Automatic Detection Models for Building Exterior Wall Cracks in Drone Imagery Based on CNN And Transformer

Yaoling Shang[1], Ying Ge[2], Yuqing Ma[3], Yingying Zhang[4], Shilin Lv[5]

[1] National Quality Inspection and Testing Center for Surveying and Mapping Products, People's Republic of China - shangyl@vip.sina.com

[2] Hohai University, People's Republic of China - geying@hhu.edu.cn

[3] State Grid Zhejiang Electric Power Co.,Ltd. Logistics Service Company, People's Republic of China – 89410100@qq.com

[4] State Grid Zhejiang Electric Power Co.,Ltd. Logistics Service Company, People's Republic of China – 28488836@qq.com

[5] State Grid Zhejiang Electric Power Co.,Ltd. Logistics Service Company, People's Republic of China – 392022437@qq.com

Keywords: Building Facades, Drone Imagery, Crack Detection, U-Net Model, Transformer Architecture, Image Segmentation.

Abstract

This study constructs a comprehensive evaluation framework comprising six representative models: standard U-Net, Resnet34-U-Net, UNet-Attention, UNet-Residual, HybridUNet, and TransUNet. We performed systematic ablation experiments to analyse the contributions of different architectural components, including residual connections, attention mechanisms, and Transformer modules. The models were trained and validated on a dedicated dataset of building exterior crack images captured by drones, with careful consideration of the challenges posed by complex backgrounds, varying lighting conditions, and fine crack features. Multiple loss functions - F1 Loss, Focal-Dice-Loss, and BCE-Dice-Loss - were evaluated to determine their impact on model performance. The evaluation employed comprehensive metrics including Accuracy, F1 Score, IoU, Precision, Recall, and Loss values to ensure thorough performance assessment.

Experimental results demonstrate that TransUNet achieved the best overall performance with F1 Score of 87.66%, Precision of 90.43%, and Recall of 89.99%, leveraging its Transformer module's global context modelling capability. In loss function comparisons, F1 Loss yielded the most balanced performance on TransUNet with F1 Score of 87.50%, while Focal-Dice-Loss showed exceptional optimization stability with the lowest loss value (0.1008) and high recall (96.05%). Interestingly, the performance gap among the six models was relatively small, with the difference in F1 Score between the optimal TransUNet and baseline standard U-Net being less than 0.5%. Qualitative analysis revealed that while complex models like TransUNet excel in overall metrics, simpler architectures like UNet-Attention and UNet-Residual demonstrate better robustness in challenging scenarios with complex textures, highlighting the importance of context-specific model selection.

This research provides comprehensive insights into deep learning approaches for building exterior crack detection. TransUNet with F1 Loss emerges as the optimal solution for high-accuracy requirements, while standard U-Net and its attention-enhanced variants offer cost-effective alternatives for large-scale applications. The minimal performance gap among different architectures suggests that model complexity alone doesn't guarantee superior performance for this specific task. The study emphasizes the importance of balancing accuracy needs with computational efficiency in practical engineering applications. These findings offer valuable guidance for model selection in real-world building maintenance scenarios and contribute to the advancement of intelligent detection technologies in structural health monitoring. Future work should focus on enhancing model robustness across diverse environmental conditions and optimizing computational efficiency for broader implementation.

1. Introduction

Traditional crack detection relies on manual inspection and periodic maintenance, which is not only time-consuming and labour-intensive but also highly subjective. Furthermore, it is often difficult to implement for high-rise buildings, leading to potential hazards that cannot be identified and resolved promptly. Therefore, there is an urgent need to develop a rapid method for detecting cracks on external walls, enabling the daily inspection and maintenance of buildings, especially high-rise structures, thereby facilitating early risk identification and preventing greater losses. In recent years, the widespread application of technologies such as unmanned aerial vehicles (UAVs) and artificial intelligence (AI) has made the automatic detection of cracks in high-rise buildings possible (Chen et al., 2025). By employing UAVs for autonomous inspection around high-rise buildings, combined with image recognition, big data analysis, and deep learning algorithms, the health condition of buildings can be assessed quickly and accurately. This significantly reduces the risks for workers performing high-altitude tasks, enhances the reliability of health monitoring for high-rise buildings, and provides advanced technical support for

the intelligent management and maintenance of building structures.

Among CNN architectures, the U-Net model has become a research focus in many fields due to its powerful capabilities in image feature learning and representation (Lian et al., 2023; Zhang et al., 2004; Liu et al., 2025; Xu et al., 2025). Since Zhang et al. (2016) first applied convolutional neural networks to road crack detection, CNNs have been widely used for detecting surface cracks on various concrete structures, such as tunnels (Liu et al., 2018; Chang et al., 2020), roads (Weng et al., 2019; Li et al., 2023), bridges (Zhu et al., 2019; Qiao et al., 2024), hydraulic structures (Hu et al., 2023; Zhu et al., 2024), and building exteriors (Chaiyasarn et al., 2018; Cai et al., 2022; Liu et al., 2022; Loverdos and Sarhosis, 2022).

The U-Net model employs a symmetric encoder-decoder structure and utilizes skip connections to achieve multi-scale feature fusion. This enables effective capture of local contextual information of cracks, achieves precise pixel-level localization, and thereby enhances the ability to extract crack details. However, Ji et al. (2020) suggested that the Deeplabv3+ model performs better than the U-Net model in crack detection applications. Nevertheless, the core architecture of U-Net is based on CNNs, and the local receptive field characteristic of convolutional operations inherently limits the model's ability to establish long-range dependencies between image patches, leading to shortcomings in integrating global contextual information from the entire image (Strudel et al., 2021). Conversely, the Deeplabv3+ model struggles to effectively reduce the false positive rate, which is a challenge currently faced in crack semantic segmentation (Hsieh et al., 2020).

In light of these limitations, researchers have begun to explore architectures with global modelling capabilities. Dosovitskiy et al. (2020) first proposed the Vision Transformer (ViT) architecture and applied it to image classification. Its core self-attention mechanism can directly establish global dependencies on sequences of image patches, offering a new approach to overcome U-Net's limitations in long-range modelling. Introducing the self-attention mechanism into U-Net-like models can not only enhance the model's grasp of hierarchical semantic information in images, but its built-in positional encoding also helps to more accurately retain spatial positional information of features, thereby effectively mitigating the attenuation of spatial structural information that occurs as depth increases in traditional deep convolutional networks.

In the field of building exterior wall crack detection, deep learning-based methods continue to evolve. Yang et al. (2018) pioneered the use of Fully Convolutional Networks (FCN) to achieve end-to-end pixel-level crack segmentation, significantly improving detection efficiency. Subsequently, Liu et al. (2019) proposed the DeepCrack model specifically designed for crack segmentation, which maintains superior performance even in complex backgrounds by fusing multi-scale features. To systematically evaluate the effectiveness of different methods, Loverdos and Sarhosis (2022) conducted a systematic review of various image segmentation models, including FCN, U-Net, DeepLab, and FPN, for exterior wall inspection tasks. They summarized common evaluation metrics and proposed an improved DeepCrack model suitable for traditional English bond masonry walls. With the rise of Vision Transformers, Shamsabadi et al. (2022) combined Transformer with U-Net to construct the TransUNet model. Experiments showed that in asphalt and concrete crack detection, its IoU metric significantly increased by 61% and 3.8% compared to DeepLabv3+ and U-Net, respectively, demonstrating the clear potential of models combining global perception and local details for crack detection.

In related research, Liu et al. (2022) and Cai et al. (2022) divided building facades into 3720 sub-regions and used the deep residual network ResNet-101 for crack identification. Their model achieved training phase accuracy and recall rates of 99.15% and 97.38%, respectively. In overall facade testing, the model correctly identified 226 cracked areas, with 13 misclassifications and 14 missed detections, achieving crack identification for the entire building facade. The research indicates that this method can effectively achieve crack identification, validating the feasibility of this technical approach.

