Journal of New Media

Tech Science Press

https://doi.org/10.32604/jnm.2026.088090

ARTICLE

Efficient Deepfake Geolocation Detection via Comprehensive Features

Xiaochen Hao[1,2], **Wei Liu**[3,*], **Yuhuan Liu**[4] **and Zheng Kong**[5]

[1]College of Computer Science and Engineering, Guilin University of Technology, Guilin, China
[2]Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin, China
[3]School of Cyberspace Security, Guangdong Polytechnic Normal University, Guangzhou, China
[4]School of Computer Science and Engineering, Macau University of Science and Technology, Macau, China
[5]School of Cyberspace Security, Guangzhou University, Guangzhou, China

*Corresponding Author: Wei Liu. Email: weiliu823@163.com

Received: 28 June 2026; Accepted: 17 September 2026; Published: 12 October 2026

ABSTRACT: In recent years, deepfake technology has increasingly extended to satellite imagery, raising serious concerns regarding the authenticity and reliability of geospatial information. To balance localized forgery artifact capture, global semantic modeling, and model computational efficiency, this paper proposes an efficient cascaded fake satellite image detection model based on comprehensive features. Specifically, a depthwise separable self-calibrated convolution module is first designed. It replaces conventional convolutions with depthwise and pointwise convolutions, reducing parameter volume and computational complexity while boosting the capability of extracting local detailed features. Subsequently, a comprehensive feature extraction module is constructed, incorporating multi-head self-attention and adaptive attention mechanisms to model global dependencies and strengthen critical local features for Deepfake Geolocation Detection. Experimental results demonstrate that the proposed method achieves an F1 score of 0.988 on the public University of Washington deepfake satellite image dataset (UW dataset) with merely 7.66 million (M) parameters, and achieves a favorable balance between detection performance and model size, indicating potential for resource-constrained deployment. Further tests under common degradation scenarios including Joint Photographic Experts Group (JPEG) compression, Gaussian noise addition, and blurring verify the robustness of the proposed method, which validates its reliability and practicability under various harsh conditions.

KEYWORDS: Digital forensics; deepfake geography; fake satellite image detection; multi-head self-attention mechanism; comprehensive feature extraction

1 Introduction

With the continuous innovation of artificial intelligence technologies and the substantial improvement of hardware computing power, deep learning algorithms [1] can now process more complex and larger-scale datasets and perform increasingly sophisticated computational tasks. This progress has greatly promoted the development of artificial intelligence-generated content (AIGC). In particular, high-performance image generation methods, such as generative adversarial networks (GANs) [2] and denoising diffusion probabilistic models (DDPMs) [3], have created extensive possibilities for high-quality generated content. However, advances in AIGC have also accelerated the rapid development of deepfake technology [4], posing serious threats to politics, the economy, and national security [5]. For example, in 2022, a fabricated video showing Ukrainian President Volodymyr Zelenskyy surrendering was widely circulated. Although it was eventually debunked, it had already caused social unrest within a short period. As deepfakes become increasingly

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

prevalent, the misuse of these technologies is no longer merely a prank or a form of entertainment, but has evolved into serious criminal behavior; accordingly, the problems associated with generative artificial intelligence (AI) have gradually shifted toward risk governance. Active forensics of deepfake content has therefore become a major focus for governments, enterprises, and researchers worldwide. At present, deepfake technology is gradually extending into the field of satellite imaging, giving rise to a new problem known as "deepfake geography" [6]. Satellite images are widely used in environmental monitoring, disaster warning, national defense, and many other fields [7–9]. Forged satellite images may be used to spread false information, facilitate illegal activities, or interfere with military decision-making, thus creating substantial security risks.

The main contributions of this study are summarized as follows:

(1) A depthwise separable self-calibrated convolution (DSSC) module is designed by decomposing conventional convolution operations into depthwise and pointwise convolutions. This design aims to preserve the local feature extraction capability of self-calibrated convolution while reducing parameter redundancy.
(2) An integrated feature extraction module (IFEM) is developed by cascading multi-head self-attention (MHSA) with adaptive attention. MHSA is used to model long-range dependencies, while adaptive attention further emphasizes informative feature responses.
(3) A cascaded local-to-global detection architecture is constructed by sequentially connecting DSSC and IFEM. The architecture avoids directly parallelizing convolutional neural network (CNN) and Transformer branches and achieves a favorable balance between detection performance and model size.
(4) Extensive experiments on the UW dataset and several representative image degradation conditions demonstrate the effectiveness and robustness of the proposed architecture, while the ablation experiments verify the contribution of the DSSC redesign and the integrated feature extraction stage.

2 Related Work

False satellite images generated by AIGC are highly realistic and are almost indistinguishable to the naked eye. Although progress has been made in deepfake detection, research on detecting deepfake satellite images remains limited because of the lack of relevant datasets. To address this problem, Zhao et al. [6] proposed a method based on a cycle-consistent generative adversarial network (CycleGAN) [10] for generating fake satellite images and created the first UW dataset for deepfake satellite images. In this dataset, fake satellite images are generated by CycleGAN from satellite images of Tacoma, Seattle, and Beijing; real satellite images are collected from Google Earth. Several samples from the UW dataset are shown in Fig. 1. More recently, Ding et al. [11] provided a comprehensive survey of satellite image forensics, summarizing recent advances in forgery detection and localization. Aslan and Özüpak [12] also demonstrated the effectiveness of deep learning for extracting complex spatial features from satellite images.

Meanwhile, based on the theory found in earlier studies that images generated by GANs differ from real images in visual features such as color, texture, and detail, as well as in frequency-domain features, the team selected 26 handcrafted features from three categories: color space, histogram, and frequency, using independent t-tests to identify significant mean differences between all real and forged satellite images. These features were then fed into a support vector machine (SVM) to detect fake satellite images. However, detection methods based on traditional handcrafted features are often unable to cope with increasingly complex and diverse forgery techniques [13]. Therefore, Fezza et al. [14] attempted to detect forged satellite images using convolutional neural networks (CNNs), employing four CNN models pretrained on ImageNet: the 50-layer residual network (ResNet50) [15], the 16-layer Visual Geometry Group network (VGG16) [16], InceptionV3 [17], and Xception [18]. Experiments were conducted to evaluate the performance

and feasibility of these models for forged satellite image detection. Nevertheless, pure CNNs are constrained by local receptive fields and have limited ability to model global context. Zhao et al. [19] recently explored Transformer-based model fusion for multi-scenario remote sensing image forgery detection, demonstrating the potential of global feature modeling. In contrast, Transformer models based on self-attention are good at integrating global information and capturing long-range dependencies, but they suffer from high computational complexity and insufficient inductive capability for local details.

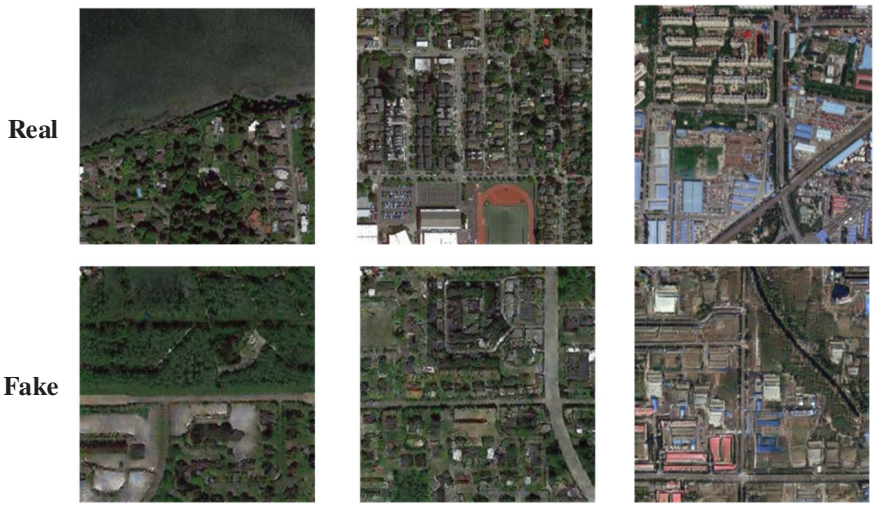


Figure 1: Some samples from the UW dataset.

