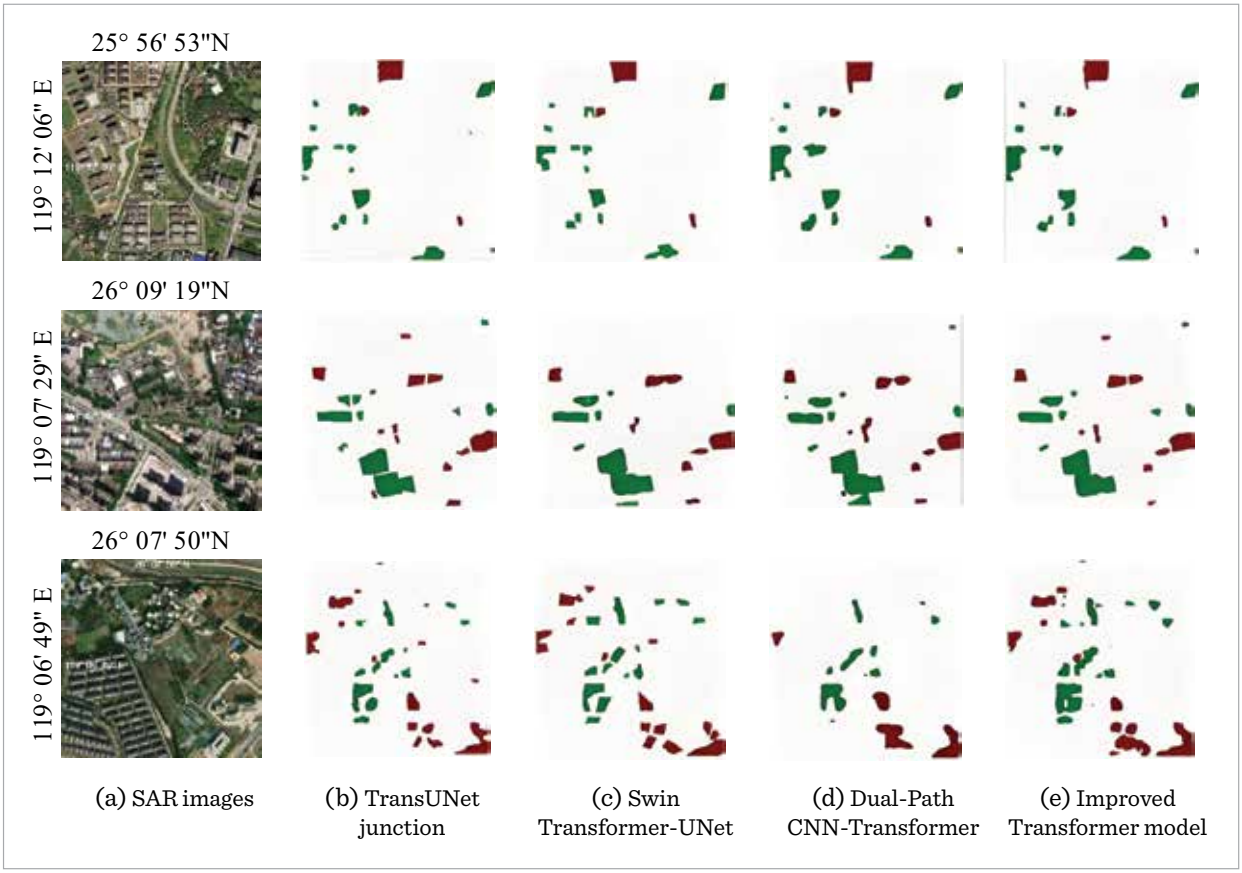
certain improvements in the accuracy and Kappa of change detection. The findings show that combining CNN with Transformer is an effective method that can achieve good results. Compared with Transformer-UNet, the model proposed by the research had higher accuracy with a Kappa of 0.795, making it more suitable for building change detection in SAR data.

