Membrane Technology
ISSN (online): 1873-4049

A Residual-Self-Attention Fusion Network for Identification and Classification of Apparent Defects in Fair-faced Concrete

Bing Ling[1], Yan Li[1,*], Zhexiong Shang[1], Long Dang[2]
[1]*School of Civil and Architecture Engineering, Hainan University, Haikou, China*
[2]*Hainan Haikong Zhongneng Construction Engineering Co., Haikou, China*
**Corresponding Author.*

Abstract:
This paper proposes a fair-faced concrete surface defect classification method based on a residual-self-attention deep fusion network (RSAFuser), aiming to solve the issues of low accuracy and high time consumption in the task. The method uses an "Expert Voting"-like feature extraction approach to identify surface defects of fair-faced concrete, such as Pores, Surface Contamination, Cracks, and Repairs. The residual expert network handles small-area local point defects, such as Pores and Repairs, while the self-attention expert network processes larger-area global linear and block-shaped defects, like cracks and Surface Contamination, extracting and reflecting the overall features of the input image from different perspectives. The fused network not only retains the advantages of the residual network in local information recognition and mitigating network degradation, but also uses the self-attention network for block processing and effectively captures global long-distance information through weight calculation. Experimental results show that compared to ResNet and Swin-Transformer, RSAFuser improves accuracy by 19.19% and 21.51%, respectively. Experimental results validate the effectiveness and accuracy of the proposed method in the fair-faced concrete defect classification task.

Keywords: fair-faced concrete; surface defect recognition and classification; self-attention; residual network; expert voting.

INTRODUCTION

Fair-faced concrete is widely used in landmark buildings, such as the Lenovo Research and Development Center [1], Jingang Cultural in China and the Wolfsburg Pheno Science Center in Germany, due to its unique texture and aesthetic appeal [2-3]. However, due to the lack of external protection on fair-faced concrete surfaces, defects such as Pores, cracks, Surface Contamination, and artificial repairs are common during and after construction [4-5]. Due to significant differences in defect types, traditional classification methods struggle to balance local and global features, resulting in low efficiency when a single network model is used. Therefore, there is an urgent need to develop more intelligent and efficient defect classification methods. With the advancements in deep learning, surface defect classification of concrete has become feasible [6-8]. Kim et al. [9] used machine learning to classify cracks and non-cracks on concrete surfaces, reducing manual inspection costs, but their method is limited to a single defect type and does not account for defect diversity. Fan et al. [10] used support vector machine-based image clustering to detect multiple damages in reinforced concrete, achieving an accuracy of 94.9%. However, the clustering algorithm has weak damage-type determination ability and is highly sensitive to parameters. With advancements in deep learning, convolutional neural network (CNN)-based defect recognition methods have gradually been applied to the concrete field [11-12]. Zhu et al. [13] utilized the VGGNet16 model for precise classification of surface defects on cement concrete bridges, reaching an accuracy of 83.03%. However, due to the large number of parameters in VGGNet, long training times, and susceptibility to overfitting [14], its practical applications are limited. Consequently, researchers have turned to deeper and more efficient residual networks (ResNet). Li et al. [15] introduced an automated crack classification method for concrete dam structures using deep residual neural networks. Liu et al. [16] enhanced concrete defect recognition by fine-tuning the ResNet50 network. However, since ResNet primarily focuses on local feature learning, it struggles to effectively capture global features [17].

In recent years, the Self-Attention mechanism has demonstrated exceptional performance in modeling long-range dependencies and has been widely applied in image recognition tasks [18]. For instance, Vision Transformer (ViT) [19] successfully learns global image features by dividing images into patches and leveraging the self-attention mechanism to capture inter-patch dependencies. However, the high computational complexity of ViT limits its applicability. To improve computational efficiency while maintaining strong performance, Microsoft Research introduced the Swin Transformer model [20], which utilizes a hierarchical shifting window mechanism to significantly reduce computational complexity and enhance recognition accuracy, making it a new benchmark in self-attention networks for computer vision.

The success of Swin Transformer provided valuable insights for further research on fair-faced concrete surface defect detection. By applying its hierarchical shifted window mechanism, we can better achieve the combination of local and global feature

extraction, effectively addressing the complex defect detection requirements of fair-faced concrete surfaces. In fair-faced concrete surface defect detection, capturing both local and global information is equally important. For example, in detecting pore defects, the detection network should focus on the local details of the image (as shown in Figure 1(a)); whereas, in detecting crack defects, the network needs to have the ability to capture long-range dependencies (as shown in Figure 1(b)). Therefore, this paper proposes a Residual-Self-Attention Fusion Network (RSAFuser), which combines the advantages of residual networks and self-attention networks. It effectively captures local features while also grasping global dependencies, overcoming the limitations of a single model in detecting fair-faced concrete defects, and providing a new solution to improve detection accuracy and efficiency.

