IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 March 2024, accepted 2 April 2024, date of publication 4 April 2024, date of current version 15 April 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3385090*

APPLIED RESEARCH

# Research on Non-Destructive Testing of Curtain Wall Hangings Based on Efficient Channel Attention Vision Transformer Network

**CHONGLIANG GAO, JIE GAO, AND YAJUN CAO**
China Construction Shenzhen Decoration Co., Ltd., Guangdong, Shenzhen 518001, China

Corresponding author: Chongliang Gao (qyjszxcscec@163.com)

This work was supported by the Science and Technology Research and Development Program of CSCEC under Grant CSCEC-2022-Z-27.

**ABSTRACT** Traditional inspection of curtain wall metal hangings usually relies on manual visual inspection, which is costly, slow, and limited in coverage. To reduce cost and improve efficiency and accuracy, a non-destructive automatic inspection system for architectural curtain walls based on millimeter wave imaging was designed. The system is designed as a single-side reflective point-frequency imaging, which can effectively solve the problem of reflective waves generated on the upper and lower surfaces of the curtain wall during millimeter-wave penetration affecting the echo signal. To address the fine-grained classification problem of metal hangings, we propose an efficient channel attention Vision Transformer (ECA-ViT) lightweight classification network based on a hybrid architecture of convolutional neural network and Transformer. Among them, the inverted residual attention module (IRAM) improves the network's attention weight on the image foreground, and the low-rank MobileViT module (LR-ViT) can provide global modeling for the network and making the whole model more lightweight by reducing the computational complexity of the self-attention mechanism. The experimental results demonstrate that the proposed method achieves an accuracy of 95.66% with fewer model parameters and computational complexity, demonstrating good performance advantages.

**INDEX TERMS** Curtain wall metal pendant, deep learning, ECA-ViT, IRAM.

## I. INTRODUCTION

Curtain wall metal hangings are common components in modern architectural design, which play an important role in building appearance decoration and structural support. In recent years, with the wide application of curtain walls [1], the maintenance of curtain walls has become more and more important. Lin et al. [2] proposed an infrared thermography method based on scanning laser depth heating, which can effectively detect the defective problems of structural adhesive in glass curtain walls; Xu et al. [3] proposed a deformation monitoring method for glass curtain walls based on fiber optic sensing technology. Traditional metal pendant detection methods are mainly based on visual inspection and external equipment detection technology, but these methods have many shortcomings. Visual inspection is affected by the environmental light, and there are problems of leakage and misdetection. The external equipment detection technology requires specialized equipment and technicians, which is costly and complicated to operate. Therefore, it is imperative to develop a convenient, efficient and accurate metal pendant detection technology.

Millimeter-wave imaging technology, as an emerging nondestructive testing method, has many advantages. Among them, millimeter-wave has the property of penetrating non-metallic materials [4], [5], such as stone, which can be detected without destroying the curtain wall structure, and the non-diffraction millimeter beam can improve the inspection depth [6]. In addition, millimeter-wave imaging technology has high resolution and high sensitivity, which can detect tiny defects on the surface of metal hangings [7]. Combined with computer vision technology, automated, efficient, and highly accurate nondestructive inspection can be achieved by analyzing and recognizing millimeter-wave images [8], [9].

The associate editor coordinating the review of this manuscript and approving it for publication was Wanchen Yang.

Wu and Dahnoun [10] presented a millimeter-wave radar-based health monitoring system that is capable of posture estimation and heart rate detection. With millimeter-wave radar, this study demonstrated the potential of non-contact health monitoring, providing a convenient and interference-free means of monitoring elderly people or patients who require continuous monitoring. Wagner et al. [11] explored how non-contact millimeter-wave radar can be used to detect intrusive drilling in secure transportation containers. This technology is expected to be widely used in the security field to ensure the integrity of transportation containers and reduce potential threats. In addition, millimeter wave technology has important applications in medical imaging. Iliopoulos et al. [12] used field-focusing techniques to enhance millimeter wave penetration in breast tissue to improve the accuracy of breast cancer detection. This study provides a new tool for early cancer detection that can improve patient survival. Yang et al. [13] introduced the Transformer technique for anchorless detection of passive millimeter wave images. Their work is potentially valuable for occluded object detection.