The visualization results of the improved method and the latest method are shown in Figure 12. It can be seen from the figure that compared with the original TransUNet model, the last two columns have a better effect in the detection of changes in small-area buildings. The detection effect of the method proposed in this paper is significantly better than that of

Table 2
Building modeling performance comparison.

Model	FCN	ResUnet	UNet++	Transformer-UNet	UNet	Proposal
mIoU (%)	71.82	72.93	73.42	74.11	73.21	76.39
Average F1 value (%)	82.22	83.51	83.71	84.1	83.62	86.23
Accuracy (%)	98.22	98.92	99.09	99.15	99.01	99.36
Kappa coefficient	0.709	0.7493	0.7605	0.7662	0.7532	0.7955
Total number of errors	26047	15423	9432	7823	11620	5644

Figure 12
Visualization results of the improved method and the latest method (Picture (a) Source from: ASF DAAC: https://search.asf.alaska.edu/#/?zoom=3.495¢er=130.110,20.186)



Note: Picture (a) shows the SAR images of the regions with latitudes and longitude [119°12 ′06 ″E, 25°56′ 53″ N], [119°07 ′29 ″E, 26°09′ 19″ N], and [119°06′49 ″E, 26°07′ 50″ N].

the Dual-Path CNN-Transformer. It can be seen that the selection of the sampling rate depends on the data set and even the model itself. The results show that the method proposed in this paper can capture multi-scale context information with the receptive field remaining unchanged, improve the detection ability of the image semantic segmentation network for the changes of small-area buildings, and enhance the detection accuracy. However, extreme weather conditions (such as heavy rain or snow accumulation) may exacerbate image blurring and affect the detection accuracy; Furthermore, the model relies on fully supervised learning and requires a large amount of labeled data. However, in actual scenarios, obtaining high-quality labeled SAR data is costly. Although lightweight design reduces computational complexity, it is still necessary to optimize hardware compatibility and computing latency issues when dealing with large-scale urban monitoring in real time. In the future, multi-sensor fusion and self-supervised learning need to be combined to enhance robustness.

4. Discussion and Conclusion

A deep learning-based SAR urban building area overlay accurate detection method was proposed by combining the advantages of Transformer model and CNN. The method effectively extracted global and long-range features through SAM, and CNN had a strong ability to extract local features. It also mined inter channel overlay features and interferometric phase overlay features to raise the accuracy and robustness. The loss value of the proposed model by gradually stabilized after more than 80 iterations, and the accuracy reached over 95% after 100 epochs. In terms of mIOU, the addition of the CA model improved the efficacy of the model by 1.37%. In terms of evaluating the Kappa coefficient, the addition of the CA model improved the accuracy of monitoring changes by 1.29%. Replacing traditional convolution with DSC reduced the total FLOPs to about one-third of the original model. When the CNN in the optimized Transformer model was replaced by the Ghost module, it could lessen the number of parameters in the neural network and decrease computational complexity. The improved Transformer model had the least number of parameters and the lowest computational complexity, with a Kappa value of 0.795. Experimental results showed that this algorithm was more suitable for detecting building changes in SAR images.

By using DSC and Ghost modules to achieve model lightweighting and reduce the computational complexity and parameter count of the model, the proposed method becomes more feasible for deployment on resource constrained devices. For example, in urban planning, lightweight models can be integrated into mobile map systems to provide real-time updates on building changes, helping to quickly assess urban development or illegal construction activities. The reduction of computational overhead not only lowers deployment costs, but also speeds up inference time, making the model suitable for real-time applications. However, despite the advantages of SAR data in all-weather and all-day imaging capabilities, it is inherently noisy due to speckle interference. This type of noise can reduce the quality of input data, especially in densely populated urban areas where buildings are located in close proximity, leading to potential misclassification or missed detections. In addition, the performance of the model may be affected by extreme weather conditions such as heavy rain or snow, which may further blur SAR images and reduce the model's ability to accurately detect changes.