In deepfake geography detection, local forgeries are often manifested as subtle signals such as texture anomalies and abrupt edge changes, whereas global forgeries may appear as inconsistent semantic relationships among ground objects or distortions in the overall structure. A single model is therefore difficult to use for both types of evidence. Although simply parallelizing CNNs and Transformers can combine the strengths of both, it sharply increases computational cost and is not conducive to future mobile deployment or real-time detection. To address this issue, this paper designs an efficient CNN-Transformer cascaded model. The rich local semantic features extracted by the CNN are sequentially passed to the Transformer for global relation modeling, which improves detection accuracy while substantially reducing computational overhead. In this way, the proposed model achieves a balance between local features and global semantics, as well as between accuracy and efficiency.

Recent studies have increasingly investigated hybrid CNN-Transformer architectures for visual representation learning, aiming to combine the local inductive bias of convolutional networks with the global dependency modeling capability of Transformer-based architectures. Lightweight Transformer variants have also been developed to reduce computational overhead and improve their suitability for resource-constrained applications. These studies motivate the use of a cascaded CNN-Transformer architecture in the present work. Unlike architectures that directly parallelize CNN and Transformer branches, the proposed method adopts a sequential local-to-global strategy, in which DSSC first extracts efficient local representations and IFEM subsequently performs global dependency modeling and adaptive feature refinement.

3 Related Techniques

3.1 Self-Calibrated Convolutional Neural Network

The classification performance of downstream tasks depends on the feature extraction capability of the upstream convolutional network. The self-calibrated convolutional neural network (Self-Calibration Net, SCNet) [20] improves model performance by enhancing basic convolutional modules without modifying the overall network architecture. Based on the residual structure, SCNet designs a plug-and-play self-calibration module (SCModule) to replace conventional convolutional blocks. Unlike traditional two-dimensional (2D) convolution, the self-calibration (SC) module evenly splits the input data and feeds them into two branches. The upper branch is a self-calibration branch, which performs mean downsampling, convolutional feature transformation, bilinear downsampling, and spatial-attention feature-map fusion on the input data. Because the self-calibration branch considers only spatial positional information and avoids interference from irrelevant information, it can greatly enlarge the receptive field of the output features. The other part of the input data is fed into the lower branch, which is a conventional convolutional transformation branch that preserves the background relationships and context of the data. Finally, the processed features are concatenated as the final output. The receptive field of each spatial position in the SC module is not controlled by a predefined convolution-kernel size. Without introducing additional parameters, it helps the model extract features more completely and accurately and enhances the discriminability of the extracted features. Based on these advantages, this paper selects SCModule as the baseline and improves it.

3.2 Multi-Head Self-Attention Mechanism (MHSA)

The multi-head self-attention mechanism is an effective feature-modeling approach that captures dependencies among different spatial positions, provides global contextual information, and serves as a core component of Transformer architectures [21]. The principle of multi-head self-attention is shown in Fig. 2:

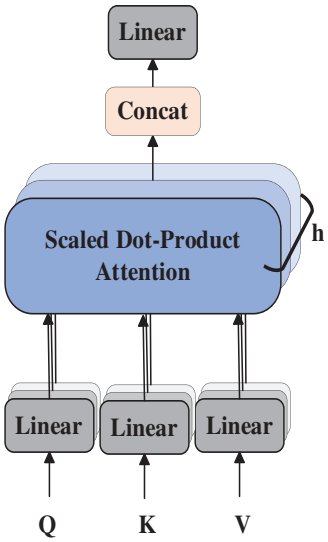


Figure 2: The multi-head self-attention mechanism.

The mechanism first reshapes the input feature map into a form suitable for attention computation and generates query (Q), key (K), and value (V) vectors through linear transformations. The calculation is shown in Eq. (1):

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{1}$$

Then, Q and K are multiplied by dot product to compute attention scores, which are scaled to reduce numerical instability. The Softmax function is then used to convert the attention scores into weights, indicating the importance of each position in the feature map. By multiplying the attention weights with V, the weighted output is obtained, thereby capturing the relationships among different positions in the feature map, as shown in Eq. (2), where d_k is the dimension of the key vector:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2}$$

Finally, through the mapping of the linear layer, the output is restored to the original channel dimension, as shown in Eq. (3):

$$MHSA(X) = Concat\left(head_1, \ldots, head_h\right)W_O \tag{3}$$

This multi-head mechanism enables the model to learn multiple attention features in parallel, thereby improving the richness and accuracy of feature extraction.

4 Model Construction

4.1 Depthwise Separable Self-Calibrating Convolution (DSSC)

Building a complete feature extraction channel requires the cascading of multiple self-calibrating convolutions, which multiplies the number of parameters and makes the model unsuitable for resource-constrained environments. To optimize this problem, this paper designs a depthwise separable self-calibrating convolution (DSSC), which adopts a decomposed-convolution strategy by decomposing the standard convolution operations in SCModule into depthwise convolution and pointwise convolution. The improvement strategy is shown in Fig. 3:

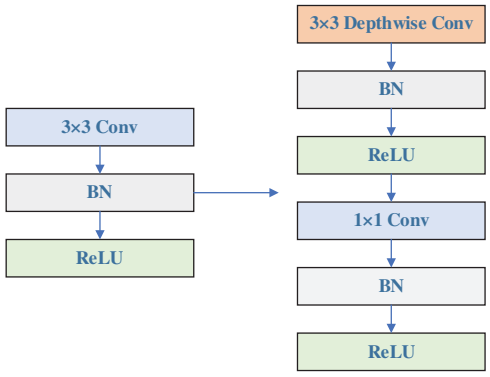


Figure 3: The principle of the improvement strategy.

Depthwise convolution is first introduced to perform convolution independently for each input channel. Specifically, depthwise convolution applies an independent spatial kernel to each input channel, such that each output channel depends only on its corresponding input channel. This operation significantly reduces the number of parameters and computational cost, because each kernel interacts only with its corresponding channel and does not involve information from other channels. The convolution operation is simplified into independent convolution over each input channel; each output channel depends only on the corresponding

input channel, and the convolution kernel W is associated with only one input channel. The definition of depthwise convolution is given in Eq. (4):

$$Y_{o,c} = \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} X_{o+i,c+j} \times W_{i,j,c} \tag{4}$$

Next, pointwise convolution with a 1 × 1 kernel is used to fuse the output of depthwise convolution. It implements a linear combination of feature maps, aggregates the channel information generated by depthwise convolution, and recomposes new feature maps. The definition is shown in Eq. (5), where Z is the final output feature map, C_{in} is the number of input channels.

$$Z_o = \sum_{c=0}^{C_{in}-1} Y_{o,c} \cdot W_{c,k} \tag{5}$$

In summary, depthwise convolution reduces the parameter count of SCModule through independent channel-wise convolution, whereas pointwise convolution is used for cross-channel feature fusion to ensure full utilization of the channel information output by each path. Through this decomposition, the model maintains accuracy while achieving higher computational efficiency.