As an important part of the building curtain wall system, the quality of the curtain wall metal pendant directly affects the service life and safety of the building. In recent years, accidents caused by curtain wall detachment have occurred frequently, especially in high-rise buildings where curtain wall detachment causes significant damage to ground personnel and other facilities. Therefore, how to detect the components inside the curtain has become a hot research topic. Millimeter-wave imaging technology uses microwave radiation to capture signals reflected from target objects to produce high-resolution images. These images provide information about the metal hangings of the stone curtain wall, including their shape, location and structural characteristics. By combining computer vision and deep learning techniques [14], [15], [16], [17], the ability to analyze and interpret millimeter-wave images of stone curtain wall metal pendants can be further improved, thereby automating the pendant classification process and increasing efficiency and accuracy. However, the millimeter-wave images of a wide variety of metal pendants are extremely similar, which poses a great challenge to the judgment of pendant types. Classification is a critical step in processing millimeter-wave images of curtain wall metal hangings, which can help identify the condition and problems of the hangings, thus ensuring the safety and maintainability of the building structure.

In this paper, we study the classification of curtain wall metal hangings based on millimeter wave imaging data of curtain wall metal hangings. We designed a unilateral reflective millimeter-wave point-frequency imaging device, which has a simple structure and can effectively solve the problem of reflective waves generated on the upper and lower surfaces of the curtain wall during millimeter-wave penetration affecting the echo signal. It greatly improves the accuracy of the echo signal and improves the imaging quality. Aiming at the problem of similar imaging of metal hangings, a method of fine-grained image classification based on efficient channel attention Vision Transformer (ECA-ViT) is proposed. The method can effectively improve the classification accuracy and the whole neural network more lightweight. Deep learning algorithms and millimeter wave imaging technology are combined to complete the visual inspection of curtain wall metal components without damaging the curtain wall.

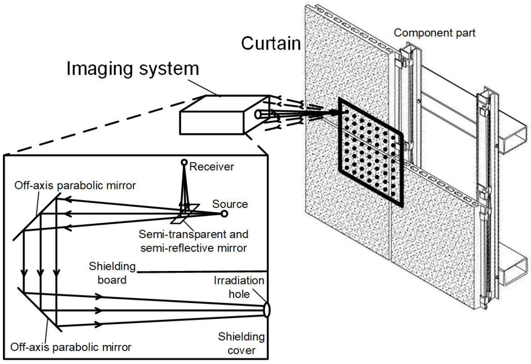

**FIGURE 1. Single side reflection millimeter wave point frequency imaging device.**

## II. IMAGING DEVICE

The single side reflective millimeter-wave wave point frequency imaging system can be loaded on the high-altitude wall climbing robot, and adopts the point-to-point scanning method with strong penetration ability to irradiate and image each point on the curtain wall, to realize the detection of the metal hangings of the curtain wall. The structure of the whole device is shown in Figure 1, the resolution of the device can reach 3mm, and the aperture size of the mm-wave radiation source is 10cm. The imaging device consists of a 35Ghz transmitter, receiver, semi-transparent semi-reflective mirrors, and two off-axis parabolic mirrors with an off-axis angle of 90 °. The sources and detectors we use are both linearly polarized. The polarization direction must be consistent, and the two also need to align the polarization direction. The device changes the emission path of the transmitter by two off-axis parabolic mirrors and focuses on the detected object. When the millimeter-wave penetrates the curtain wall, reflection and transmission will occur. Transmission makes the light path move a certain distance laterally. At this time, the incident and exit angles are the same. After penetration, the wave is focused on the element and reflected. Due to the multi-angle reflection of the signal, the whole system only retains one illumination hole, so that the echo signal with the same path as the transmitted signal passes through the hole and returns. Other reflected and transmitted waves are shielded to ensure the accuracy of the echo signal, which can effectively solve the problem that the reflected wave generated on the upper and lower surfaces of the curtain wall affects the echo signal during the millimeter-wave penetration process. A semi-transparent semi-reflective mirror is placed at the transmitting port of the millimeter-wave transmitter to reflect part of the echo signal upward by 90°, which is received by the receiver and used to detect the hangings. The structure of the whole device is simple, and by changing the power of the transmitter, it can penetrate different thicknesses of curtain walls for imaging metal hangings.