(a) Holes (b) Cracks

Figure 1. Apparent defects in fair-faced concrete

THE OVERALL ARCHITECTURE OF THE RSAFUSER NETWORK

With the development of automatic concrete surface defect recognition technology, effectively combining local detail features with global structural information has become crucial for improving recognition accuracy and model robustness. To address this, this paper proposes the RSAFuser fusion network model, which integrates residual networks and self-attention networks to extract image features from both local and global levels, and implements dynamic feature fusion through an expert voting module.

RSAFuser Model

The RSAFuser model consists of three components: (1) a local information residual block, serving as the local feature expert (Local Expert); (2) a multi-head self-attention module, acting as the global feature expert (Global Expert); and (3) an expert voting module, which enables dynamic weighted fusion of local and global feature information. The structure and key modules of the RSAFuser model are discussed in detail below.

Residual expert network

The Residual Expert Network module, based on the ResNet architecture, is primarily responsible for extracting local detail features from images, such as those in Pores and repair areas. Through convolution operations, it effectively captures texture features in the input image, while the design of the residual block, shown in Figure 2, addresses the vanishing gradient problem in deep networks, ensuring the effective extraction of local features.

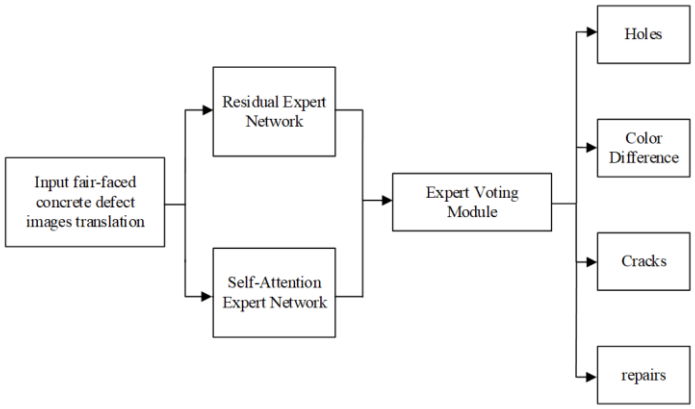


Figure 2. RSAFuser flowchart

The residual expert network module, based on the ResNet architecture, focuses on extracting local detail features from images, such as Pores and patch areas. By using convolution operations, it effectively captures texture features from the input image. Additionally, the design of residual blocks (shown in Figure 3) alleviates the vanishing gradient problem in deep networks, ensuring the effective extraction of local features.

To enhance the modeling capability of local features, the ResNet branch receives high-resolution input images without significant downsampling, retaining rich detail information. In addition, during training, the network uses sliding windows or random cropping to input local image patches, thereby strengthening its ability to learn local regions and reducing the interference of global information in feature extraction. Moreover, region-level annotation information is incorporated to indicate the location of local defects, further improving the accuracy of local feature extraction. The mathematical expression of the residual block is shown in Equation (1).

$$y=F\left((X,W_R\right)+x \tag{1}$$

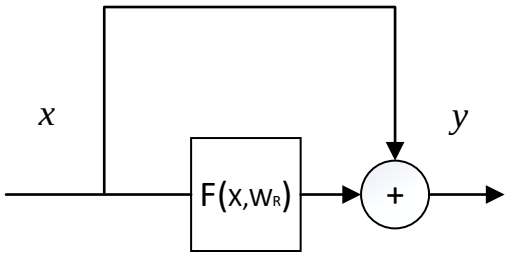


Figure 3. Schematic diagram of the residual block

Here, x represents the hierarchical input, $F(X,W_R)$ denotes the convolution operation applied to the input x, and W_R represents the weight parameters of the convolution layer.

Self-Attention expert network

The Multi-Head Self-Attention Module [21] focuses on extracting global features of images, which is suitable for identifying large-scale defects like cracks and Surface Contamination, distinguishing it from ResNet. By calculating pixel relationships across the entire image, the model captures global structural information and long-range dependencies. To ensure the transformer focuses on global feature extraction, the following strategies are adopted. (1) Input downsampling and block processing. (2) Random masking during training, where parts of regions or specific pixels are randomly masked to encourage the network to learn long-range dependencies, rather than being confined to local features. (3) Global loss guidance, which directs the model to learn feature representations related to global defects. The core of this mechanism involves parallel computation of multiple attention heads, allowing the model to simultaneously capture information from different subspaces, thereby enhancing its ability to model global features. The computation can be broken down into the following steps:

(1) Mapping of queries, keys, and values.