In recent years, image segmentation techniques based on improvements and extensions of the U-Net model have continued to develop. Based on a systematic review of relevant methods in this field, this paper first categorically elaborates on the core modules, loss functions, and evaluation metrics commonly used in image segmentation. Subsequently, it specifically discusses the U-Net model and its improved variants, as well as strategies for integrating them with the Transformer architecture. Finally, six summarized models are applied to building exterior wall crack detection and evaluated in detail, aiming to provide a reference for subsequent research.

2. Methodology

By reviewing the technologies of the U-Net model and its Transformer architecture, a comparative framework comprising six models—Standard U-Net, ResNet34-UNet, UNet-Attention, UNet-Residual, HybridUNet, and TransUNet—was constructed. Different loss functions and evaluation parameters were compared to detect and analyse cracks on building exteriors using UAV aerial imagery.

2.1 Standard U-Net Model and Its Improvements

The standard U-Net model employs a symmetric encoder-decoder structure (Ronneberger et al., 2015). The encoder progressively extracts contextual features of the image through four down-sampling operations and two convolutional layers, compressing the spatial dimensions. The decoder gradually restores detail and resolution through up-sampling and convolutional operations, ultimately outputting a pixel-level segmentation result. The core innovation of this model lies in its introduction of skip connections, which fuse the high-resolution shallow features from the encoder stage with the deep semantic features from the decoder stage. This mechanism helps mitigate gradient vanishing, reduces information loss, and significantly enhances model convergence speed and segmentation accuracy. Due to its well-designed architecture and excellent performance, the standard U-Net model has become a widely adopted foundational framework for small-sample image segmentation tasks.

However, the standard U-Net model faces several challenges when directly applied to building exterior crack detection. On one hand, cracks typically occupy a very low pixel proportion in the overall image, leading to a severe class imbalance problem. On the other hand, building exterior backgrounds are complex, often containing interference such as brick patterns, stains, and shadows, while the cracks themselves exhibit

diverse morphologies and large scale variations, further complicating accurate segmentation. Furthermore, as the network depth increases, the model encounters the problem of network degradation, where the training error increases with depth. This limits the standard U-Net model's ability to extract image features.

To address these issues, this paper improves the standard U-Net model from both the data and model perspectives. At the data level, augmentation strategies are employed to enhance sample diversity, model generalization ability, and robustness. At the model level, optimizations focus on the U-Net architecture itself. Firstly, an attention mechanism is introduced into the skip connections to enhance the screening and fusion of crack-related features while suppressing background interference. Secondly, stacked residual modules are added to the bottleneck layer to strengthen gradient propagation and deep feature extraction capability, thereby alleviating network degradation. These improvements aim to further enhance the model's perception and segmentation performance for faint crack features while preserving the inherent advantages of the standard U-Net model.

Additionally, using ResNet as the U-Net's backbone network is another improvement strategy. By utilizing residual connection networks (e.g., ResNet-34, ResNet-50, or even ResNet-101), the network can learn residual mappings. The primary reasons are: (1) To solve the deep network degradation problem, enabling the construction of deeper U-Net models, thereby obtaining larger receptive fields and stronger feature extraction capabilities; (2) To alleviate the vanishing gradient problem, meaning parameters in the lower layers of the network (responsible for extracting basic features like edges and textures) can also be fully trained, ensuring the model captures subtle crack features; (3) To leverage powerful pre-trained weights for efficient transfer learning.

Specific improvement measures are as follows:

(1) Data Augmentation

Data augmentation involves applying geometric transformations, adding noise, interpolation, etc., to the original sample images to increase the sample size, thereby improving the model's generalization ability and robustness. Given that cracks are thin and sparse, data augmentation in this study is limited to geometric and illumination variations to increase data diversity. The data augmentation methods used in this paper are listed in Table 1.

Category	Specific Operation	Probability
Flipping	Horizontal flip / vertical flip	0.5 / 0.1
Affine transformation	Translation, scaling, rotation	0.5
Brightness/contrast adjustment	Random brightness contrast	0.3
Noise	Gauss noise	0.1
Size unification	Resize	1.0
Normalization	Normalize	1.0

Table 1. Image data augmentation methods.

(2) Incorporation of Residual Modules

The core idea behind the residual module is to address the issues of vanishing gradients and network degradation when training deeper U-Net models by introducing a residual mechanism, while simultaneously preserving more image feature information during feature propagation (He et al., 2016). Consequently, in the field of image segmentation, residual modules are widely used in improving the U-Net structure, primarily through the following three methods:

a) Replacing Standard Convolutional Blocks. This method was first used in the field of image segmentation. It retains the overall encoder-decoder architecture and skip connections of U-Net but replaces each "double convolutional layer" unit in the encoder and decoder with a "residual block," thereby constructing Res-UNet (Xiao et al., 2019). For targets like cracks, which have an extremely small pixel proportion and subtle features, the residual structure alleviates the vanishing gradient problem, enabling deep networks to more effectively learn the feature differences between cracks and the background. This mitigates the negative impact of class imbalance and significantly improves crack segmentation accuracy.

b) Stacking Residual Modules in the Bottleneck Layer (Ibtehaz and Rahman, 2019). Stacking two or more residual modules at the deepest part of the network enhances its representational capacity and nonlinear fitting ability. This design allows the model to better understand crack structures with global semantic features (e.g., cracks penetrating the wall surface), thereby reducing the misclassification of isolated short cracks or noise as cracks and lowering the false positive rate.

c) Introducing a Residual Connection Mechanism between the Encoder and Decoder. By introducing residual modules into the skip connections, the global semantic information of deep feature maps can be preserved, making feature fusion more stable and further enhancing the understanding of the crack context.

For building exterior crack detection, an effective and cost-efficient strategy is method 2, stacking residual modules in the bottleneck layer. Its unique identity mapping mechanism effectively alleviates the gradient attenuation problem in deep networks, ensuring training stability. Simultaneously, constructing a deeper bottleneck structure significantly enhances the model's ability to understand and integrate global semantic information, enabling the network to better capture long-range crack features and complex contextual relationships across image regions. This notably improves the network's robustness and generalization ability, effectively addressing the subtlety and complexity of crack features.

(3) Incorporation of Attention Modules

The primary function of attention modules is to enhance target features and suppress background interference. In the field of image segmentation, common forms mainly include the following four categories:

a) Gated Attention Module at Skip Connections (Oktay et al., 2018). This method was among the first used in image segmentation. By introducing a gating mechanism into the skip connections, it effectively learns and suppresses noise features unrelated to cracks, such as mortar joints, stains, shadows, and insulation layer textures. It allows only crack-related edge and linear structural features to pass through, significantly improving the effectiveness of feature transfer.

b) Dual Attention (CBAM) Module for Spatial and Channel Dimensions in the Bottleneck Layer (Park et al., 2018). Cracks often exhibit long extensions, requiring reliance on contextual information from a large receptive field for identification. Features in the bottleneck layer have the broadest receptive field. Using a parallelly added CBAM module integrates spatial and channel attention information, reducing the misclassification of brick joints or scratches as cracks.

c) Channel Attention Module (Hu et al., 2020). The Channel Attention Module (i.e., Squeeze-and-Excitation module) can adaptively evaluate the importance of each feature channel, enhancing the responses of channels related to cracks and suppressing features that contribute less to the segmentation task. Typically, adding channel attention modules to the encoder and bottleneck layer can significantly enhance the extraction of crack features.

d) Dual Attention Module in the Decoder (Tian et al., 2019). Incorporating an attention mechanism during the decoding stage allows the model to focus on the spatial regions and semantic information most relevant to cracks in each up-sampling step, thereby progressively refining the segmentation results and improving boundary recognition accuracy.