Due to the fact that in actual engineering operations, the curtain wall is perpendicular to the ground and the metal pendant is perpendicular to the curtain wall, the imaging results of the SE type, T type, back bolt, butterfly type and oblique cantilever hanging metal pendants are shown in Figure 2. The millimeter wave images of these five metal pendants were generated by this imaging device, and data enhancement was performed on the images to obtain 2940 images of the SE type pendant, 2730 images of the T type pendant, 2730 images of the back bolt pendant, 2870 images of the butterfly type pendant, and 2760 images of the oblique cantilever pendant, of which 70% were used for training and 30% were used for testing.

## III. METHODS

The millimeter-wave images of curtain wall metal pendants are mainly grayscale images, and the color characteristics are not obvious. In practical engineering, metal pendants are fixed on the keel to undertake the curtain wall, and both the keel and the pendants belong to steel structures. It is not easy to distinguish the two in imaging, and the similarity of various metal pendants in imaging is very high. Hence, there are limited distinguishing characteristics in the millimeter wave images of different types of metal hangings on the curtain wall, and a neural network is needed to capture the fine details of the images of different types of metal hangings.

### A. NETWORK FRAMEWORK

In this paper, a fine-grained classification method based on the ECA-ViT model is proposed. The network model adopts a hybrid architecture of convolutional neural network and Transformer. The efficient convergence and local feature ability of CNN are combined with the global feature correlation ability of Transformer to achieve accurate recognition of millimeter-wave images of metal hangings. The complete end-to-end architecture of the model is shown in Figure 3. It consists of a linear mapping layer, multiple inverted residual attention modules (IRAM), multiple low-rank MobileViT modules (LR-ViT), and a final classifier. The proposed IRAM is to modify the small convolutional kernel in the inverted residual module into a large convolutional kernel and add the channel attention mechanism to improve the sensory field of the network and obtain a larger range of local features. The LR-ViT module mainly reduces the amount of calculation of the self-attention mechanism in the Transformer module, and lightens the Transformer module so that the overall network can reduce hardware requirements.

### B. INVERSE RESIDUAL ATTENTION MODULE (LRAM)

Considering that the millimeter-wave images of metal components have relatively few identifiable features, and the local differences in the images are not large, a deep network with stronger expression ability is needed to fit. In theory, the greater the network depth, the higher the degree of fitting to the training set. However, in actual training, too deep network is difficult to train, which often leads to greater errors. The residual module realizes direct connection by skipping a certain number of layers, which can effectively solve the problems of gradient disappearance and gradient explosion, and the performance of the neural network can be improved at a deeper level.

The inverted residual attention module (LRAM) is considered to be proposed for millimeter-wave images. Compared with the conventional inverted residual module in Fig. 4(a), this module combines a large-scale convolutional layer and an efficient channel attention mechanism to enhance the effective receptive field of feature extraction and the acuity of details. The large-scale convolutional layer takes reference from the Swin Transformer [18] model of using a large 7 × 7 window for the feature map slicing method, replacing the original 3 × 3 convolutional kernel with a 7 × 7 convolutional kernel, and thus enhancing the sensory field of the network. Since the use of large-scale convolution kernels will increase the number of parameters and the amount of calculation, the amount of calculation is reduced by reducing the activation function layer and the normalization layer. The proposed IRAM only retains one normalization layer and one activation function layer. The structure is shown in Figure 4(b). After the 7 × 7 convolution layer, the efficient channel attention module is connected. The module implements a local cross-channel interaction strategy without dimensionality reduction through one-dimensional convolution, which can improve the classification accuracy of fine-grained images without increasing the amount of calculation.