4.2 Adaptive Attention Mechanism (Ada)

Adaptive feature extraction is the core idea of the SE (Squeeze-and-Excitation, SE) module [22]. It is intended to solve the problem of low feature extraction efficiency caused by the use of convolutional and pooling layers in traditional CNN architectures for image feature extraction. Referring to the SE module, this paper adopts an adaptive attention mechanism (Adaptive Attention, Ada) to extract local features, aiming to dynamically modulate different regions of the input feature map and thereby enhance the model's focus on important features.

During forward propagation, the input feature map is first processed by the first 1 × 1 convolutional layer, and the generated intermediate feature map is activated by the rectified linear unit (ReLU) function. Subsequently, the merged feature map is processed by the second 1 × 1 convolutional layer to generate an attention feature map with the same number of channels as the input feature map. To ensure the validity of attention weights, the Sigmoid activation function is then applied to this feature map, compressing it into the range $[0, 1]$ to form attention weights and modulate the original feature map. Finally, the original feature map is multiplied element-wise by the calculated attention weights to generate the weighted output feature map. This weighting operation enables the model to dynamically attend to important features, effectively improving its sensitivity to key features and promoting the accuracy of feature extraction.

Fig. 4 shows the structure of adaptive attention and clearly illustrates the layer-by-layer transformation of the feature map. In terms of data flow, the input feature map sequentially undergoes convolution, activation, a second convolution, and Sigmoid processing.

The specific calculation is shown in Eq. (6):

$$A = \sigma\left(W_2 \cdot ReLU\left(W_1 \cdot X\right)\right) \tag{6}$$

where W_1 and W_2 are the parameters of the 1 × 1 convolution kernels, σ denotes the Sigmoid activation function, and the generated attention-weight matrix is A.

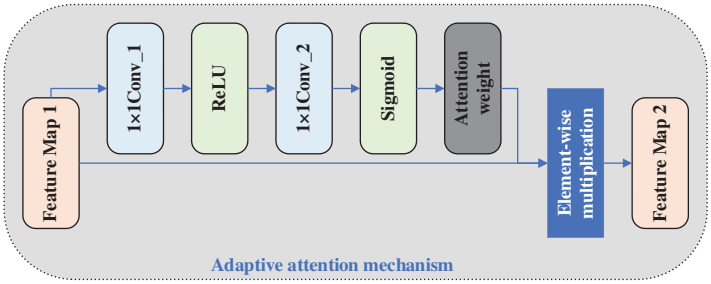


Figure 4: Adaptive attention.

Finally, it is multiplied by the original input to weight the input features and form the output, as shown in Eq. (7):

$$X_{\text{out}} = X \cdot A \tag{7}$$

Through the above design, the adaptive attention mechanism effectively integrates local information and global context, providing a more detailed and richer representation for feature extraction and helping improve the performance of deep learning models in various tasks.

4.3 Integrated Feature Extraction Module (IFEM)

In deepfake geography detection, both the local feature extraction capability of CNNs and the global information integration capability of Transformers are required. To improve the extraction capability of comprehensive features and enhance computational efficiency, this paper designs an Integrated Feature Extraction Module (IFEM) to make full use of comprehensive features. The core idea is to introduce multi-head self-attention, the core idea of Transformer, into CNNs and cascade it with an adaptive attention mechanism. This enhances global feature extraction while further refining local features. Finally, the feature vector is output through an average-pooling operation. The structure of IFEM is shown in Fig. 5.

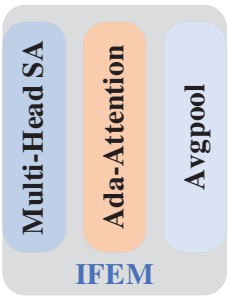


Figure 5: Integrated feature extraction module.

The proposed framework follows a local-to-global processing paradigm. DSSC first extracts local forgery traces and structural inconsistencies from the input image. The resulting high-level feature representation is then passed to IFEM, where MHSA models long-range dependencies and adaptive attention further emphasizes task-relevant features. Therefore, DSSC and IFEM are sequentially connected and perform complementary feature-extraction operations.

First, the multi-head self-attention mechanism captures long-range dependencies among features from a global perspective, enabling the model to better understand the overall structure and contextual

information of the input data. This stage enhances the model's ability to attend to different feature dimensions by processing multiple attention heads in parallel. Subsequently, the adaptive attention mechanism weightedly modulates the features processed by multi-head self-attention, emphasizing important features and suppressing unimportant ones, thereby achieving a more refined feature representation. The overall calculation is shown in Eq. (8), where MHSA(X) is the global feature obtained by multi-head self-attention, and AdaptiveAttention (MHSA(X)) is the feature further weighted by the adaptive attention mechanism.

$$X_{out} = AdaptiveAttention\,(MHSA(X)) \tag{8}$$

The advantage of this design lies in capturing global features through multi-head self-attention while maintaining sensitivity to local features. At the same time, the adaptive attention mechanism further optimizes the feature map, allowing the model to dynamically focus on information relevant to the final task. This local-to-global processing workflow makes feature extraction not only efficient but also highly expressive, laying a solid foundation for subsequent classification or detection tasks.

4.4 Overall Architecture of the Cascaded Detection Model

This paper designs an efficient detection model for deepfake geography by cascading the depthwise separable self-calibrating convolution module with the integrated feature extraction module. The model consists of three parts: data processing, the feature extraction channel, and fully connected classification.

In the data processing part, random cropping is used to crop the input data to 224 × 224, thereby increasing the diversity of the dataset.

The feature extraction channel is composed of one standard convolutional layer, three depthwise separable self-calibrating convolutional layers, and one integrated feature extraction module in cascade. The depthwise separable self-calibrating convolutional layers use the ReLU activation function to introduce nonlinearity into the output and use batch normalization (Batch_Norm) to stabilize the network, prevent overfitting, accelerate convergence, and improve network performance. The image in the input feature channel is first converted into a 64 × 64 image using the Transform. Resize function and then passed through 64 filters of size 7 × 7 with a stride of 2. After feature representation is extracted, downsampling is performed through the max-pooling layer. To match the input data dimensions and help the pooling layer sample more effectively, the stride of the max-pooling layer (Max_Pool) is changed to 1. Next, feature representations are extracted sequentially through Layer1, a depthwise separable self-calibrating convolutional layer containing one DSSC module; Layer2, a depthwise separable self-calibrating convolutional layer containing one DSSC module; and Layer3, a depthwise separable self-calibrating convolutional layer containing two DSSC modules. The extracted feature maps are then fed into the IFEM module for comprehensive feature extraction, and a 1024 × 1 × 1 feature vector is output after average pooling.

In the cascaded feature extraction channel, DSSC is positioned before IFEM. The hierarchical DSSC layers progressively extract local and mid-level representations, while the subsequent IFEM operates on the resulting high-level feature map to establish global relationships and perform adaptive feature refinement.

The fully connected classification layer receives the feature map from the feature extraction channel. In this layer, the feature vector is mapped through two fully connected layers, reducing its dimension from 1024 to 512 and finally outputting the prediction results for two classes. The complete structure of the proposed method is shown in Fig. 6.