For the UAV crack detection application scenario in this paper, the core attention modules are: (1) Gated Attention at skip connections (local purification); and (2) SE modules embedded in the encoder (source enhancement). These two are the priority "essential" modules. When computational resources allow, the CBAM module in the bottleneck layer (global analysis) can be further introduced. The combination of these three forms a complete optimization path: "Local Purification (Gated) → Source Enhancement (SE) → Global Analysis (CBAM)", capable of comprehensively improving the accuracy and robustness of crack detection from local to global levels, achieving an optimal balance between feature representation and segmentation performance.

(4) Combined Model Integrating Residual and Attention Modules

Based on the standard U-Net model and targeting the morphological characteristics and detection challenges of building exterior cracks, this paper adopts a minimized structural improvement strategy. Namely, a gated attention mechanism is embedded at the skip connections, while stacked residual modules are introduced in the bottleneck layer. This design aims to strengthen the model's capability for selective fusion of crack features and deep semantic extraction, thereby improving segmentation accuracy and robustness. Specifically:

a) Gated Attention Skip Connections: The gating mechanism adaptively calibrates the multi-scale features transmitted from the encoder to the decoder, suppressing interference from irrelevant information like wall background textures, and highlighting structure and detail features related to cracks, thereby improving feature fusion quality.

b) Residual Stacking in the Bottleneck Layer: Stacking residual modules in the deepest part of the network (the bottleneck) enhances gradient flow and model capacity, improves the understanding of global context and complex patterns of cracks, and maintains training stability.

This combined structure achieves effective feature selection between the encoder and decoder and enhanced learning of deep representations at a relatively low computational cost, providing an efficient and reliable solution for the accurate segmentation of exterior wall cracks.

To compare the performance of the U-Net models, the following sets of ablative experiments were established: (a) Baseline: Standard U-Net; (b) Attention Only: Standard U-Net + Gated Attention Skip Connections; (c) Residual Only: U-Net + Stacked Residual Modules in the Bottleneck Layer; (d) Combined Model; (e) Adding a ResNet Backbone. By comparing the metrics (F1-score, Precision, Recall, etc.) of the above groups on the validation/test sets, the performance of each module and their combinations was observed. For all experiments, the network depth was set to 4 and the number of convolutional layers to 2.

2.2 TransUNet Model: U-Net Embedded with Transformer

The Transformer architecture was first applied to image semantic segmentation by Zheng et al. (2021) and has since attracted extensive attention due to its powerful feature extraction capabilities and ability to model long-range dependencies. Traditional CNN models, represented by U-Net, rely on convolutional kernels to capture features. The convolutional operation is inherently local, and its receptive field is limited. To acquire broader contextual information, U-Net must progressively expand the receptive field by stacking multiple convolutional layers, introducing pooling layers, or using dilated convolutions, among other operations.

In contrast, the self-attention mechanism of the Vision Transformer (ViT) framework fundamentally addresses this limitation (Carion et al., 2020). This mechanism can compute the relationships between all image patches in the sequence from the very first layer. Regardless of the actual distance between two patches in the image, their relationship can be directly modelled. This provides immediate and pure global contextual information for image segmentation tasks, granting it an inherent advantage in understanding complex scene structures.

The TransUNet model embeds the Vision Transformer (ViT) framework as an encoder into the classic U-Net architecture, forming an efficient hybrid model (Ma et al., 2024). This hybrid model overcomes the limitations of the traditional U-Net, such as its restricted receptive field and insufficient capability in modelling long-range dependencies caused by its inherent locality. Simultaneously, it mitigates the shortcomings of pure Vision Transformer (ViT), such as the loss of spatial structural details and weakened positional information resulting from the sequential processing of images. By integrating the local feature extraction strengths of CNNs with the global contextual modelling capabilities of the Transformer, TransUNet achieves a balance between global semantic understanding and local detail recovery, making it widely applicable in the field of image segmentation.

As shown in Figure 1, the core of the TransUNet model lies in image serialization (Li et al., 2024). This process addresses the primary challenge of adapting the successful Transformer architecture from Natural Language Processing to the computer vision domain, particularly for tasks involving 2D image data processing. Specifically, it involves reshaping the image into a

sequence of patches and transforming them into sequential data, subsequently leveraging a pre-trained Vision Transformer (ViT) as the encoder.

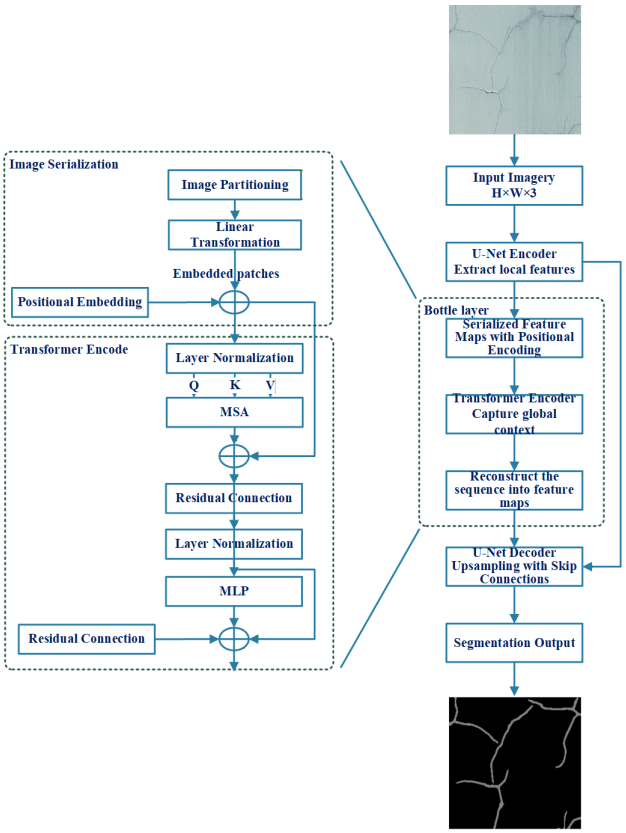


Figure 1. TransUNet model architecture.

The Image Serialization Process: (1) Patching. The input image with dimensions H × W × C is divided into N = (H × W) / P^2 non-overlapping regular patches, each of size P × P × C. (2) Embedding. Each image patch is flattened into a 1D vector and projected into a latent embedding space via a linear projection layer, forming a sequence of patch embeddings. (3) Positioning. Learnable position encoding vectors are introduced to compensate for the inherent lack of positional information in the self-attention mechanism. (4) Enhancement. The position encodings and patch embeddings are combined using element-wise addition, generating a position-augmented sequence representation. (5) Transformation. This position-augmented sequence serves as the input to the encoder, completing the mathematical conversion from a 2D image to a 1D sequence, which is suitable for input into the Transformer encoder.

The Transformer Encoder is composed of l (where l =1,…, L) identical layers stacked upon each other. Each layer consists of two core modules, that is, Multi-headed Self-Attention (MSA) and Multilayer Perceptron (MLP). The former is to establish long-range dependencies, capturing the complete context of cracks across the entire image. The latter is to perform non-linear transformation and feature refinement on the output of the MSA, enhancing the model's ability to distinguish between "crack" and "non-crack" features. It is important to note that Layer Normalization (LayerNorm) is applied before the MSA module. The output of the MSA module is then combined with its initial input via a residual connection, followed by another LayerNorm. This normalized output is subsequently fed into the MLP layer. Finally, the output of the MLP module is combined with the MLP module's input through another residual connection.