There are two kinds of inverted residual attention modules in the network, namely inverted residual attention module I (IRAM-I) and inverted residual attention module II (IRAM-II). The difference between the two is that the intermediate large-scale deep convolution layer uses different strides. The IRAM-I uses a stride of 1, and there is a shortcut connection. The stride distance used in the IRAM-I is 2, that is, it needs to be down-sampled, and there is no shortcut connection.

### C. LOW-RANK MOBILEVIT MODULE (LR-ViT)

The low-rank MobileViT (LR-ViT) module is the core part of the ECA-ViT network. The input feature map is locally modeled and adjusted the number of channels through a 3 × 3 convolution layer and a 1 × 1 convolution layer. The obtained feature map is converted into the input data dimensions required by the Transformer module, and the specific operation is shown in Figure 5. Firstly, the feature map is divided into patches, and the size of the patch in Figure 5 is 2 × 2, that is, each patch consists of 4 pixels.

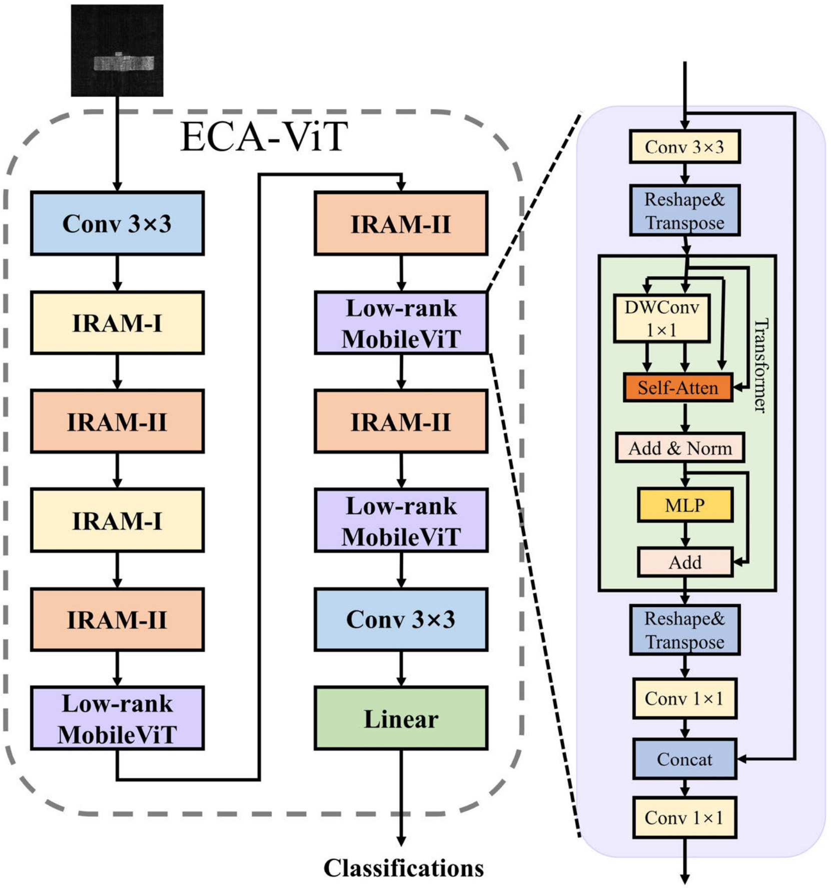

**FIGURE 2.** Network structure diagram.

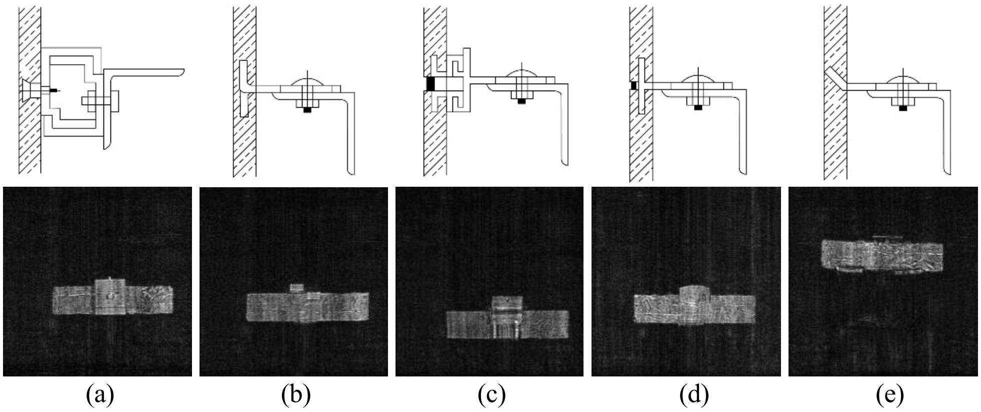