Although the proposed architecture integrates DSSC and IFEM, these modules are designed to provide complementary rather than redundant feature representations. DSSC focuses on efficient extraction of local and fine-grained forgery traces, while IFEM is applied once at the high-level feature stage to model global

dependencies and adaptively refine discriminative features. Therefore, the proposed architecture follows a lightweight cascaded local-to-global feature extraction strategy rather than simply increasing the number of network components.

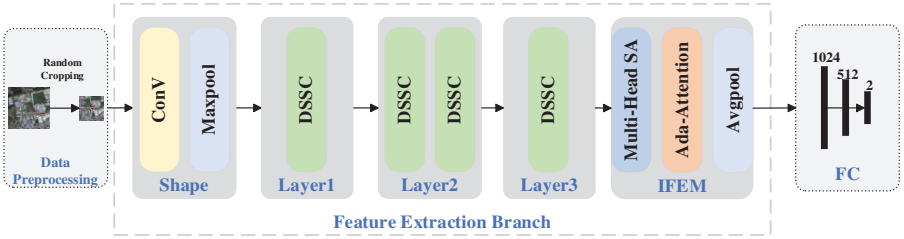


Figure 6: Cascaded forged satellite image detection model.

The advantage of the cascaded model based on comprehensive features is that, by combining convolutional neural networks with attention mechanisms, it aims to achieve higher efficiency and precision in feature extraction. Depthwise separable self-calibrating convolution reduces computational cost, while the integrated feature module enhances the model's representation capability for complex features through global and adaptive feature modeling.

5 Experiments and Result Analysis

To evaluate the effectiveness of the proposed method, we compare it with several representative existing methods. These methods include the handcrafted-feature-based method proposed by Zhao et al. and CNN methods based on transfer learning (TL), including VGG16, InceptionV3, and Xception. In addition, to verify that the proposed method can achieve an effective balance between performance and parameter count, transfer-learning-based Vision Transformer (ViT; base variant, ViT-B) [23] and MobileNetV2 [24] methods are added for comparison. Furthermore, to evaluate the robustness of the model when processing images with different post-processing operations, further experiments are conducted on test conditions: Gaussian blur, JPEG compression, and Gaussian noise.

5.1 Experimental Environment Settings

The current experiments are conducted on the public UW dataset to maintain consistency with existing deepfake satellite image detection studies. Although this setting enables a fair comparison among different methods, evaluation on additional satellite image datasets is still required to more comprehensively assess cross-dataset generalization.

The experimental configuration was as follows: the central processing unit (CPU) configuration was dual Intel(R) Silver 4210R @ 2.40 GHz/2.39 GHz processors, and the graphics processing unit (GPU) was an RTX 3060 12 GB graphics card. The training and testing platform was Python 3.8 with the PyTorch deep learning library.

The experimental settings were as follows: during training and validation, the learning rate was set to 0.001 and the batch size to 32; during testing, the batch size was set to 64. This paper uses the public UW dataset released by Zhao et al., which contains 8064 satellite images. Among them, 4032 are real 256 × 256 color satellite images, and 4032 are fake 256 × 256 color satellite images generated by CycleGAN. The UW dataset was divided into training and testing sets at a ratio of 9:1. To reduce the influence of random variation during evaluation, the testing procedure was repeated for 20 rounds, and the final performance results

were calculated as the average values across the repeated evaluations. The results were retained to three decimal places.

5.2 Evaluation Metrics

For the image binary classification task, the three most widely used performance metrics are adopted to evaluate the model: Precision (P), Recall (R), and F1-score (F). The definitions of these three evaluation metrics are shown in Eq. (9):

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ F1 = 2 \times \dfrac{Precision \times Recall}{Precision + Recall} \end{cases} \tag{9}$$

where true positive (*TP*) denotes the number of forged satellite images correctly predicted as forged, false positive (*FP*) denotes the number of real satellite images incorrectly predicted as forged, and false negative (*FN*) denotes the number of forged satellite images incorrectly predicted as real. True negative (*TN*) denotes the number of real satellite images correctly predicted as real.

5.3 Comparative Experimental Results

5.3.1 Model Performance Comparison Experiment

The experimental results on the UW test set are shown in Table 1, and the statistical comparison chart is shown in Fig. 7.

Table 1: Experimental results on the original dataset.

Method	P	R	F1
Zhao.	0.827	0.919	0.871
InceptionV3 TL	0.927	0.926	0.926
VGG16 TL	0.965	0.965	0.965
Xception TL	0.953	0.953	0.953
MobileNetV2 TL	0.969	0.968	0.968
ViT-B TL	0.985	0.985	0.985
Ours	0.990	0.987	0.988

The experimental results show that the method of Zhao et al. performs poorly in forged satellite image detection because its handcrafted features are limited in complexity and representation capability. In contrast, CNN models based on transfer learning, such as VGG16 and MobileNetV2, have relatively strong feature extraction capability. The ViT-B model also performs well in feature extraction and classification accuracy. The detection model proposed in this paper, which cascades depthwise separable self-calibrating convolution with the integrated feature extraction module, achieves outstanding performance in the forged satellite image detection task. Compared with other methods, the proposed method combines global and local feature extraction to achieve more refined and comprehensive feature extraction. As a result, it outperforms other methods on all evaluation metrics and obtains an F1-score of 0.988.

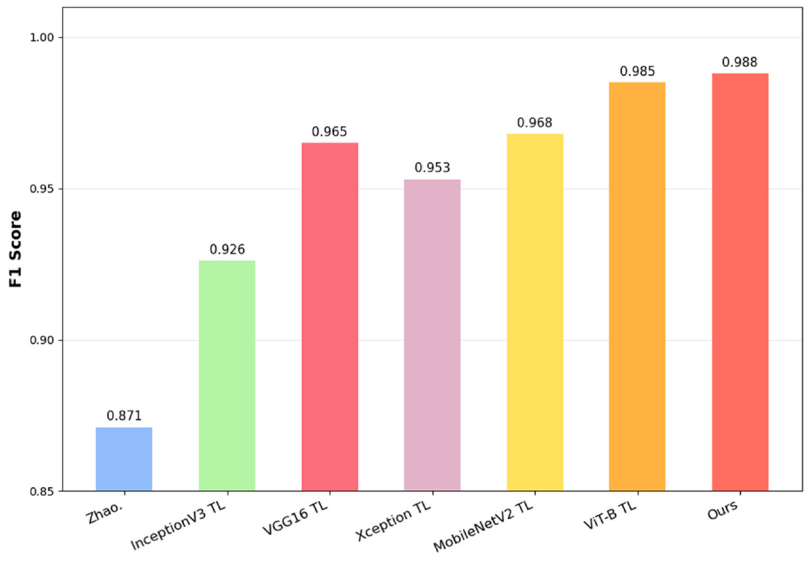


Figure 7: Experimental results on the original dataset.

5.3.2 Robustness Comparison Experiments

(1) Comparative experiment on the JPEG compressed dataset

During image transmission, data usually undergo JPEG compression. JPEG compression balances image quality and coding bit rate by adjusting the quality factor (QF), thereby controlling the relationship between image quality and file size during compression. QF is usually an integer ranging from 1 to 100. A higher QF indicates a lower compression ratio, higher image quality, and a larger file size; conversely, a lower QF indicates a higher compression ratio. Although the file size is reduced, image quality decreases significantly and distortion increases. To evaluate the influence of compression distortion on model robustness, QF values of 75, 85, and 95 were used to apply JPEG compression to the test images, which were divided into three groups for experiments. The experimental results are shown in Table 2, and the comparison line chart is shown in Fig. 8.