In TransUNet, the U-Net depth, feature map size, and number of Transformer layers are three hyperparameters that critically determine the model's training effectiveness. For instance, given an input image size of 256×256, a typical configuration uses a U-Net depth of 4 layers, a feature map size of 16×16, and 6 Transformer layers.

2.3 Model Evaluation

Leave two blank lines under the keywords. Type "**Abstract**" flush left in bold, followed by one blank line. Note that the abstract should be concise (100 - 250 words), present briefly the content and very importantly, the scientific contribution and results of the paper in words understandable also to non-specialists.

2.3.1 Evaluation Metrics: The model evaluation metrics selected in this paper primarily address the binary classification problem of "crack" versus "non-crack". These include Precision, Recall, Accuracy, and the F1-score. Here, TP = True Positive, FP = False Positive, FN = False Negative, and TN = True Negative.

(1) Precision indicates the proportion of samples predicted as positive that are correctly predicted. It is calculated as the ratio of true positive predictions to the total number of samples predicted as positive (the sum of true positives and false positives).

$$P = \frac{TP}{TP + FP}, \tag{1}$$

(2) Recall indicates the proportion of actual positive samples that are correctly predicted.

$$R = \frac{TP}{TP + FN}, \tag{2}$$

(3) Accuracy refers to the proportion of all predictions that are correctly identified, encompassing both positive and negative classes. It measures the overall correctness of the model. The formula for Accuracy is:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

(4) F1-Score is a crucial metric for comprehensively evaluating the performance of a classification model, particularly suitable for datasets with class imbalance (Lin et al., 2017). It represents the harmonic mean of Precision and Recall, providing a balanced measure between the two. The formula for the F1-Score is:

$$F1\ Score = \frac{2 \times \left(\mathrm{Pr}ecision \times \mathrm{Re}call\right)}{\left(\mathrm{Pr}ecision + \mathrm{Re}call\right)}, \tag{4}$$

(5) IoU (Intersection over Union), also known as the Jaccard Index, is a standard metric for measuring the similarity between two sets (Rezatofighi et al., 2019). In the context of image segmentation, it measures the overlap between the predicted segmentation area and the ground truth area. It is calculated as the size of the intersection of the predicted and true regions divided by the size of their union. The formula for IoU is:

$$IoU = \frac{TP}{TP + FP + FN}, \quad (5)$$

IoU and the F1-score are two related yet complementary metrics. IoU measures spatial geometric accuracy, while the F1-score provides a comprehensive performance evaluation from the perspective of pixel classification. If the focus is on the fineness of segmentation boundaries, IoU is a crucial metric as it directly reflects the degree of alignment between the predicted contour and the ground truth contour. If the priority is the accurate prediction of cracks with minimal false alarms, then the F1-score is a more suitable metric. For building facade crack detection, it is advisable to report both metrics simultaneously.

2.3.2 Loss Functions: The loss function quantifies the discrepancy between model predictions and ground truth labels. Optimization algorithms like gradient descent are used to minimize this loss, thereby automatically updating the network weights. Loverdos and Sarhosis (2022) employed multiple loss functions—including Binary Cross-Entropy (BCE), Weighted Cross-Entropy (WCE), Focal Loss (FL), and F1 Loss (F1L)—to identify the optimal model for evaluating the difference between predictions and targets for cracks in traditional British masonry walls.

(1) Binary Cross-Entropy (BCE) is specifically designed for binary classification tasks. It measures the dissimilarity between the probability distribution of the model's predictions and the distribution of the true labels. For a binary classification where the true label y is either 0 or 1, and the predicted probability is p, the BCE loss for a single sample is defined as:

$$BCE = -\left[y \times \ln\left(p\right) + \left(1-y\right) \times \ln\left(1-p\right)\right], \quad (6)$$

(2) Weighted Cross-Entropy (WCE), or Weighted Binary Cross-Entropy, builds upon the standard Binary Cross-Entropy by assigning different weights to different classes to address the issue of class imbalance. In tasks like crack detection where the positive class (cracks) is significantly underrepresented, WCE increases the cost of misclassifying the minority class. This encourages the model to pay more attention to learning the features of the underrepresented class. The WCE loss for a single sample is defined as:

$$WCE = -\left[w\_pos \times y \times \ln\left(p\right) + w\_neg \times \left(1-y\right) \times \ln\left(1-p\right)\right], \quad (7)$$

where y is the true label (0 or 1); p is the model's predicted probability for the positive class; w_pos is the weight assigned to the positive class; w_neg is the weight assigned to the negative class. A common practice is to set the weights inversely proportional to the class frequencies in the training data.

(3) Focal Loss (FL) is a modified loss function based on the standard cross-entropy loss (Lin et al., 2017). By introducing a dynamically scaling factor, it effectively addresses the issue of severe foreground-background class imbalance common in image segmentation tasks. Its core idea is to reduce the relative loss contribution from well-classified, easy examples (which are typically the abundant negative/background samples), forcing the model to focus its learning on hard-to-classify examples. This implements a form of automatic hard example mining. The Focal Loss for a binary classification task is defined as:

$$FL\left(p_t\right) = -\alpha_t\left(1-p_t\right)^{\gamma} \ln\left(p_t\right), \quad (8)$$

where p_t is the model's estimated probability for the true class. When the true label is the positive class (e.g., "crack"), $p_t = p$; when the true label is the negative class (e.g., "background"), $p_t = 1\text{-}p$.; α_t is a balancing factor, often set inversely proportional to class frequency. Typically, for the positive class, α (or α_t when y=1) is set to 0.75, and for the negative class, α is set to 0.25.; -ln(p_t) is the standard cross-entropy loss component; γ is the focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted. A larger γ value increases the suppression effect on easy-to-classify samples. It is typically set to 2.

(4) Dice Loss (DL) is designed to measure the overlap between the region predicted by the model and the ground truth segmentation region (Milletari et al., 2016). It is directly derived from the Dice Similarity Coefficient (DSC), a metric commonly used to evaluate segmentation performance. The core idea of Dice Loss is to directly optimize this overlap measure. It is particularly effective for tasks with imbalanced class distributions, such as medical image segmentation or crack detection, where the target region (e.g., a crack) occupies only a small fraction of the entire image. The Dice Loss is defined as:

$$L_{Dice} = 1 - Dice, \quad (9)$$

where

$$Dice = \frac{\left(2 \times \left|x \cap y\right|\right)}{\left(\left|x\right| + \left|y\right|\right)}, \quad (10)$$

where x represents the set of pixels predicted by the model as the "positive class" (e.g., crack). y represents the set of pixels in the ground truth label that belong to the "positive class".

(5) F1 Loss (F1L) is a loss function designed to directly optimize the F1-Score, which is a key metric for evaluating classification models, especially those trained on imbalanced datasets. By transforming the F1-Score into a differentiable loss function, it directly uses the evaluation metric as the optimization objective, ensuring that the model's training goal is highly consistent with the final evaluation criterion. This alignment often leads to improved performance on the target metric.

Given the characteristics of UAV-captured building facade images, where there is a severe class imbalance between crack and non-crack pixels, four loss functions are selected for this

study: Focal-Dice Loss (FDL), BCE-Dice Loss (BDL), F1 Loss (F1L), and Weighted Cross-Entropy (WCE). Focal Loss and Dice Loss are particularly suitable for handling the class imbalance inherent in crack data. Focal Loss addresses imbalance by down-weighting easy examples, while Dice Loss directly optimizes for spatial overlap, which is crucial for segmentation tasks. Combining them or using F1 Loss aims to directly improve the primary evaluation metric.

(6) Focal-Dice Loss (FDL) combines the advantage of Focal Loss in addressing severe class imbalance with the strength of Dice Loss in focusing on the precise contours of cracks. This combination often provides more stable gradient signals during training, which can lead to improved model convergence and performance. The composite loss function can be formulated as a weighted sum:

$$Focal-Dice-Loss = \alpha \times Focal\ Loss + \beta \times Dice\ Loss, \quad (11)$$

where α = 0.3 ~ 0.7, β = 1.0.