The current models rely on fully supervised learning, which requires a large amount of labeled data. Future work can explore weakly supervised or unsupervised learning techniques to reduce reliance on labeled data. For example, self-supervised learning methods can be used to pre-train the model on a large amount of unlabeled SAR data, and then fine tune it on a smaller labeled dataset. Although the model has been optimized for SAR images, future research can explore its adaptability to other types of remote sensing data, such as optical images or LiDAR. Combining data from multiple sensors can provide complementary information and improve the model's ability to detect environmental changes in different cities.

In summary, the proposed Transformer-UNet model has been enhanced through coordinated attention mechanism and lightweight technology, providing a robust solution for urban building change detection using SAR images. The high accuracy and reduced

computational complexity of this model make it highly suitable for practical applications in urban planning and disaster monitoring.

Availability of data and materials

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The author declares that there is no conflict of interest.

Declaration of generative AI in scientific writing

The author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

1. Cao, H., Wu, Y., Bao, Y., Feng, X., Wan, S., Qian, C. UTrans-Net: A Model for Short-Term Precipitation Prediction. Artificial Intelligence and Applications, 2023, 1(2), 106-113. https://doi.org/10.47852/bonviewAIA2202337
2. Falegari, S., Shirzadi Javid, A. A. Integrating Building Information Modeling and Life Cycle Assessment to Analyze the Role of Climate and Passive Design Parameters in Energy Consumption. Energy and Environment, 2024, 35(4), 2087-2106. https://doi.org/10.1177/0958305X221145923
3. Gibril, M. B. A., Al-Ruzouq, R., Shanableh, A., Jena, R., Bolcek, J., Shafri, H. Z. M., et al. Transformer-Based Semantic Segmentation for Large-Scale Building Footprint Extraction from Very-High Resolution Satellite Images. Advances in Space Research, 2024, 73(10), 4937-4954. https://doi.org/10.1016/j.asr.2024.01.012
4. Huo, H., Yu, Y. L., Liu, Z. H. Facial Expression Recognition Based on Improved Depthwise Separable Convolutional Network. Multimedia Tools and Applications, 2023, 82(12), 18635-18652. https://doi.org/10.1007/s11042-022-14066-6
5. Ismail, Z. A. Improving Failure Risk by Better Planning and Safety for Precast Beam-to-Column Connection Elements Using Physical Internet-Enabled Building Information Modeling Technology: A Malaysian Case Study. International Journal of Building Pathology and Adaptation, 2023, 41(3), 517-532. https://doi.org/10.1108/IJBPA-08-2021-0104
6. Liang, S., Hua, Z., Li, J. Hybrid Transformer-CNN Networks Using Superpixel Segmentation for Remote Sensing Building Change Detection. International Journal of Remote Sensing, 2023, 44(8), 2754-2780. https://doi.org/10.1080/01431161.2023.2208711
7. Liu, Y., Jiang, H., Nader, G., Wu, Z., Tanasnitikul, T., Lasang, P. Local Topology Constrained Point Cloud Registration in Building Information Modelling. IEEE Sensors Journal, 2023, 24(3), 4036-4046. https://doi.org/10.1109/JSEN.2023.3341218
8. Ni, K., Yuan, C., Zheng, Z., Zhang, B., Wang, P. MPT-SFANet: Multi-Order Pooling Transformer-Based Semantic Feature Aggregation Network for SAR Image Classification. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60(4), 4923-4938. https://doi.org/10.1109/TAES.2024.3382622
9. Omaran, S. M., Al-Zuheriy, A. S. J. Integrating Building Information Modeling and Virtual Reality to Develop Real-Time Suitable Cost Estimates Using Building Visualization. International Journal of Engineering, 2023, 36(5), 858-869. https://doi.org/10.5829/IJE.2023.36.05B.12
10. Rajkumar, R., Shanthi, D., Manivannan, K. Efficient Guided Grad-CAM Tuned Patch Neural Network for Accurate Anomaly Detection in Full Images. Information Technology and Control, 2024, 53(2), 355-371. https://doi.org/10.5755/j01.itc.53.2.34525
11. Schacht, G., Fritsch, C., Voigt, C., Ewert, E., Arndt, R. Structural Information Modeling - The Digital Transformation of Building Diagnostics. Bautechnik, 2022, 99(3), 213-221. https://doi.org/10.1002/bate.202100121
12. Sreekumar, S. P., Palanisamy, R., Swaminathan, R. An Approach to Segment Nuclei and Cytoplasm in Lung Cancer Brightfield Images Using Hybrid Swin-Unet Transformer. Journal of Medical and Biological Engineering, 2024, 44(3), 448-459. https://doi.org/10.1007/s40846-024-00873-9
13. Sun, Y., Zhao, Y., Han, X., Gao, W., Hu, Y., Zhang, Y. A Feature Enhancement Network Combining UNet and Vision Transformer for Building Change Detection in High-Resolution Remote Sensing Images. Neural Computing and Applications, 2025, 37(3), 1429-1456. https://doi.org/10.1007/s00521-024-10666-5