Table 2: Experimental results on the JPEG compressed dataset.

Factor	Method	P	R	F1
75	Zhao.	0.884	0.920	0.902
	InceptionV3 TL	0.892	0.880	0.879
	VGG16 TL	0.958	0.958	0.958
	Xception TL	0.913	0.911	0.911
	MobileNetV2 TL	0.951	0.948	0.948
	ViT-B TL	0.907	0.887	0.886
	Ours	0.961	0.992	0.976
85	Zhao.	0.858	0.936	0.895
	InceptionV3 TL	0.905	0.900	0.899
	VGG16 TL	0.960	0.960	0.960
	Xception TL	0.941	0.941	0.941
	MobileNetV2 TL	0.967	0.967	0.967

(Continued)

Table 2 (continued)

Factor	Method	P	R	F1
	ViT-B TL	0.931	0.922	0.922
	Ours	0.977	0.990	0.983
95	Zhao.	0.850	0.889	0.869
	InceptionV3 TL	0.924	0.923	0.923
	VGG16 TL	0.962	0.962	0.962
	Xception TL	0.946	0.946	0.946
	MobileNetV2 TL	0.968	0.968	0.968
	ViT-B TL	0.959	0.956	0.956
	Ours	0.985	0.985	0.985

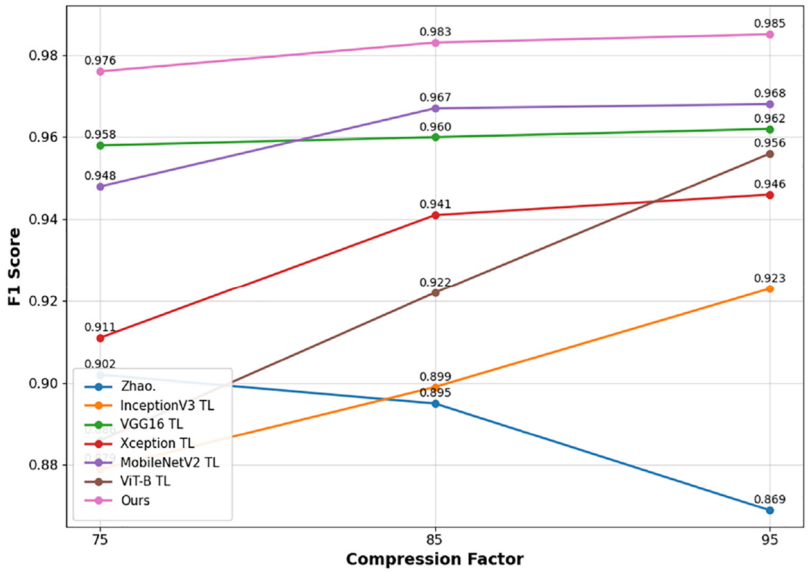


Figure 8: Experimental results on the JPEG compressed dataset.

By introducing depthwise separable self-calibrating convolution and the integrated feature extraction module, the model designed in this paper significantly improves the depth and breadth of feature extraction. The experimental results show that under test conditions with different compression quality factors, the proposed method consistently performs well and surpasses several existing detection models, achieving the highest Precision, Recall, and F1-score.

The comparison line chart intuitively shows that when QF = 75, the F1-score of the proposed method reaches 0.976, which is clearly higher than those of Zhao et al.'s method and the transfer-learning-based InceptionV3 and Xception methods. In the experiment with QF = 85, the F1-score of the proposed method reaches 0.983, outperforming methods including MobileNetV2 and ViT-B. When QF = 95, the proposed method maintains near-perfect detection capability, with an F1-score of 0.985, exceeding that of other methods. These results indicate that the proposed model performs well not only under lower-compression conditions but also demonstrates strong adaptability and robustness in high-compression environments.

This strategy of combining global and local features makes the model more flexible and efficient in capturing subtle changes in forged satellite images, significantly improving detection accuracy and stability.

(2) Comparative experiment on the Gaussian noise dataset

During the acquisition and transmission of satellite images, Gaussian noise may be introduced because of the influence of device channels. The intensity of Gaussian noise is usually quantified and described by a noise factor, which reflects the degree of disturbance of the noise signal relative to the ideal signal and defines the distribution characteristics of the noise around the mean value. In Section 3, the principle of Gaussian noise was defined, explaining in detail that a larger standard deviation of the noise factor corresponds to a greater noise amplitude and thus has a stronger impact on high-frequency image features. This phenomenon poses a serious challenge to detection models. To evaluate the model performance under noise interference, a noise factor with a standard deviation of 0.02 was used for testing. The experimental results are shown in Table 3, and the statistical comparison chart is shown in Fig. 9.

Table 3: Experimental results on the Gaussian noise dataset.

Factor	Method	P	R	F1
0.02	Zhao.	0.814	0.851	0.832
	InceptionV3 TL	0.895	0.887	0.886
	VGG16 TL	0.948	0.946	0.946
	Xception TL	0.884	0.865	0.863
	MobileNetV2 TL	0.922	0.922	0.922
	ViT-B TL	0.857	0.815	0.809
	Ours	0.961	0.987	0.974

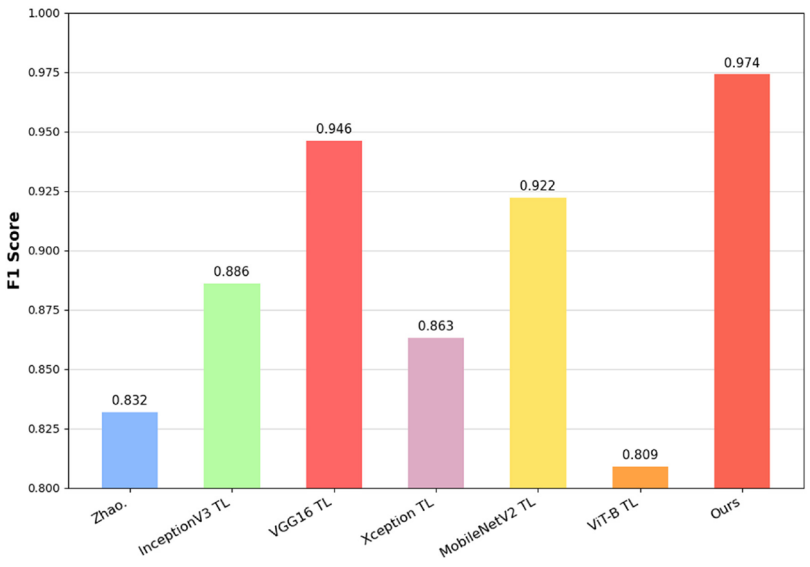


Figure 9: Experimental results on the Gaussian noise dataset.

The experimental results show that, although Gaussian noise was added to the test set, the proposed method still achieves the highest detection performance, indicating that the model has excellent robustness in noisy environments. When the noise factor is 0.02, the F1-score of the proposed model reaches 0.974.

Compared with Zhao et al.'s method and transfer-learning-based models such as InceptionV3, VGG16, and MobileNetV2, the proposed method maintains superior performance.