**FIGURE 3.** Imaging image of curtain wall metal structure. (a) Back bolt; (b) Butterfly type; (c) SE type; (d) T type; (e) Oblique cantilever.

The pixel of the same color is flattened in a sequence, so that self-attention can be used directly to compute the attention of each sequence in parallel. Then, the global feature modeling is performed through the Transformer module. Finally, the output feature map and the original input feature map are spliced by a method similar to the residual structure. The main amount of computation in the Transformer module comes from the self-attention layer. The LR-ViT module we proposed mainly reduces the computation of the self-attention mechanism by reducing the dimensions of the key and value matrices in the self-attention layer.

In the original self-attention mechanism, the input feature sequence is trans-formed to obtain the query matrix $\boldsymbol{Q}$, the key matrix $\boldsymbol{K}$ and the numerical matrix $\boldsymbol{V}$. Each matrix has

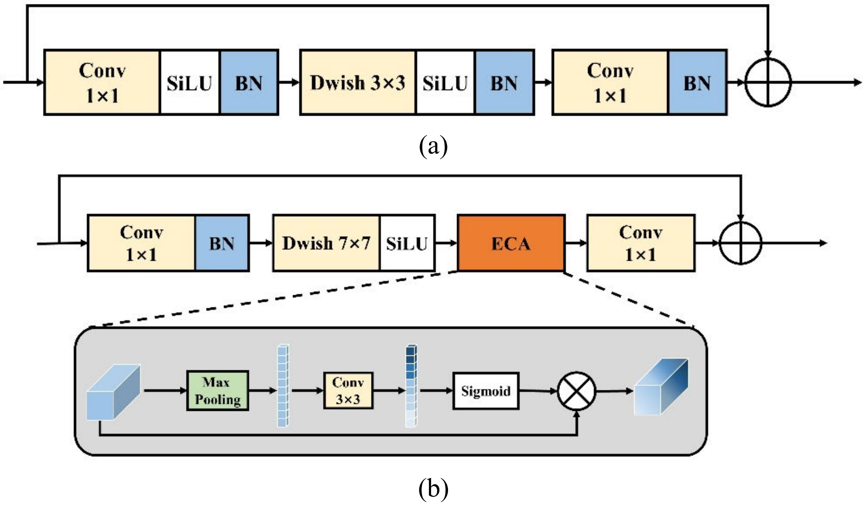

**FIGURE 4.** Inverted residual block. (a) Original inverted residual block; (b) Large-scale Inverted residual block.

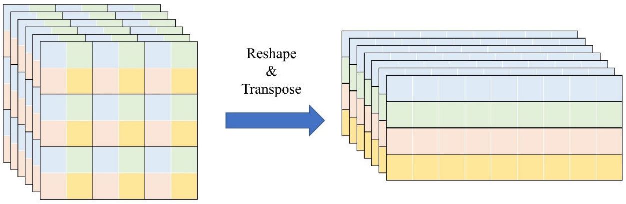

**FIGURE 5.** Feature map dimension transformation.

the same dimension N × C, where N = H/2×W/2. The calculation formula of the self-attention mechanism is as follows:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\frac{(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})}{\sqrt{d}})\boldsymbol{V} \tag{1}$$