(7) BCE-Dice Loss (BDL) leverages the strengths of both Binary Cross-Entropy (BCE) and Dice Loss. The BCE component provides stable and smooth gradient signals, especially during the initial stages of training, which helps the model quickly establish fundamental classification capability. The Dice Loss component directly optimizes the overlap of the segmented regions; it is less sensitive to class imbalance and effectively improves the model's F1-score for the target class. This composite loss is typically formulated as a weighted sum of the two individual losses:

$$BCE-Dice-Loss = \alpha \times BCE\ Loss + \beta \times Dice\ Loss, \quad (12)$$

where the weighting coefficients can be adjusted based on the training focus: if greater emphasis is placed on the Dice component (e.g., when segmentation quality is prioritized), set α = 0.5 and β = 1.0; if greater emphasis is placed on the BCE component (e.g., when training stability is the primary concern), set α = 1.0 and β = 0.5.

3. Experiments and Results Analysis

3.1 Experimental Setup

This study selected over ten buildings constructed around the year 2000 for exterior wall crack detection experiments. Data acquisition was performed using a DJI Mavic 3T UAV. The image resolution was 4000 × 3000 pixels, with a Ground Sample Distance (GSD) of 0.03 cm/pixel. Automated route flights were conducted to capture data from the building facades, resulting in a total collection of 6933 high-resolution UAV aerial images. Among these, 4853 images were used for training, 1040 for validation, and 1040 for testing. Given the extreme class imbalance in the facade crack dataset, where crack pixels constitute a very low proportion, the original dataset exclusively included images containing cracks. Furthermore, data augmentation techniques such as geometric transformations and noise addition were applied to the original sample images to increase the sample size during model training and enhance the model's generalization capability.

The deep learning framework was implemented using Python 3.9 and PyTorch 1.13.1. The U-Net models had a depth of 4 and 2 convolutional layers per block. All experiments utilized the Adam optimizer. The training process employed a learning rate decay strategy to optimize network parameters. The initial learning rate was set to 0.0005, with a decay rate of 1e-6, dropout of 0.25, a batch size of 8, and training was conducted for 30 epochs. Four loss functions were selected to evaluate the training results of all models in this study: Focal-Dice Loss (FDL), BCE-Dice Loss (BDL), F1 Loss (F1L), and Weighted Cross-Entropy (WCE).

This paper employed a series of U-Net-based models for detecting cracks in building exteriors, as listed in Table 2.

Model	Composition	Description
Standard U-Net		Consists of a symmetric 4-layer encoder-decoder structure.
ResNet34 -UNet	ResNet34 + standard U-Net	Utilizes a ResNet34 encoder (fully residual encoder) and a standard U-Net decoder.
UNet-Residual	Standard U-Net + residual modules	Employs residual modules in the bottleneck layer (required), with optional use in the encoder and decoder.
UNet-Attention	Standard U-Net + attention modules	Uses attention modules in the skip connections (required), with optional dual-attention modules in the bottleneck layer and SE attention modules in the encoder.
HybridUNet	Standard U-Net + residual modules + attention modules	Uses standard convolutions in the encoder/decoder, residual modules in the bottleneck layer, and attention models in the skip connections.

Table 2. The U-Net series models.

3.2 Results Analysis

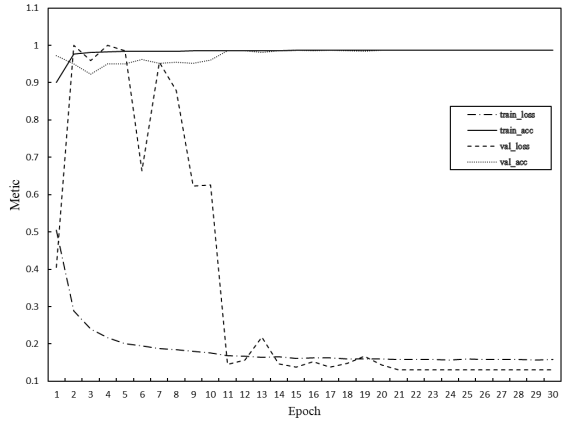
To compare the performance of the six typical models—Standard U-Net, ResNet34-UNet, UNet-Attention, UNet-Residual, HybridUNet, and TransUNet—a systematic evaluation was conducted from the following three aspects.

(1) Model Training Process Evaluation

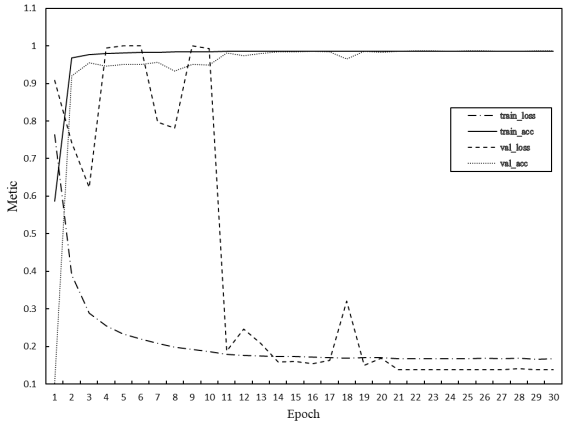
The training process of the U-Net models was evaluated using four parameters: training loss (train loss), training accuracy (train acc), validation loss (val loss), and validation accuracy (val acc). The training curves of three representative U-Net models — the Standard U-Net, the HybridUNet, and the TransUNet—are compared and illustrated in Figure 2 as a case study.

Figure 2 reveals distinct training dynamics among the three models. The Standard U-Net model demonstrates a relatively slow convergence rate, with its validation loss showing early signs of plateauing, which suggests limited learning capacity. In contrast, the HybridUNet model, which incorporates both residual and attention modules, exhibits a more stable descent in both training and validation loss, indicating improved gradient flow and feature learning. The TransUNet model, leveraging the global contextual understanding of the Transformer in its bottleneck, achieves the lowest final

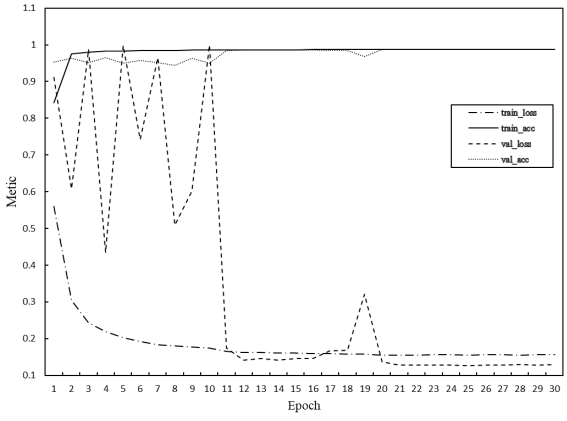
validation loss and highest validation accuracy among the three, highlighting its superior modelling capability. However, it also shows a slightly larger gap between training and validation accuracy compared to HybridUNet, hinting at a need for more careful regularization to prevent potential overfitting on smaller datasets. Overall, the integration of advanced modules correlates with enhanced learning efficiency and final performance in this task.



(a) U-Net model.



(b) HybridUNet model.



(c) TransUNet model.

Figure 2. Three U-Net model with F1L loss function and learning rate of 1e-4.

Analysis reveals that the training process exhibits distinct phase characteristics:

During the initial training phase (epoch <= 10), the validation loss shows short-term fluctuations. This is likely due to the model exploring the optimization direction and is a normal part of the learning process. As training progresses, these fluctuations quickly subside, indicating the model is gradually converging to a stable state.