The input sequence $X= [x_1,x_2,\ldots,x_n,]$ is mapped to query, key, and value spaces using distinct weight matrices.

$$K_i=XW_i^K,\ Q_i=XW_i^Q,\ V_i=XW_i^V,\text{for=i=1},\ 2,\ \ldots,\ h \tag{2}$$

Here, W_i^K, W_i^Q, and W_i^V are the weight matrices for each head, h denotes the number of heads, and d_k represents the dimensionality of the queries and keys for each head.

(2) Calculating attention for each head.

For each head iii, the attention representation is computed as:

$$\text{Attention}\left(Q_i,K_i,V_i\right)=\text{softmax}\left(\frac{Q_i\cdot K_i^T}{\sqrt{d_k}}\right)\cdot V \tag{3}$$

Here, the similarity is computed using the dot product of queries and keys, followed by normalization using the softmax function, yielding attention weights that are then multiplied with the value matrix V_i to obtain the weighted representation.

(3) Concatenating the outputs of all heads.

The outputs of all heads are concatenated as follows:

$$\text{Concat}(\text{Attention}_1,\text{Attention}_2\ldots,\text{Attention}_h) \quad (4)$$

The concatenated result has dimensions $n\times h\cdot d_v$

(4). Linear Transformation

A linear transformation is applied to the concatenated output to obtain the final multi-head self-attention output:

$$\text{Output}(\text{Attention}_1,\ldots,\text{Attention}_h WW^O \quad (5)$$

Here, W^O represents the output weight matrix, with dimensions $h\cdot d_v\times d_{model}$

Expert voting module

The Expert Voting Module, as one of the core components of the RSAFuser model, primarily functions to integrate the feature outputs from the Local Residual Network and the Self-Attention Expert Network, thereby providing robust feature representations for classification tasks. This integration mechanism is derived from the mixed expert model but has been optimized to meet the specific requirements of image feature fusion. The detailed working principles and processes are as follows:

(1) Feature Input. The Local Expert Network extracts features represented as f_{local}, primarily encompassing detailed image features such as Pores and repair areas. The Global Expert Network extracts features represented as f_{global}, primarily encompassing overall structural features of the image, such as cracks and color variations. These two types of features collectively serve as inputs to the EV-Module.

(2) Weight Allocation. The EV-Module dynamically assigns weights to the local and global features, denoted by μ and v respectively. The weight allocation adheres to the following constraint:

$$\mu+v=1 \quad (6)$$

Here, μ represents the weight of the local features, and v represents the weight of the global features.

(3) Dynamic Weight Learning. Through an adaptive mechanism, the EV-Module dynamically adjusts μ and v based on the characteristics of the input image. Specifically, when the image defects are predominantly local, such as densely distributed Pores, the EV-Module assigns a larger μ value to enhance the contribution of local features. Conversely, when the defects are predominantly global, such as cracks spanning the entire image, a larger v value is assigned to enhance the contribution of global features.

Information Encoding and Fusion

The RSAFuser network employs the Residual Expert Network and the Self-Attention Expert Network to perform parallel encoding of both local and global information from Fair-faced concrete defect images. This results in the generation of local features f_{Local} and global features f_{Global}. These features represent a multi-scale characterization of the input image, capturing both detailed textures and overall structural information. In the later stages, the Information Fusion Module deeply integrates these highly abstracted features to generate a latent vector f_{Latent} which is used for the final classification task.

Local and global feature fusion mechanism

In the Information Fusion Module, local and global features are combined into a unified high-dimensional feature representation through feature stacking. This design of feature stacking maintains the original feature information while enabling subsequent convolution operations to extract cross-dimensional joint feature relationships, thereby fully integrating the advantages of both local and global features. The fused features are represented as shown in Equation (7).

$$f_{Concat}=\text{Concat}(f_{Local,}f_{Global}) \quad (7)$$

Here, f_{Concat} represents the feature representation after stacking, and $\text{Concat}(\cdot)$ denotes the feature stacking operation that concatenates local and global features along the channel dimension.

Non-linear feature fusion

To further explore the potential relationships between local and global features, the information fusion module employs a non-linear convolutional neural network to jointly model the stacked features. Compared to traditional linear feature fusion methods,

such as simple weighted addition, non-linear convolution operations can capture complex relationships between features, thereby avoiding redundancy caused by simple information addition. The non-linear feature fusion is illustrated in Equation (8).

$$f_{Latent}=Conv(f_{Concat,}) \tag{8}$$

Here, f_{Latent} represents the latent vector after fusion, which is a high-level feature representation processed through convolution. Conv(·) denotes the convolution operation, which extracts local and global correlations between the stacked features using convolutional kernels.

Classification output

The fused latent vector f_{Latent} is input into the subsequent classification module, where high-dimensional features are mapped to a probability distribution through a softmax activation function, thereby completing the final prediction of defect categories. This is illustrated by Equation (9).

$$P(y=c|f_{Latent})=\frac{\exp(f_{Latent},c)}{\sum c'^{(\exp(f_{Latent},c')}} \tag{9}$$

Here, $P(y=c|f_{Latent})$ represents the predicted probability that an image belongs to defect category c, and f_{Latent} denotes the feature value in the latent vector associated with category c.