According to the similarity or correlation between $\boldsymbol{Q}$ and $\boldsymbol{K}$, the original score is normalized, and the value is weighted and summed according to the weight coefficient to obtain attention value. Its computational complexity increases exponentially with the increase of the resolution of the input image. Therefore, the dimensions of the $\boldsymbol{K}$ and $\boldsymbol{V}$ matrices are reduced by scaling the coefficient S. Firstly, the dimensions of $\boldsymbol{K}$ and $\boldsymbol{V}$ matrices are transformed into H × W × C, and the deep separable convolution is used to reduce the dimension of the matrix, taking $\boldsymbol{K}$ matrix as an example.

$$\boldsymbol{K'} = DWConv(\boldsymbol{K}) \tag{2}$$

The convolution kernel size of the convolution layer is S×S, and the step size is S. The feature map is divided into windows. These windows do not overlap with each other, and the dimension of each window is S. A total of H/S×W/S windows are obtained, and the dimension of the obtained $\boldsymbol{K}'$ matrix is H/S × W/S × C, subsequently transformed into N/S$^2$ × C. The V matrix goes through the same procedure to obtain $\boldsymbol{V}'$. The formula for the new attention mechanism obtained is:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\frac{(\boldsymbol{Q}(DWConv(\boldsymbol{K}))^{\mathrm{T}})}{\sqrt{d}})DWConv(\boldsymbol{V}) \tag{3}$$

Therefore, the self-attention mechanism complexity is reduced from $\boldsymbol{O}(\mathrm{N}^2)$ to $\boldsymbol{O}(\mathrm{N}^2/\mathrm{S}^2)$. The LR-ViT can improve the classification accuracy while reducing the computational complexity of the whole model.

### D. LOSS FUNCTION

The most commonly used loss function in image classification is the cross-entropy loss function, which is used to measure the difference between the output of the model and the real label. The expression of the loss function is:

$$Loss_{CE} = -\frac{1}{L}\sum_{l\in L}\sum_{i=1}^{N} y_l^i \log(\hat{y}_l^i) \tag{4}$$

where L denotes the number of samples, N denotes the number of categories, and $y_l^i$ denotes the category corresponding to the true label of the lth sample, taken as 1 or 0. $y''$ denotes the probability value predicted by the model. Minimizing the cross-entropy loss function is equivalent to maximizing the model's probability of predicting the correct category. During the training process, the model parameters are adjusted by the optimization algorithm to reduce the value of the loss function so that the model predicts the true labels more accurately.

### E. IMPLEMENTATION DETAIL

In this paper, the proposed model is implemented in the Pytorch framework and the network is trained end-to-end using the Adam optimizer. The initial learning rate of the model is 1 × 10$^{-4}$, the minimum learning rate is 1 × 10$^{-6}$, and the learning rate is adjusted by cosine annealing. The model was trained on NVIDIA GeForce RTX 3090 GPUs for 100 iterations with a batch size of 32.

## IV. RESULTS

### A. EVALUATION METRICS

Test images are used to test the weights obtained from the network training. The performance parameters of the test network are mainly precision, recall rate and F1.

$$precision = \frac{TP}{TP + FP} \tag{5}$$

where Precision is the precision rate, which is the ratio of correctly retrieved targets to all actually retrieved targets, TP is the number of positive classes predicted to be positive, and FP is the number of negative classes predicted to be positive.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

where Recall is the recall rate, which is the proportion of correctly retrieved targets to all targets that should have been retrieved, FN is the number of positive classes predicted to be negative.

$$Accuracy = \frac{TP + TN}{M} \tag{7}$$

where Accuracy is the accuracy rate, which is the proportion of correctly retrieved targets to all targets, TN predicts negative categorization as negative categorization, and S is the total number of samples.

$$F1 = 2^{*}\frac{Precision^{*}Recall}{Precision + Recall} \tag{8}$$

F1 is the harmonic mean of Precision and Recall.