After epoch 10, the training loss and validation loss converge synchronously, with the validation loss being lower than the training loss. Both training accuracy and validation accuracy stabilize and eventually converge to a high value (both Train Acc and Val Acc are greater than 98.5%).

In summary, this phenomenon indicates that the model has completed effective learning. It also verifies the rationality of the set training cycle (epoch=30), which ensures sufficient model training while avoiding unnecessary computational resource consumption. It is noteworthy that the loss values and accuracy metrics on the training and validation sets maintain a minimal and stable gap throughout, with val_loss < train_loss. This reflects consistent model performance on both training and validation data, demonstrating good generalization capability. Considering the stability, convergence, and generalization performance of the training process, it can be concluded that the model achieved an optimal training outcome and the model training parameters were set appropriately.

In the comparison of U-Net models, the parameters evaluating the training process loss and accuracy are key quantitative metrics, as shown in Table 3.

Model	Convergence performance (stable epoch)	Final accuracy (val_acc %)	Generalization ability (overfitting index)	Training stability (loss variance)	F1-Score (%)
U-Net	10	98.72	-0.0004	0.1126	87.38
Resnet34-UNet	20	98.43	0.0020	0.1207	84.38
UNet-Residual	10	98.73	-0.0005	0.1143	87.27
UNet-Attention	10	98.67	-0.0002	0.0745	86.93
HybridUNet	10	98.66	-0.0006	0.1219	86.67
TransUNet	10	98.76	-0.0006	0.1035	87.66

Table 3. A quantitative comparison of the training processes of six models.

As shown in the table above, among the six comparative models, TransUNet demonstrates the best overall performance. It not only achieved the highest validation set accuracy (98.76%) and the highest F1-score (87.66%), reflecting its ability to balance high accuracy and high completeness in the crack extraction task, but its training process also exhibited the smallest loss variance, indicating superior training stability. Furthermore, its overfitting index is negative and has the largest absolute value, suggesting the strongest generalization capability and an effective mitigation of overfitting. In summary, TransUNet not only leads in segmentation accuracy but also significantly outperforms other models in terms of training robustness,

generalization ability, and modelling of imbalanced samples, making it the optimal model choice considering comprehensive performance for the building facade crack detection task. Although the standard U-Net remains competitive in terms of simplicity and practicality, TransUNet, by virtue of its significant advantages across multiple key metrics, is the more recommended model for this scenario.

(2) Loss Function Evaluation

The loss function directly determines the search direction and convergence boundary of the algorithm in the parameter space. Especially in scenarios with imbalanced data like building crack detection, different loss functions guide the model towards distinct technical pathways. For instance, Focal Loss enhances sensitivity to fine cracks by adjusting the weights of hard examples; the Dice series focuses on optimizing the spatial overlap between the predicted region and the ground truth, often achieving higher recall but at the risk of increased false detections; while F1 Loss strives to find the Pareto optimum between precision and recall. These characteristics make the loss function not merely a mathematical tool for measuring error, but a crucial link connecting algorithmic optimization to engineering value. When a project requirement dictates "rather having false detections than missing cracks," a recall-oriented loss function should be chosen; when the reliability of detection results is paramount, an optimization strategy prioritizing precision is necessary. Therefore, a deep understanding and appropriate selection of the loss function is a key step in achieving precise alignment between algorithm performance and business needs.

Model	Loss	Val					
		Accuracy (%)	F1 (%)	IoU (%)	Precision (%)	Recall (%)	Loss
U-Net	F1 Loss	98.72	87.38	77.59	87.68	89.17	0.1298
	BCE-Dice-Loss	98.45	86.29	75.79	84.32	96.50	0.1328
	Focal-Dice-Loss	98.54	86.71	76.44	91.90	94.55	0.0989
HybridUNet	F1 Loss	98.66	86.67	76.44	87.20	86.24	0.1375
	BCE-Dice-Loss	98.41	85.62	74.74	81.03	94.60	0.1429
	Focal-Dice-Loss	98.44	86.31	75.88	80.36	95.92	0.1052
TransUNet	F1 Loss	98.73	87.50	78.02	86.82	88.26	0.1287
	BCE-Dice-Loss	98.20	85.19	74.14	77.10	99.10	0.1363
	Focal-Dice-Loss	98.41	86.34	75.89	84.83	96.05	0.1008

Table 4. A Comparative analysis of loss functions in three classical U-Net models.

Detailed analysis is as follows:

(1) Standard U-Net Model

Analysis of the three loss functions for the Standard U-Net model reveals distinct characteristics: F1 Loss demonstrated the best overall performance, achieving the highest F1-score (87.38%) and IoU (77.59%), while maintaining the optimal balance between Precision (87.68%) and Recall (89.17%). This highlights its ability to harmonize classification accuracy and spatial localization precision. In contrast, BCE-Dice-Loss ensured comprehensive crack coverage with a higher Recall (96.50%), but at the cost of a significant drop in Precision (84.32%), resulting in relatively lower F1-score (86.29%) and IoU (75.79%). Meanwhile, Focal-Dice-Loss exhibited unique advantages, achieving the highest Precision (91.90%) and the lowest loss value (0.0989) while maintaining a high Recall (94.55%). However, its F1-score (86.71%) and IoU (76.44%) still slightly trailed those of F1 Loss. From the metric relationships, all three showed the pattern F1 > IoU, confirming that IoU, as a spatial overlap metric, is more sensitive to boundary errors. In practical applications, F1 Loss should be chosen for overall performance, BCE-Dice-Loss if prioritizing complete crack detection, and Focal-Dice-Loss if prediction accuracy and training stability are paramount.

(2) HybridUNet Model

For the HybridUNet model, the three loss functions showed different traits: When using F1 Loss, the model achieved the highest F1-score (86.67%) and IoU (76.44%), alongside the best balance between Precision (87.20%) and Recall (86.24%), indicating this loss function's effectiveness in coordinating classification accuracy and boundary localization precision. Conversely, BCE-Dice-Loss, while achieving the highest Recall (94.60%) and significantly reducing the risk of missed detections, suffered a substantial drop in Precision (81.03%), leading to the lowest F1-score (85.62%) and IoU (74.74%), reflecting its inadequacy in controlling false detections. Focal-Dice-Loss performed best in terms of Recall (95.92%) and loss value (0.1052), suggesting its advantages in handling class imbalance and optimization stability. However, its unsatisfactory Precision (80.36%) limited further improvement in F1-score (86.31%) and IoU (75.88%). In summary, the F1-score and IoU showed high consistency, both validating the superior overall performance of F1 Loss. Nevertheless, if the practical application is extremely sensitive to missed detections, Focal-Dice-Loss, with its higher recall and stable optimization, is a worthwhile alternative.

(3) TransUNet Model

The three loss functions for the TransUNet model showed distinct performance differentiations: F1 Loss achieved the best overall performance with an F1-score of 87.50% and IoU of 78.02%, while maintaining the optimal balance between Precision (86.82%) and Recall (88.26%), demonstrating excellent segmentation coordination capability. Although Focal-Dice-Loss was slightly inferior in comprehensive metrics (F1 86.34%, IoU 75.89%), it exhibited outstanding optimization stability (lowest loss value of 0.1008) and crack coverage capability (excellent Recall of 96.05%). Meanwhile, BCE-Dice-Loss achieved a near-perfect Recall (99.10%), but suffered a significant drop in Precision (77.10%), resulting in the lowest F1-score (85.19%) and IoU (74.14%). The substantial 11.05 percentage point gap between its F1-score and IoU further highlights the severe negative impact of numerous false

detections on the spatial overlap metric. In conclusion, for crack detection tasks, F1 Loss is most suitable for scenarios pursuing balanced overall performance; Focal-Dice-Loss strikes a good balance between stability and high recall requirements; whereas BCE-Dice-Loss is only applicable for special working conditions with extreme demands on low missed detection rates.