(3) Comparative experiment on the Gaussian blur dataset

In addition to issues occurring during acquisition and transmission, satellite images may also suffer from blurring due to unstable external environmental conditions, such as satellite motion or movement of the acquisition equipment. Such blurring can lead to the loss of important feature points, thereby affecting the performance of detection models. In Gaussian blur processing, the window refers to the pixel area used for the blurring operation, and its size and shape directly affect the intensity and range of the blurring effect. The window size is usually expressed as an odd number to ensure that there is always a clearly defined central pixel during blurring. A larger window includes more pixels in the blur calculation, resulting in a more pronounced blurring effect and a blurrier image. To evaluate the robustness of the model under blur-processing scenarios, Gaussian blur techniques with 5 × 5 and 9 × 9 windows were applied to the test set, and experiments were conducted. The results are shown in Table 4, and the comparison line chart is shown in Fig. 10.

Table 4: Experimental results on the Gaussian blur dataset.

Window	Method	P	R	F1
5 × 5	InceptionV3 TL [17]	0.876	0.865	0.864
	VGG16 TL [16]	0.911	0.958	0.901
	Xception TL [18]	0.880	0.856	0.853
	MobileNetV2 TL	0.952	0.952	0.952
	ViT-B TL	0.946	0.943	0.943
	Ours	0.989	0.892	0.938
9 × 9	InceptionV3 TL [17]	0.815	0.772	0.764
	VGG16 TL [16]	0.883	0.867	0.866
	Xception TL [18]	0.816	0.711	0.698
	MobileNetV2 TL	0.907	0.905	0.904
	ViT-B TL	0.912	0.910	0.909
	Ours	0.996	0.732	0.844

Under the 5 × 5 window blur processing, the Precision of the proposed method is significantly higher than those of other models, reaching 0.989 and demonstrating excellent detection precision. However, its Recall is only 0.892, which is relatively low and indicates a certain degree of missed detection. Nevertheless, the overall F1-score still reaches 0.938, showing the model's advantage in comprehensive performance. Under the 9 × 9 Gaussian blur condition, the proposed method achieves a Precision of 0.996, while Recall decreases to 0.732 and the corresponding F1-score decreases to 0.844. This result indicates that severe blur can substantially suppress local texture and structural cues that are useful for forgery detection. The proposed model therefore maintains a high level of prediction precision but may miss a proportion of forged samples under extreme degradation. Compared with MobileNetV2 TL and ViT-B TL, which obtain F1-scores of 0.904 and 0.909, respectively, the proposed model is less effective under this specific severe-blur condition. This result represents an important limitation of the current architecture and indicates that further improvement is required for extreme image degradation.

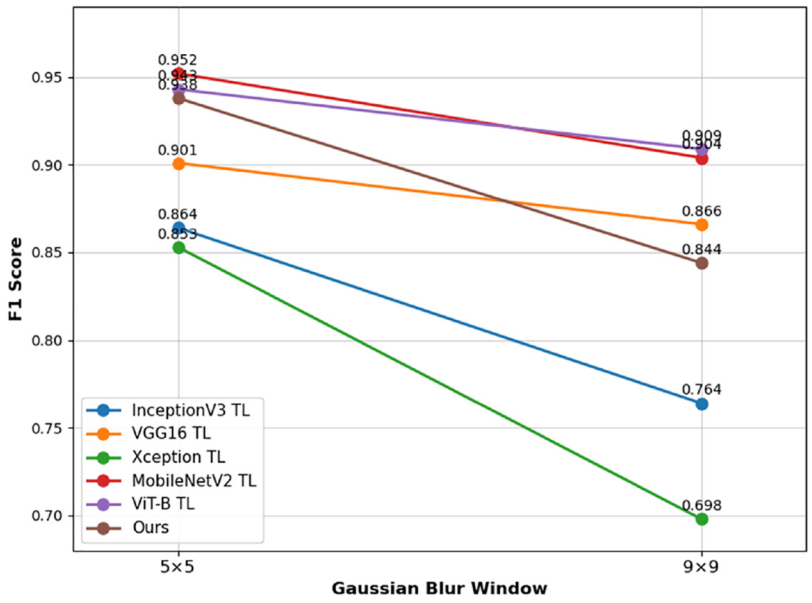


Figure 10: Experimental results on the Gaussian blur dataset.

The results show that the detection capability of the model weakens under larger-window blur processing, indicating that the parameter-reduction design of the proposed method leads to a certain loss of robustness. By comparison, under the 5 × 5 window blur processing, the Precision values of other methods are generally lower than that of the proposed model. In particular, the Precision values of InceptionV3 and Xception are only 0.876 and 0.880, respectively, showing their limitations in capturing detailed features. Under the 9 × 9 window blur processing, the performance of most models declines significantly. For example, the Recall values of InceptionV3 and Xception decrease to 0.772 and 0.711, respectively, indicating their sensitivity to blurred environments. It is worth noting that the high-parameter ViT-B method consistently maintains a leading F1-score.

5.3.3 Comparison of Model Parameter Counts and Model-Size Efficiency

To evaluate model-size efficiency, the parameter counts of the proposed method and the comparison models are compared. Parameter count provides an important indication of model size and storage requirements, although it does not fully characterize computational cost or hardware-dependent inference efficiency. The comparison between F1-score and parameter count is shown in Table 5, and a scatter plot is drawn with parameter count as the horizontal axis and F1-score as the vertical axis, as shown in Fig. 11.

Table 5: Parameter count comparison results.

Method	F1	Parameter Count (Para) (M)
InceptionV3 TL	0.926	21.8M
VGG16 TL	0.965	134.3M
Xception TL	0.953	20.8M
MobileNetV2 TL	0.968	2.2M
ViT-B TL	0.985	85.6M
Ours	0.988	7.66M

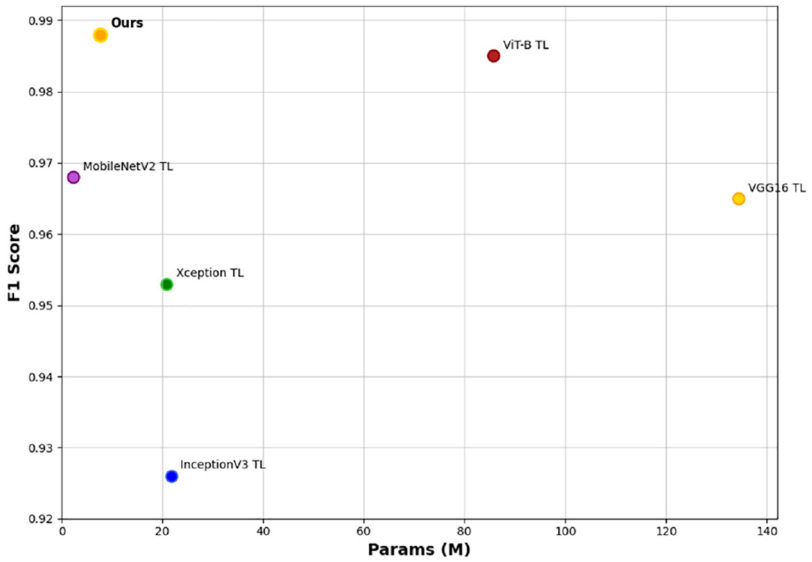


Figure 11: Parameter count comparison results.