### B. RESULT ANALYSIS

In the experiment, we compare the proposed model with the previous state-of-the-art image classification model. The model is analyzed from the classification accuracy and computational complexity, and the superiority of our proposed model is demonstrated. Through ablation experiments, we illustrate the effectiveness of the inverted residual attention module (IRAM) and the low-rank MobileViT module (LR-ViT) for model classification.

We performed comparative experiments using Mobilenetv2, Efficientnet-b0, and MobileViT network models on the collected dataset. Figure 6 illustrates the variation in loss values for each network. As depicted in Figure 6, our proposed efficient channel attention vision Transformer (ECA-ViT) network model demonstrated superior convergence and the lowest loss value during the process of model training.

To compare objectively the classification performance of various networks, this paper statistics their performance metrics on the self-constructed dataset, as shown in Table 1. It is evident from Table 1 that our proposed network model enhances the F1 value by 2.54% - 3.87% and the accuracy by 2.54% - 3.88% when compared to other models, demonstrating the model's capability of detecting fine-grained image distinctions and its strong feature extraction ability. The respective amounts of parameters and computational operations in the network are compared and analyzed. An evaluation of the computational cost is then carried out by quantifying the number of parameters and floating-point operations (FLOPs) used. The parameter quantity denotes the number of parameters requiring learning during the process of training, while FLOPs represents the number of floating-point operations executed during the inference stage. The models in this study underwent evaluation by single-scale inference using an input image resolution of 224 × 224. It can be seen from Table 1 that our model parameters and calculations are lower than the comparison model, which can be better deployed on the curtain wall metal hanging detection system.

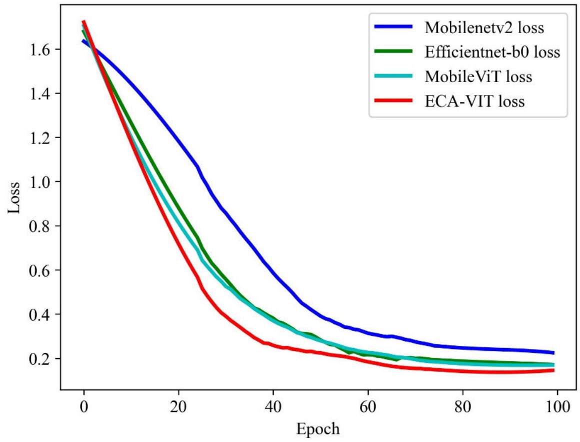

**FIGURE 6.** Loss value variation chart.

**TABLE 1.** Comparison of the performance of each network for image classification on self-built datasets.

| Network model | F1 | Accuracy | parameters | FLOPs |
|---|---|---|---|---|
| Mobilenetv2 | 91.81% | 91.78% | 3.51 M | 654.97 M |
| Efficientnet-b0 | 92.70% | 92.64% | 5.29 M | 825.66 M |
| MobileViT | 93.14% | 93.12% | 1.27 M | 547.34 M |
| ECA-ViT | 95.68% | 95.66% | 1.09 M | 470.56 M |

To validate the effectiveness of the proposed inverted residual attention module and the low-rank MobileViT module, we conducted ablation experiments on them separately, all of which were carried out on self-constructed datasets. Specifically, we use the MobileViT model as a baseline for comparison. We compare Baseline-I, which contains only the inverted residual attention module, to the baseline model. As can be seen in the second row of Table 2, the inverted residual attention module improves classification accuracy, and the reduced activation and normalization layers significantly reduce model parameters. Similarly, we compare Baseline-II, which contains only the low-rank MobileViT module, to the baseline model. As can be seen in the third row of Table 2, the low-rank MobileViT module improves model