In the comprehensive evaluation across the three U-Net architecture variants, the choice of loss function revealed a highly consistent performance pattern: Whether for the Standard U-Net, HybridUNet, or the top-performing TransUNet, F1 Loss consistently demonstrated the best balanced performance, achieving the highest F1-scores (87.38%, 86.67%, and 87.50% respectively) and IoU metrics (77.59%, 76.44%, and 78.02% respectively) on their respective models. This underscores its exceptional ability to coordinate between classification accuracy and boundary localization precision. Focal-Dice-Loss exhibited excellent optimization stability (lowest loss value reaching 0.1008) and high-recall characteristics (up to 96.05%) across all three model types, but its insufficient Precision limited further improvement in the comprehensive metrics. BCE-Dice-Loss, while achieving very high Recall (up to 99.10%) across all three models, did so at the cost of a significant drop in Precision, leading to markedly lower F1-scores and IoU. This was particularly evident in the TransUNet, where the gap between the F1-score and IoU reached 11.05 percentage points, highlighting that it is only suitable for special scenarios with extreme requirements for low missed detections. Overall, TransUNet combined with F1 Loss achieved the best performance among all model-loss function combinations (F1 87.50%, IoU 78.02%), validating the synergistic advantages of the Transformer architecture's global modeling and F1 Loss's balanced optimization, providing a best-practice solution for the crack detection task.

(3) Prediction Results for Exterior Wall Cracks

To intuitively evaluate the generalization performance of the six models, a qualitative comparative analysis was conducted on a unified dataset, with results shown in Figure 4. In the figure, red rectangles indicate areas of missed crack detection, while green rectangles highlight areas of false detection. It is noteworthy that the visualization results show some differences from the quantitative analysis based on the entire dataset: When dealing with complex exterior wall crack scenarios, UNet-Residual and UNet-Attention demonstrated more complete crack capture capability and lower false detection rates, with their prediction results aligning most closely with the ground truth. In contrast, the structurally more complex TransUNet and Resnet34-UNet showed significant false detections in some areas with complex textures. The Standard U-Net exhibited unstable detection performance, with numerous errors in fine crack areas. This phenomenon reveals an adaptation gap between the models trained on the data distribution and the real-world complex scenes: some models that perform excellently on overall metrics might underperform in specific complex scenarios due to overfitting to features in the training set. These findings not only deepen the understanding of the crack semantic comprehension capabilities of different models but also emphasize the need for model selection based on specific scene characteristics in practical applications, providing important references for model optimization aimed at complex environments.

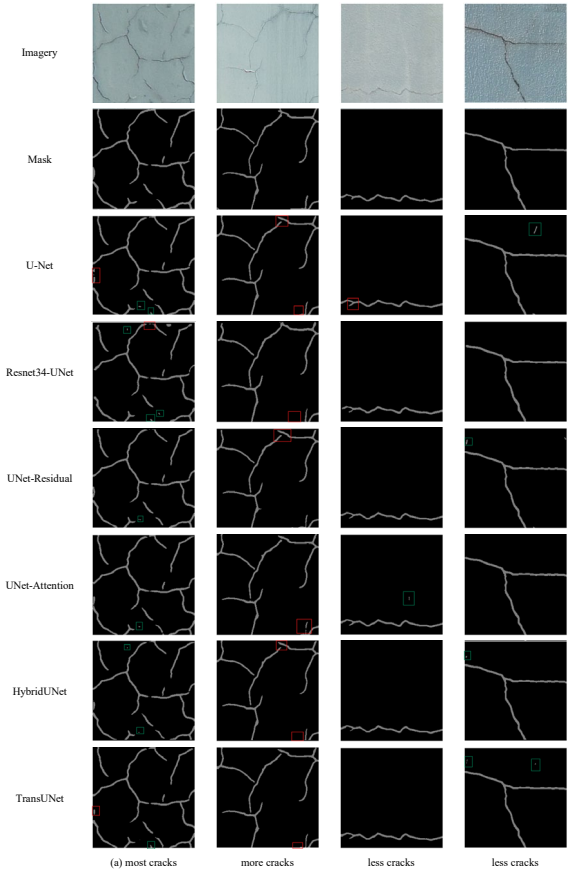


Figure 3. Crack detection results from different models.

In summary, while TransUNet maintains a leading position in overall performance, its advantage is relatively limited. Furthermore, the performance differences among the models in complex scenarios are also quite subtle. Although TransUNet exhibited some false detections in areas with complex textures, and the structurally simpler UNet-Residual and UNet-Attention demonstrated slightly better robustness, none of the models established a decisive performance gap. This outcome indicates that, despite having distinguishable design orientations in feature extraction and generalization capabilities, the practical performance of the various models on the current task converges, with no single solution demonstrating absolute dominance. The choice of model may therefore depend more heavily on specific application constraints, such as computational resources or tolerance for certain types of error, rather than a clear-cut performance superiority.

4. Conclusion

This study, through systematic experimental investigation, reveals the performance characteristics and applicability of different deep learning models for the task of building exterior wall crack detection. The main conclusions are as follows:

(1) Regarding the relationship between model architecture and performance, the study finds that although TransUNet achieved optimal comprehensive metrics due to the global context modeling capability of its Transformer module, the performance differences among the six comparative models

were limited (F1 Score gap < 0.5%). This indicates that, within the context of the current task, increasing model complexity did not yield significant performance gains. Relatively simpler models like the standard U-Net and its UNet-Attention variant maintained high accuracy while offering better practical utility for engineering applications.

(2) Concerning the selection of loss functions, F1 Loss consistently demonstrated the best balanced performance across the three primary model architectures. Focal-Dice-Loss excelled in optimization stability, while BCE-Dice-Loss proved suitable for specific scenarios with extreme requirements for low missed-detection rates. This pattern provides clear guidance for selecting loss functions tailored to different application scenarios.

(3) In terms of methodological contribution, the six-model comparative framework and the systematic design of ablative experiments established a reproducible benchmark for model evaluation in the field of building crack detection. The research identified a significant complementary relationship between the global modeling capacity of Transformer and the local feature extraction capability of U-Net. The effective integration of these two aspects is key to enhancing crack detection accuracy.

The limitations of this study primarily lie in the limited diversity of the dataset's background. The generalization capability of the models under extreme lighting conditions and on wall surfaces with special materials still requires further validation. Furthermore, the current research has not yet fully considered the balance between computational efficiency and energy consumption in practical deployment scenarios.

Based on the findings, future work will focus on the following research directions: constructing more diverse datasets of building exterior wall cracks; developing lightweight and efficient model architectures that balance accuracy and computational cost; exploring cross-domain adaptation methods to improve model generalization across different building types and environmental conditions; and promoting the translation of research findings into practical engineering applications to provide more reliable technical support for building structural health monitoring.

Acknowledgements

This work was supported by National Key R&D Program of China ‘‘Joint Research, Development and Application Demonstration of Remote Sensing Monitoring Technology for Typical Natural Resources Features’’ (Grant No. 2023YFE0207900).

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. In: *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, Glasgow, UK, pp. 213-229.

Cai, T.C., Liu, C., Zhou, X.T., Xu, Z.R., and Chen, C., 2022. Detection of Cracks in Building Façade Based on UAV. *Geotechnical Investigation & Surveying*, (4), pp. 45-51 (in Chinese).

Chang, H., Rao, Z.Q., Zhao, Y.L., and Li, Y.C., 2021. Research on Tunnel Crack Segmentation Algorithm Based on Improved U-Net Network. *Computer Engineering and Applications*, 57(22), pp. 215-222 (in Chinese).