In the scatter plot, the proposed method is located in the upper-left region, indicating that it achieves a superior balance between performance and parameter count compared with other models. Specifically, the proposed method has an F1-score of 0.988, outperforming InceptionV3, Xception, and VGG16, while its parameter count is only 7.66M, far lower than the 134.3M of VGG16 and the 85.6M of ViT-B. This result indicates that the proposed method effectively reduces model complexity while maintaining high detection performance. Although the proposed method has more parameters than MobileNetV2 (F1 = 0.968, parameter count = 2.2M), it has a clear advantage in F1-score. With only 5.4M more parameters, the proposed method improves the F1-score by 0.02, demonstrating reasonable control over the parameter count and stronger performance advantages. In addition, the current efficiency evaluation mainly considers parameter count. Floating-point operations (FLOPs), memory consumption, and inference latency were not systematically evaluated and should be investigated in future work to provide a more comprehensive assessment of practical deployment efficiency.

5.4 Ablation Experiment Results

To demonstrate the rationality and effectiveness of the design, this paper sets up two ablation experiments: a layer-design experiment and a component experiment.

(1) Layer-design ablation experiment

The channel ablation experiment evaluates the rationality of the cascaded network architecture design by comparing the F1-scores and parameter counts of SCNet50, SCNet101, and the SCNet designed in this paper with a 1:1:2:1 ratio. The results are analyzed accordingly. The results are shown in Table 6, and a scatter plot is drawn with parameter count as the horizontal axis and F1-score as the vertical axis, as shown in Fig. 12.

The layer-design experiment shows that, for the forged satellite image detection task, redesigning SCNet according to the 1:1:2:1 ratio can reduce the parameter count while ensuring model performance, and it obtains the highest F1-score. In the scatter plot, the newly designed SCNet is located in the upper-left region. Compared with SCNet50 and SCNet101, it achieves a more effective balance between parameter count and performance.

Table 6: Ablation study results.

Layers	F1	Para (M)
SCNET50	0.975	25.62M
SCNET101	0.976	44.62M
SCNet (1:1:2:1)	0.984	11.26M

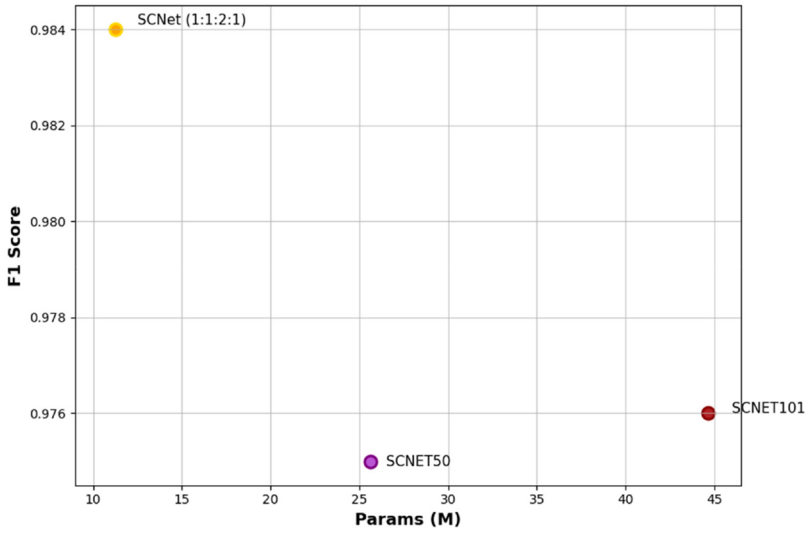


Figure 12: Parameter count and F1-score comparison for the layer-design ablation study.

(2) Module ablation experiment

The component ablation experiment evaluates the influence of different components on detection performance. Controlled-variable ablation experiments are conducted to compare the SCNet designed with a 1:1:2:1 layer ratio, the depthwise separable self-calibrating convolutional network (DSSCNet) designed with depthwise separable self-calibrating convolution according to the 1:1:2:1 layer ratio, and the proposed comprehensive-feature-based detection model. The results are shown in Table 7, and a scatter plot is drawn with parameter count as the horizontal axis and F1-score as the vertical axis, as shown in Fig. 13.

Table 7: Ablation study results.

Model	F1	Para (M)
SCNet	0.984	11.26M
DSSCNet	0.986	7.97M
Ours	0.988	7.66M

The ablation experiments show that the complete architecture cascading the DSSC module and IFEM achieves the best performance and the lowest parameter count. This indicates that the proposed design achieves an effective balance between parameter count and performance and can fully utilize comprehensive features while sufficiently reducing model computational complexity. It is worth noting that, through the improvement of SCModule, the DSSCNet constructed by the designed DSSC reduces the parameter count by 3.29M while increasing the F1-score by 0.002. This demonstrates that the redesign of SCModule can

maintain the model's sensitivity to local features while effectively reducing the number of parameters. The current ablation study primarily evaluates the contribution of the DSSC redesign and the complete IFEM-based feature extraction stage. It does not independently remove MHSA and adaptive attention from IFEM. Therefore, the individual contributions of these two attention components cannot be quantitatively separated based on the current experiments. Fine-grained ablation experiments will be conducted in future work to further investigate the independent contribution of each component.

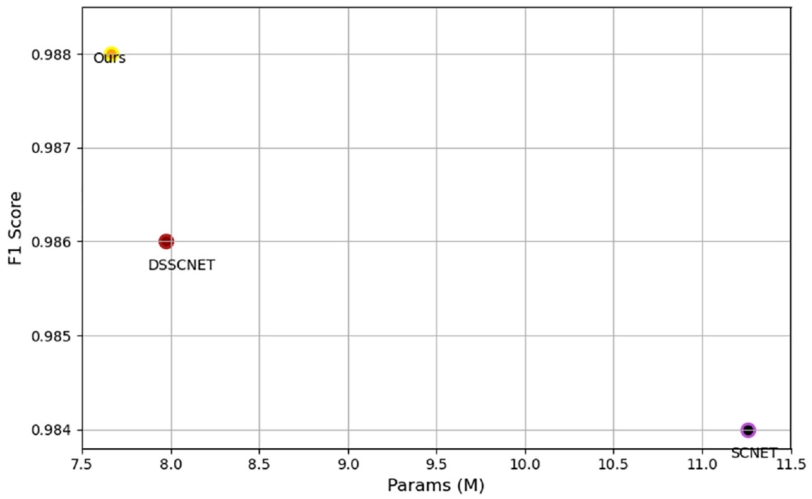


Figure 13: Parameter count and F1-score comparison for the module ablation study.

Precision, Recall, and F1-score are adopted as the primary evaluation metrics in the present experiments. Nevertheless, the area under the receiver operating characteristic curve (ROC-AUC), the area under the precision-recall curve (PR-AUC), and the Matthews correlation coefficient (MCC) can provide additional information regarding ranking performance, precision-recall behavior, and classification correlation, respectively. These metrics will be incorporated into future evaluations to provide a more comprehensive assessment.

6 Conclusions

This paper addresses the limitations of conventional convolutional neural networks (CNNs) in modeling global feature dependencies, the insufficient local-detail inductive capability of Transformer-based models, and the relatively high computational cost associated with combining these architectures. To this end, an efficient cascaded deepfake detection model is proposed for distinguishing real and forged satellite images. The proposed framework combines a depthwise separable self-calibrating convolution (DSSC) module with an integrated feature extraction module (IFEM) in a sequential architecture. DSSC is employed to extract subtle local forgery traces while reducing parameter redundancy through depthwise and pointwise convolutions. Subsequently, IFEM employs multi-head self-attention (MHSA) to model long-range dependencies among spatial features and uses an adaptive attention mechanism to further refine task-relevant feature responses. This local-to-global feature extraction strategy enables the proposed model to jointly exploit fine-grained local information and global contextual relationships.