RSAFuser Network Classification Module

The classification module of the RSAFuser network consists of the following key steps: local feature extraction, global feature extraction, the Expert Voting Module, and the classifier, as illustrated in Figure 4. This module provides efficient feature representation and prediction for classifying Fair-faced concrete defects by integrating both local and global features.

Local information extraction branch

As shown in Figure 4, the left-side local information extraction branch uses the classic ResNet-152 network. The network consists of four depth-increasing Residual Blocks. After each residual block, the input image size is reduced to $\frac{1}{2^i}$ of the original size, where $i\in\{1,2,3,4\}$

The layer-by-layer downsampling mechanism ensures the progressive extraction of local detailed features. Each residual block first extracts local features through multiple convolutional layers. Simultaneously, shortcut connections are employed to retain input information, preventing the vanishing gradient problem. Finally, a global average pooling layer compresses the spatial information of the local features into a fixed-size vector, yielding the local feature f_{Local}.

Global information extraction branch

As shown in Figure 4, the right-side global information extraction branch on the right utilizes a self-attention-based network architecture composed of four self-attention modules. It performs dimensionality reduction and patching operations on the input image. This module captures long-range dependencies between pixels across the entire image. First, the dimensionally reduced image features are fed into the self-attention modules, which compute global correlations between pixels using a multi-head self-attention mechanism. Next, the generated features further extract overall pattern information during subsequent processing. For example, features such as crack penetration and Surface Contamination distribution are extracted, and ultimately, the global feature vector f_{Global} is produced through the integration of global features.

Classifier design

The fused classification feature f_{cls} is fed into a classifier for defect category prediction. The classifier consists of a fully connected layer and a Softmax layer. First, f_{cls} is mapped to an output dimension consistent with the number of defect categories through the fully connected layer, as shown in Equation (10).

$$Y_c=W_c\cdot f_{cls}+b_c \tag{10}$$

Here, Y_C represents the predicted score for the C-th defect category, W_c denotes the weight matrix of the fully connected layer, and b_c refers to the bias term of the fully connected layer.

The output scores from the fully connected layer are then converted into a probability distribution using the Softmax function, which characterizes the predicted probability for each class, as shown in Equation (11).

$$P(y=c|f_{cls})=\frac{\exp(\hat{y}_c)}{\sum c'^{(\exp(\hat{y}_{c'}))}} \quad (11)$$

Here, $P(y=c|f_{cls})$ denotes the probability that the image belongs to the c-th defect category, and $\sum c'^{(\exp(\hat{y}_{c'}))}$ represents the normalization term, which ensures that the sum of predicted probabilities for all categories is 1.

Loss function optimization

The error between the model's predicted output $P(y=c|f_{cls})$ and the true class label y_c is calculated using the cross-entropy function, as shown in Equation (12).

$$Loss=-\sum_c y_c \cdot \log(P(y=c|f_{cls})) \quad (12)$$

Here, y_c represents the true label of the actual class, with 1 indicating membership in the class and 0 indicating non-membership, and $P(y=c|f_{cls})$ represents the predicted class probability by the model.