Chaiyasarn, K., Sharma, M., Ali, L., Khan, W., and Poovarodom, N., 2018. Crack Detection in Historical Structures Based on Convolutional Neural Network. *International Journal of Geomate*, 15(51), pp. 240-251.

Chen, J., Gao, Y., Guo, C.Y., Tang, J.H., Liao, X.H., Jiang, J., Zhang, S.Q., and Liu, W.Z., 2025. Harnessing 3D Realistic Geospatial Landscape Model to Empower the Low-Altitude Economy: Fundamental Problems and Major Tasks. *Journal of Spatio-temporal Information*, 32(1), pp. 1-10 (in Chinese).

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., and Zhou, Y., 2021. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5), pp. 6393-6405.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations* (ICLR).

He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), Las Vegas, NV, USA, pp. 770-778.

Hsieh, Y.A., and Tsai, Y.J., 2020. Machine Learning for Crack Detection: *Review and Model Performance Comparison. Journal of Computing in Civil Engineering*, 34(5), 04020038.

Hu, J., Shen, L., Albanie, S., Sun, G., and Wu, E., 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8), pp. 2011-2023.

Hu, Q., Chen, H.D., Li, Y.J., and Li, H.K., 2023. A Review of Structural Crack Detection Based on Deep Learning. *Jiangxi Hydraulic Science & Technology*, 49(4), pp. 241-246 (in Chinese).

Ibtehaz, N., and Rahman, M.S., 2020. MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation. Neural Networks, 121, pp. 74-87.

Ji, A., Xue, X., Wang, Y., Luo, X., and Xue, W., 2020. An Integrated Approach to Automatic Pixel-Level Crack Detection and Quantification of Asphalt Pavement. *Automation in Construction*, 114, 103176.

Li, C.Y., Zhang, C., and Wei, H.D., 2023. Research on Road Crack Segmentation Method Based on Multi-Scale Feature Fusion Network. *Journal of Spatio-temporal Information*, 30(3), pp. 425-430 (in Chinese).

Lian, H.J., Wang, W.G., Zhu, J., Tang, R.R., and Xie, Y.K., 2023. A Highway Extraction Method from High-Resolution

Images Based on Improved U-Net. *Journal of Spatio-temporal Information*, 30(3), pp. 335-344 (in Chinese).

Li, W.S., Zhang, J., Zhuo, L., Wu, X.J., and Yan, Y., 2024. Overview of Transformer-Based Visual Segmentation Techniques. *Chinese Journal of Computers*, 47(12), pp. 2760-2782 (in Chinese).

Lin, T.Y., Goyal, P., Girshick, R., He, K., and Dollár, P., 2017. Focal Loss for Dense Object Detection. In: Proceedings of the *IEEE International Conference on Computer Vision* (ICCV), Venice, Italy, pp. 2980-2988.

Liu, C., Akbar, A., and Cai, T.C., 2022. UAV Autonomous Inspection and Crack Detection Towards Building Health Monitoring. *Journal of Tongji University* (Natural Science), 50(7), pp. 921-932 (in Chinese).

Liu, X.H., Li, Y., Ge, Y., Wang, H.Y., Lai, M.Y., Gu, Z.R., Chu, S.M., and Ding, H., 2025. An Improved U-Net for Extraction of Surface Water Resources from Remote Sensing Images in Arid Regions. *Journal of Spatio-temporal Information*, 32(2), pp. 158-167 (in Chinese).

Liu, X.G., Chen, Y.Y., Zhu, A.X., Yang, J., and He, G.H., 2018. Tunnel Crack Identification Based on Deep Learning. *Journal of Guangxi University* (Natural Science Edition), 43(6), pp. 2243-2251 (in Chinese).

Liu, Y.H., Yao, J., Lu, X.H., Xie, R.P., and Li, L., 2019. DeepCrack: A Deep Hierarchical Feature Learning Architecture for Crack Segmentation. *Neurocomputing*, 338, pp. 139-153.

Loverdos, D., and Sarhosis, V., 2022. Automatic Image-Based Brick Segmentation and Crack Detection of Masonry Walls Using Machine Learning. *Automation in Construction*, 140, 104316.

Ma, X., Zhang, X., Pun, M., and Liu, M., 2024. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 5403215.

Milletari, F., Navab, N., and Ahmadi, S.A., 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: *Proceedings of the 2016 Fourth International Conference on 3D Vision* (3DV), Stanford, CA, USA, pp. 565-571.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., and Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas. In: *Medical Imaging with Deep Learning* (MIDL), Amsterdam, The Netherlands.

Park, J., Woo, S., Lee, J., and Kweon, I.S., 2018. BAM: Bottleneck Attention Module. In: *Proceedings of the British Machine Vision Conference* (BMVC), Newcastle, UK, p. 147.

Qiao, P., Liang, Z.Q., Duan, C.J., Ma, C., Wang, S.L., and Di, J., 2024. Bridge Defects Detection and Quantifying Method Based on Modified Faster R-CNN and U-Net. *Journal of Southeast University* (Natural Science Edition), 54(3), pp. 627-638 (in Chinese).

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S., 2019. Generalized Intersection over Union: A Metric and a Loss for Bounding Box Regression. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, CA, USA, pp. 658-666.

Ronneberger, O., Fischer, P., and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Munich, Germany, pp. 234-241.

Shamsabadi, E.A., Chang, X., Rao, A.S., Nguyen, T., and Ngo, T., 2022. Vision Transformer-Based Autonomous Crack Detection on Asphalt and Concrete Surfaces. *Automation in Construction*, 140, 104316.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C., 2021. Segmenter: Transformer for Semantic Segmentation. arXiv preprint, arXiv:2105.05633.

Tian, Z., He, T., Shen, C., and Yan, Y., 2019. Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), Long Beach, CA, USA, pp. 3126-3135.

Weng, P., Lu, Y.H., Qi, X.B., and Yang, S.Y., 2019. Pavement Crack Segmentation Technology Based on Improved Fully Convolutional Networks. *Computer Engineering and Applications*, 55(16), pp. 235-239, 245 (in Chinese).

Xiao, X., Lian, S., Luo, Z., and Li, S., 2019. Weighted Res-UNet for High-Quality Retina Vessel Segmentation. In: *Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education* (ITME), Hangzhou, China, pp. 327-331.

Xu, Z.Y., Zhou, S.G., Ge, Y., and Wan, Z.H., 2025. Automatic Extraction of Unlabelled Urban Roads from Remote Sensing Images Based on Model Distillation. *Journal of Spatio-temporal Information*, 32(2), pp. 113-126 (in Chinese).

Yang, X., Li, H., Yu, Y., Luo, X., Huang, T., and Yang, X., 2018. Automatic Pixel-Level Crack Detection and Measurement Using Fully Convolutional Network. *Computer-Aided Civil and Infrastructure Engineering*, 33(12), pp. 1090-1109.

Zhang, Z.R., Fan, D.Z., Ji, S., Dong, Y., Li, D.Z., and Liu, J., 2024. Stereo Matching of Satellite Remote Sensing Images Based on Attention Mechanism. *Journal of Spatio-temporal Information*, 31(1), pp. 41-49 (in Chinese).

Zhu, P.R., Xue, R.Z., Liu, M.M., Wang, Y.M., Yin, J.L., Zhang, G., and Wang, X., 2024. Research and Development of Non-Contact High-Precision Monitoring Equipment for Concrete Cracks in Port Construction Buildings. *Journal of Waterway and Harbor*, 45(4), pp. 518-525 (in Chinese).

Zhu, S.Y., Du, J.C., Li, Y.S., and Wang, X.P., 2019. Bridge Crack Detection Method Using U-Net Convolutional Network. *Journal of Xidian University*, 46(4), pp. 35-42 (in Chinese).