14. Wahba, M., Sharaan, M., Elsadek, W. M., Kanae, S., Hassan, H. S. Building Information Modeling Integrated with Environmental Flood Hazard to Assess the Building Vulnerability to Flash Floods. Stochastic Environmental Research and Risk Assessment, 2024, 38(2), 503-520. https://doi.org/10.1007/s00477-023-02640-9

15. Wang, J., Lin, T., Zhang, C., Peng, J. FIBTNet: Building Change Detection for Remote Sensing Images Using Feature Interactive Bi-Temporal Network. Computers, Materials and Continua, 2024, 80(3), 4621-4641. https://doi.org/10.32604/cmc.2024.053206

16. Yin, L., Wang, L., Lu, S., Wang, R., Yang, Y., Yang, B., Liu, S., AlSanad, A., AlQahtani, S. A., Yin, Z., Li, X., Chen, X., Zheng, W. Convolution-Transformer for Image Feature Extraction. CMES - Computer Modeling in Engineering and Sciences, 2024, 141(1), 2-10. https://doi.org/10.32604/cmes.2024.051083

17. Yiming, T., Tang, X., Shang, H. A Shape-Aware Enhancement Vision Transformer for Building Extraction from Remote Sensing Imagery. International Journal of Remote Sensing, 2024, 45(4), 1250-1276. https://doi.org/10.1080/01431161.2024.2307325

18. Yu, B., Chen, F., Wang, N., Yang, L., Yang, H., Wang, L. MSFTrans: A Multi-Task Frequency-Spatial Learning Transformer for Building Extraction from High Spatial Resolution Remote Sensing Images. GIScience and Remote Sensing, 2022, 59(1), 1978-1996. https://doi.org/10.1080/15481603.2022.2143678

19. Yuan, Q., Xia, B. Cross-Level and Multiscale CNN-Transformer Network for Automatic Building Extraction from Remote Sensing Imagery. International Journal of Remote Sensing, 2024, 45(9), 2893-2914. https://doi.org/10.1080/01431161.2024.2339199

20. Zhang, C., Wang, J., Shi, Y., Yin, B., Ling, N. A CNN-Transformer Hybrid Network with Selective Fusion and Dual Attention for Image Super-Resolution. Multimedia Systems, 2025, 31(2), 1-17. https://doi.org/10.1007/s00530-025-01711-x

21. Zhou, X., Bu, Q., Matskevich, V. V., Nedzved, A. M. Detection System of Landscape's Unnatural Changes by Satellite Images Based on Local Areas. Pattern Recognition and Image Analysis, 2024, 34(2), 365-378. https://doi.org/10.1134/S1054661824700159

22. Zhou, Y., Yang, K., Ma, F., Hu, W., Zhang, F. Water-Land Segmentation via Structure-Aware CNN-Transformer Network on Large-Scale SAR Data. IEEE Sensors Journal, 2023, 23(2), 1408-1422. https://doi.