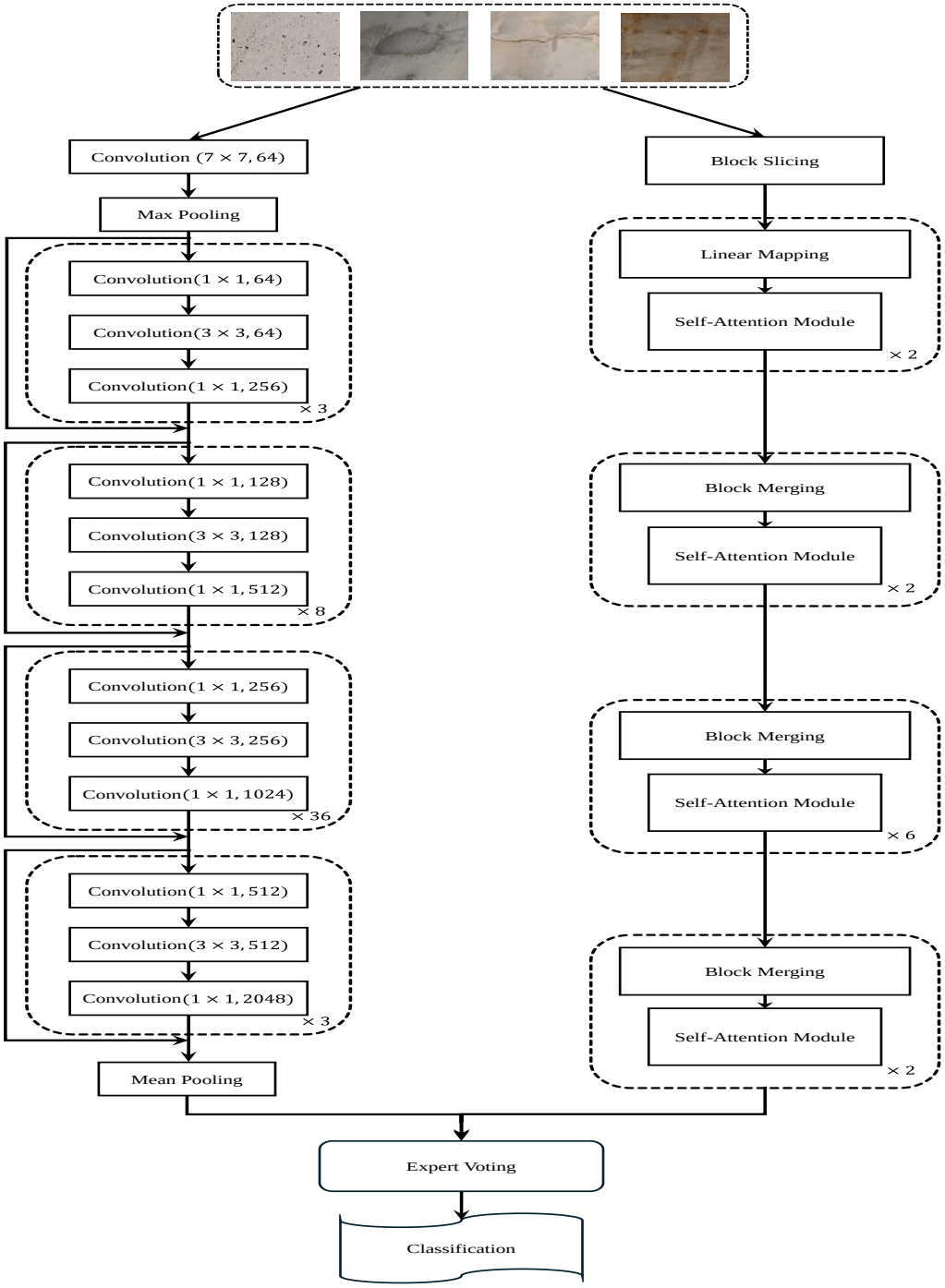


Figure 4. RSAFuser network structure

EXPERIMENT DESIGN AND RESULTS ANALYSIS

Dataset Collection and Processing

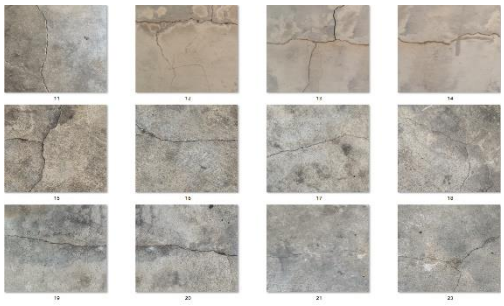
The quality of image collection for fair-faced concrete surface defects directly impacts subsequent defect identification and classification. To guarantee the data's reliability and consistency, this study established strict image collection standards, which include equipment selection and parameter settings. The Sony ZV-E10L camera was chosen as the primary collection device, paired with the Godox V860III flash. The camera settings were as follows: ISO set to 400, shutter speed to 1/50 second, and aperture value to F2.8. The camera-to-surface distance was set between 300 and 500 mm to ensure adequate detail capture. Additionally, during shooting, the camera lens was always kept parallel to the concrete surface to ensure undistorted images.

Based on the above collection plan, systematic image collection was carried out at three fair-faced concrete construction sites in Sanya and Haikou, Hainan Province, yielding a total of 2400 raw images. The images covered four common surface defects of fair-faced concrete: Pores (600 images), cracks (600 images), uneven color (600 images), and human repairs (600 images), as shown in Table 1. The collected images were then processed using Photoshop software, adjusting the resolution to 224×224 pixels for model training, with some typical sample images shown in Figure 5.

Table 1. Statistical summary of surface defects in fair-faced concrete samples

Defect Types of Fair-faced Concrete	Quantity
Pores	600
Crack	600
Surface Contamination	600
Repair	600

(a) Pores



(b) Crack

(c) Surface Contamination

(d) Repair

Figure 5. Surface defect types of fair-faced concrete

Experimental Methods

Experimental setup

The dataset is divided into training and testing sets with an 4:1 ratio, where 80% is used for training the model and 20% for evaluation, as presented in Table 2. To ensure the model's robustness across different defect types, the division strictly adhered

to random sampling principles, ensuring that samples of each defect type were evenly distributed across the training and testing sets.