Experimental results on the UW dataset demonstrate that the proposed method achieves an F1-score of 0.988 on the original test set, outperforming the comparison methods considered in this study. The proposed model contains 7.66M parameters, indicating that it maintains a relatively compact model size while achieving competitive detection performance. The ablation experiments further demonstrate the

effectiveness of replacing conventional self-calibrating convolution with DSSC and subsequently introducing the integrated feature extraction stage. In addition, the degradation experiments show that the proposed model maintains competitive detection performance under the representative JPEG compression, Gaussian noise, and Gaussian blur conditions evaluated in this study. However, the results also reveal that the robustness of the model is not uniform across all degradation levels. In particular, under severe 9×9 Gaussian blur, the model achieves a Precision of 0.996 but its Recall decreases to 0.732, resulting in an F1-score of 0.844. This indicates that severe degradation of local texture and structural information can cause the model to miss a proportion of forged samples.

Despite these promising results, several limitations remain. First, the current experiments are conducted on the UW dataset, which covers a limited number of geographical regions. Although this dataset provides a consistent benchmark for the experiments in this study, it may not fully represent the geographical and visual diversity of satellite imagery encountered in real-world applications. Therefore, the cross-dataset and cross-geographical generalization capability of the proposed method has not yet been systematically validated. Second, the current robustness evaluation focuses on three representative degradation types, namely JPEG compression, Gaussian noise, and Gaussian blur. More challenging conditions, including mixed degradations and unseen manipulation operations, have not been systematically investigated and may introduce stronger distribution shifts. Third, the current computational-efficiency analysis is mainly based on parameter count. Although the proposed model contains 7.66M parameters, FLOPs, memory consumption, and inference latency have not been systematically measured across different hardware platforms. Therefore, the practical real-time deployment efficiency of the proposed method remains to be further validated. Finally, the current ablation study evaluates the contribution of DSSC and the complete IFEM stage, but does not independently quantify the contributions of MHSA and adaptive attention.

Future research will focus on improving the generalization, robustness, interpretability, and deployment efficiency of the proposed method. First, additional satellite-image datasets covering more geographical regions and imaging conditions will be incorporated to evaluate cross-dataset and cross-domain generalization. More realistic forged satellite images generated using advanced generative models, including generative adversarial networks and diffusion models, will also be considered to improve the diversity of manipulation patterns. Second, mixed degradation conditions and unseen manipulation operations will be investigated to evaluate the robustness of the detector under more challenging distribution shifts. Third, fine-grained ablation experiments will be conducted to separately investigate the contributions of DSSC, MHSA, and adaptive attention. Fourth, explainable artificial intelligence techniques, such as gradient-weighted class activation mapping (Grad-CAM), will be explored to visualize the image regions that contribute to model predictions and to examine whether the detector relies on meaningful local forgery artifacts and global semantic inconsistencies. Finally, FLOPs, memory consumption, and inference latency will be systematically evaluated on heterogeneous hardware platforms to provide a more comprehensive assessment of computational efficiency and deployment potential. These efforts will further clarify the applicability and limitations of the proposed approach and support the development of more generalizable, interpretable, robust, and practically deployable satellite-image deepfake detection systems.

Acknowledgement: The authors gratefully acknowledge Zhao et al. for releasing the UW dataset to the public, which provides the fundamental data support for the experimental verification of this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Xiaochen Hao: conceptualization, methodology, Core algorithm design, original draft writing. Wei Liu (Corresponding author): supervision, research direction guidance, manuscript review and editing, experimental guidance. Yuhuan Liu: investigation, experimental procedure optimization, model tuning, result comparison

and analysis. Zheng Kong: experimental scheme design, dataset preprocessing, test scenario construction. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The UW deepfake satellite image dataset used in this study is a publicly available dataset proposed by Zhao et al. [6]. Dataset access link: https://figshare.com/s/eeedcd150e759ef4353c?file=2889103. Original publication of the dataset: https://doi.org/10.1080/15230406.2021.1910075.

Ethics Approval: This study only conducts algorithm research and performance verification based on publicly released satellite image datasets. It does not involve human participants, animal experiments, or any content requiring ethical review and approval.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. doi:10.1038/nature14539.
2. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139–44. doi:10.1145/3422622.
3. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst. 2020;33:6840–51. doi:10.48550/arxiv.2006.11239.
4. Chadha A, Kumar V, Kashyap S, Gupta M. Deepfake: an overview. In: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020. Singapore: Springer; 2021. p. 557–66. doi:10.1007/978-981-16-0733-2_39.
5. Wang R, Chu B, Yang Z, Zhou L. An overview of visual DeepFake detection techniques. J Image Graph. 2022;27(1):43–62. doi:10.11834/jig.210410.
6. Zhao B, Zhang S, Xu C, Sun Y, Deng C. Deep fake geography? When geospatial data encounter artificial intelligence. Cartogr Geogr Inf Sci. 2021;48(4):338–52. doi:10.1080/15230406.2021.1910075.
7. Du S-H, Li W, Xing J-H, Zhang C-Y, She C-C, Wang S-Y, et al. Change detection of open-pit mines based on FM-UNet++ and GF-2 satellite images. Coal Geol Explor. 2023;51(7):130–9.
8. Fu Q, Luo W, Lü J. Land utilization change detection of satellite remote sensing image based on AlexNet and support vector machine. Laser Optoelectron Prog. 2020;57(17):172802. doi:10.3788/lop57.172802.
9. Ma X, Xu J, Pan J, Yang J, Wu P, Meng X. Detection of marine oil spills from radar satellite images for the coastal ecological risk assessment. J Environ Manag. 2023;325(Pt B):116637. doi:10.1016/j.jenvman.2022.116637.
10. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 2223–32. doi:10.1109/iccv.2017.244.
11. Ding X, Nie Y, Yao J, Tang J, Lang Y. Forensic research of satellite images forgery: a comprehensive survey. Artif Intell Rev. 2024;57(9):253. doi:10.1007/s10462-024-10909-w.
12. Aslan E, Özüpak Y. Detection of road extraction from satellite images with deep learning method. Cluster Comput. 2024;28(1):72. doi:10.1007/s10586-024-04880-y.
13. Li X-R, Ji S-L, Wu C-M, Liu Z-G, Deng S-G, Cheng P, et al. Survey on deepfakes and detection techniques. J Softw. 2021;32(02):496–518. (In Chinese). doi:10.4236/jcc.2021.95003.
14. Fezza SA, Ouis MY, Kaddar B, Hamidouche W, Hadid A. Evaluation of pre-trained CNN models for geographic fake image detection. In: Proceedings of the 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP); 2022 Sep 26–28; Shanghai, China. p. 1–6. doi:10.1109/MMSP55362.2022.9949282.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
16. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014. doi:10.48550/arxiv.1409.1556.

17. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 2818–26. doi:10.1109/CVPR.2016.308.
18. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1251–8. doi:10.1109/cvpr.2017.195.
19. Zhao J, Shi Z, Yu C, Liu Y. Multi-scenario remote sensing image forgery detection based on transformer and model fusion. Remote Sens. 2024;16(22):4311. doi:10.3390/rs16224311.
20. Liu JJ, Hou Q, Cheng MM, Wang C, Feng J. Improving convolutional networks with self-calibrated convolutions. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 10096–105. doi:10.1109/cvpr42600.2020.01011.
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008. doi:10.65215/r5bs2d54.
22. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 7132–41. doi:10.1109/CVPR.2018.00745.
23. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
24. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4510–20. doi:10.1109/CVPR.2018.00474.