In addition, the model was implemented using the PyTorch deep learning framework and trained on an Nvidia GeForce RTX 4090 GPU. To validate the effectiveness of the RSAFuser network, the model underwent parameter preprocessing, and pre-trained parameters were used for initialization. Subsequently, training data for surface defects in plain concrete were input into the RSAFuser network model. The Adam optimizer was employed, which adapts the learning rate. The model achieved optimal training performance and highest accuracy with a batch size of 128 and 30 epochs.

Table 2. Division of the plain concrete defect dataset

Defect Types of Fair-faced Concrete	Train set	Test set
Pores	480	120
Crack	480	120
Surface Contamination	480	120
Repair	480	120

Model algorithm building

After dividing the dataset, the images are input into the neural network model in RGB three-channel color format, with the size adjusted to 224×224 pixels. Since the pixel values of the original RGB images range from $I \in (0,255)$, they need to be normalized to meet the computational requirements of the neural network. The specific normalization method is as follows:

$$I' = \frac{I}{255} \tag{13}$$

Here, I denotes the original pixel values of the image, and I' represents the normalized pixel values. After the normalization process $I' \epsilon (0.0,1.0)$, the image values fall within the computational range suitable for the neural network.

The normalized images I' are input into the local expert network f_{Local} and the global expert network f_{Global}, respectively, to extract features across multiple levels. The local features f_{Local} and global features f_{Global} are then sent to the expert voting module for feature fusion f_{cls}, which generates the final classification features. Based on the characteristics of these classification features, the model makes the final classification decision. The algorithm flow is shown below:

Algorithm 1：Residual Self-Attention (RSAFuser) Algorithm
Input: Defect image **I** of Fair-faced concrete Output: Defect classification types **c** 1: The defect image **I** is normalized to the input image $\mathbf{I'=\frac{1}{255}}$ 2: The input image is fed into the local expert network $\mathbf{E_{Local}}$ to obtain the local features $\mathbf{f_{Local}=E_{Local}(I')}$ 3: The input image is fed into the global expert network. $\mathbf{E_{Global}}$ to obtain the global features $f_{Global}=E_{Global}(I')$ 4: The local features $\mathbf{f_{Local}}$ and global features $\mathbf{f_{Global}}$ are sent to the Expert Voting Module (EV-Module) for feature fusion, resulting in the classification features. $f_{cls}=\mu \cdot f_{Local}+v \cdot f_{Global}$ 5: The classified features $\mathbf{f_{cls}}$ are sent to the model's classification head to obtain the final classification result **c**
return **c**

Evaluation metrics establishment

To evaluate the performance of the RSAFuser network in recognizing surface defects in fair-faced concrete, overall and category-specific evaluation metrics were established, using Accuracy and Loss Rate as the performance indicators.

1. Accuracy

Accuracy quantifies the model's overall performance by calculating the ratio of correctly identified defect samples to the total number of samples. The formula is as follows:

$$acc_c = \frac{N_{correct}}{N_{all}} \times 100\% \tag{14}$$

Here, $N_{correct}$ represents the number of correctly identified defect samples, while N_{all} denotes the total number of samples

2. Loss Rate

The Loss Rate quantifies the proportion of samples that the model incorrectly identifies. The formula is:

$$\text{Loss Rate} = \frac{N_{incorrect}}{N_{all}} \quad (15)$$

Here, $N_{incorrect}$ denotes the number of defect samples that were misidentified.

Experimental Results

Experimental results of dataset input

The defect recognition results of the RSAFuser network are shown in Table 3. The results indicate that the RSAFuser network achieved an overall defect recognition accuracy of 98.55%. For each type of defect, the accuracy remained above 95%, with the recognition rate for crack defects reaching 100%.

Table 3. Defect recognition accuracy of RSAFuser network.

Defect Type	Pores	Crack	Surface Contamination	Repair	Overall
RSAFuser18	97.48%	100.00%	100.00%	97.44%	98.84%
RSAFuser34	99.16%	100.00%	100.00%	97.44%	99.42%
RSAFuser50	97.48%	100.00%	100.00%	100.00%	99.13%
RSAFuser101	100.00%	100.00%	100.00%	97.44%	99.71%

Comparison of model recognition accuracy

To highlight the advantages of local-global feature fusion in defect recognition, this study compares the results of the RSAFuser expert network with the local recognition expert ResNet and the global recognition expert self-attention (Swin-Transformer) networks (using ResNet18, Swin-Transformer, and RSAFuser18 as examples), as shown in Table 4 and Figure 6. The data in the table show that, compared to the individual local recognition expert ResNet and the global recognition expert Swin-Transformer that focuses on long-range dependencies, RSAFuser achieves a significant improvement, increasing accuracy was improved by 19.19% and 21.51%, respectively., respectively. This demonstrates that extracting both local and global information from defect images and then fusing the two types of information through expert modules effectively compensates for the shortcomings of individual local or global perception, thereby significantly improving defect recognition efficiency and performance.

Table 4. Comparison of defect recognition accuracy across different expert networks

Expert Network	ResNet18	Swin-Transformer	RSAFuser18
Recognition Accuracy	79.65%	77.33%	98.84%

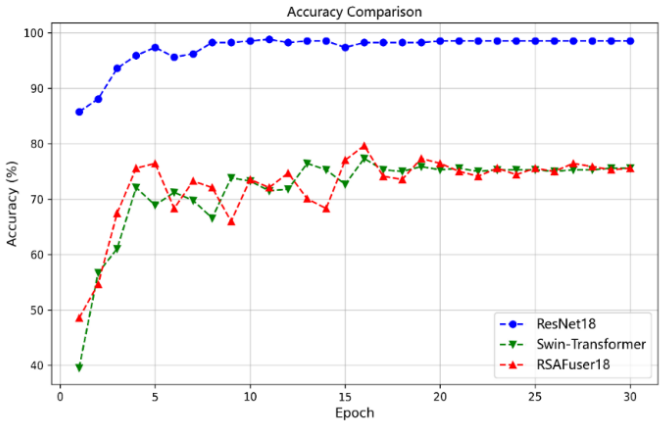
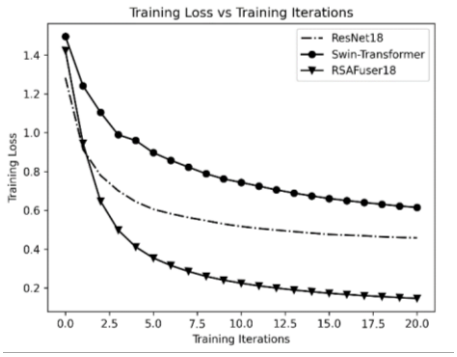


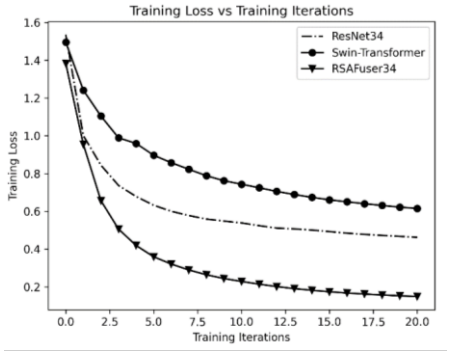
Figure 6. Comparison of training accuracy across models

Comparison of model loss decrease

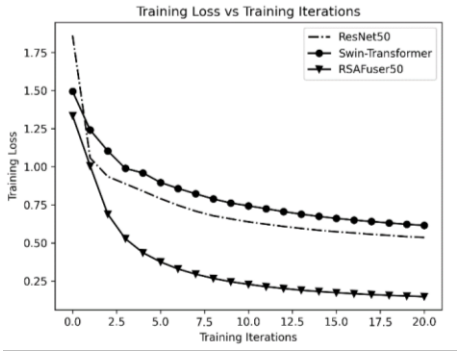
This study also specifically plotted the loss decrease curves for the three experts during the training process, as shown in Figure 7. It is evident that the loss decrease of RSAFuser is more significant compared to the local expert and the global expert. For example, with ResNet18 and RSAFuser18, after 2.5 epochs of iteration, the loss of RSAFuser decreased by about 1.1, while the local expert ResNet only decreased by 0.5, and the global expert Swin-Transformer decreased by 0.45. A similar convergence efficiency can also be observed in RSAFuser34, RSAFuser50, and RSAFuser101, which indicates that the fusion of local and global information in RSAFuser combines both effectiveness and efficiency, achieving the optimal recognition efficiency.



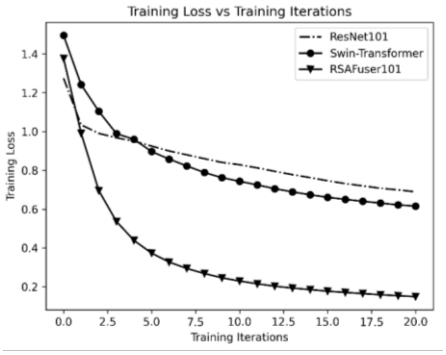
(a) Loss Curves of ResNet18, SwinT, and RSAFuser18



(b) Loss Curves of ResNet34, SwinT, and RSAFuser34.



(c) Loss Curves of ResNet50, SwinT, and RSAFuser50.



(d) Loss Curves of ResNet101, SwinT, and RSAFuser101.

Figure 7. Loss decrease curves of three experts

CONCLUSION

This study proposes the RSAFuser network, an expert identification network based on the fusion of local and global information, for the classification and identification of surface defects in fair-faced concrete. The network employs a residual block for local information and a multi-head self-attention module for global information, effectively ignoring irrelevant image features and focusing on key features, thereby balancing both local and global information. After extracting relevant information, the network fuses local and global data. Experimental results show that, compared to single-classification recognition networks, the RSAFuser network achieves faster convergence and higher accuracy, with classification performance improving by 19.19% and 21.51% over ResNet and Swin-Transformer, respectively, successfully classifying defects in fair-faced concrete. Future research will focus on developing a standard parameter system for defects to enable the automated evaluation of surface defects in fair-faced concrete.

ACKNOWLEDGEMENTS

We sincerely thank the School of Civil and Architecture Engineering, Hainan University, for their academic support and provision of resources, as well as to Hainan Haikong Zhongneng Construction Engineering Co., Ltd. for their technical assistance, without which this research would not have been possible.

FUNDING

2024 Natural Science Foundation of Hainan Province (No. 524MS033).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

All authors were fully informed of the study details and made fully informed decisions regarding their participation in this research.

REFERENCE

[1] Yu, J. Tadao Ando's Application of Fair-Faced Concrete Building Materials. ACEGE, 75:126-131.

[2] Zhang, S., Feng, C., Fan, Y., et al. Jingang Cultural Center: Complex-Shaped Fair-Faced Concrete Structure. Structural Engineering International, 2022, 33(3):425-430.

[3] Strehlein, D., Schießl, P. Dark discoloration of fair-face concrete surfaces—transport and crystallization in hardening concrete. Journal of Advanced Concrete Technology, 2008, 6(3):409-418.

[4] Fu, X., Li, Y. Strength, durability and appearance of low-carbon fair-faced concrete containing multiple mineral admixtures. Constr. Build. Mater. 2023, 392:131838.

[5] Jiang, Q., Yu, C., Zhang, Q. Progress in the formation and regulation of surface pore defects in fair-face concrete. Materials Review, 2025, 39(4):23110170.

[6] Zhang, J., Xu, Y., Lu, D. Intelligent Detection of Concrete Apparent Defects Based on a Deep Learning-Literature Review, Knowledge Gaps and Future Developments. Ceramics-Silikáty, 2024, 68(2):252-266.

[7] Tapeh AT, G., Naser, M.Z. Artificial intelligence, machine learning, and deep learning in structural engineering: A scientometrics review of trends and best practices. Arch. Comput. Methods Eng, 2023, 30:115-159.

[8] Deng, J., Singh, A., Zhou, Y., et al. Review on computer vision-based crack detection and quantification methodologies for civil structures. Constr. Build. Mater, 2022, 356:129238.

[9] Kim, H., Ahn, E., Shin, M., et al. Crack and noncrack classification from concrete surface images using machine learning. Structural Health Monitoring, 2019, 18(3):725-738. https://doi.org/10.1177/1475921718768747.

[10] Fan, C.L. Detection of multidamage to reinforced concrete using support vector machine-based clustering from digital images. Structural Control and Health Monitoring, 2021, 28(12): e2841. https://doi.org/10.1002/stc.2841.

[11] Spencer Jr, B.F., Hoskere, V., Narazaki, Y. Advances in computer vision-based civil infrastructure inspection and monitoring. Engineering, 2019, 5(2), 199-222. https://doi.org/10.1016/j.eng.2019.01.001.

[12] Zhang, J., Xu, Y., Lu, D. Intelligent detection of concrete apparent defects based on deep learning—literature review, knowledge gaps, and future developments. Ceramics-Silicates, 2024, 68(2):252-266.

[13] Zhu, J., Song, J. An intelligent classification model for surface defects on cement concrete bridges. Applied sciences, 2020, 10(3):972.

[14] Simonyan K. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[15] Li, Y., Bao, T., Xu, B., et al. A deep residual neural network framework with transfer learning for concrete dams patch-level crack classification and weakly-supervised localization. Measurement, 2022, 188: 110641.

[16] Liu, R.S., Li, Z.F., Feng, Q.H., et al. Concrete surface crack defect recognition based on ResNet50 fine-tuned network model. Journal of Henan Institute of Technology, 2023, 31(4):27-31.

[17] He, K., Zhang, X., Ren, S., et al. Identity mappings in deep residual networks//Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV 14. Springer International Publishing, 2016:630-645.

[18] Dosovitskiy, A. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv Preprint, 2020, arXiv:2010.11929.

[19] Zhou, D., Kang, B., Jin, X., et al. DeepViT: Towards Deeper Vision Transformer, 2021. DOI:10.48550/arXiv.2103.11886.

[20] Liu, Z., Lin, Y., Cao, Y., et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, 10012-10022.

[21] Vaswani, A., Shazeer, N., Parmar, N., et al. Attention Is All You Need.arXiv, 2017. DOI:10.48550/arXiv.1706.03762.