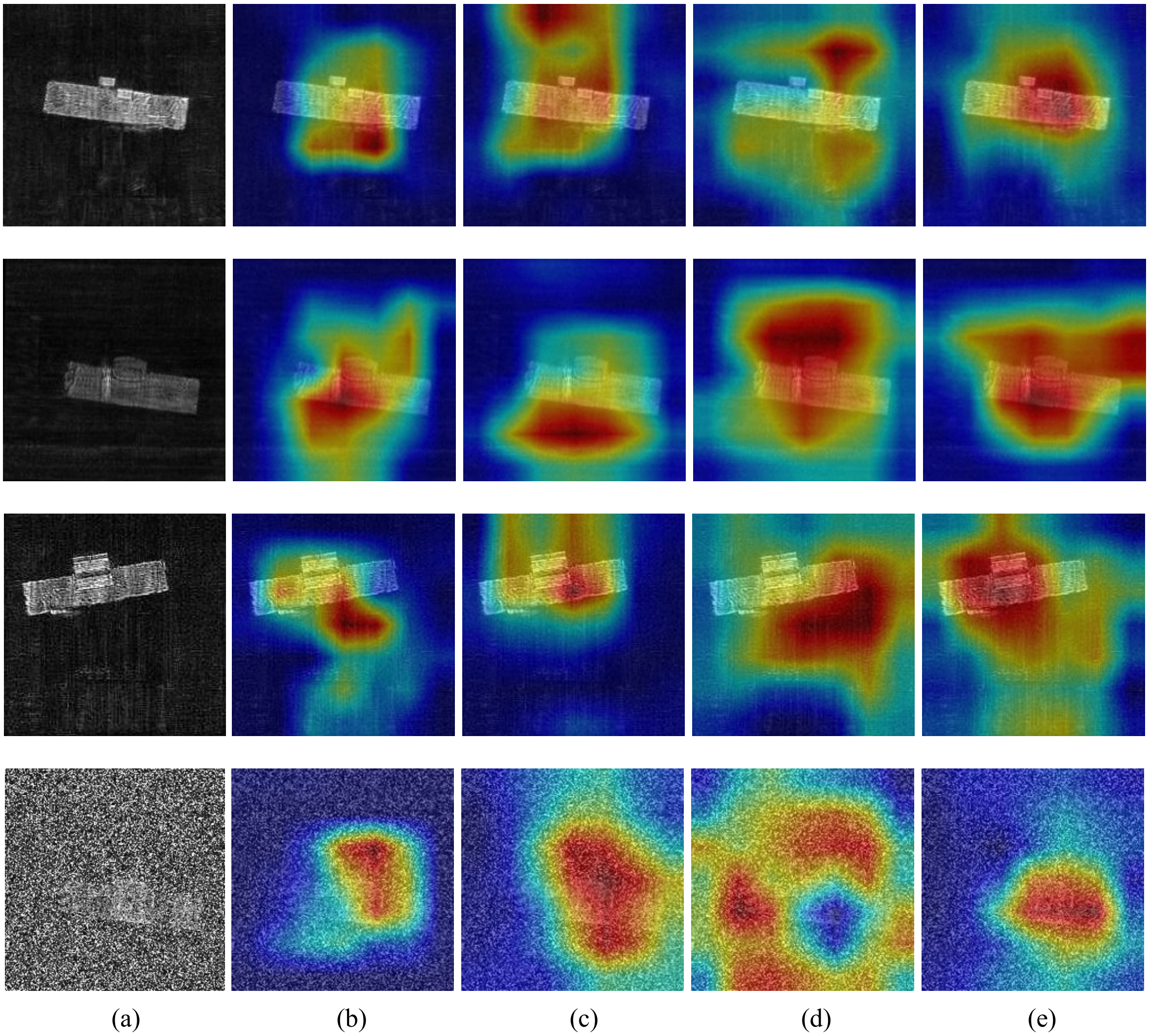

**FIGURE 7.** Visualization test results for four examples. (a) Test image; (b) The results of Mobilenetv2; (c) The results of Efficientnet-b0; (d) The results of MobileViT; (e) The results of ECA-ViT.

accuracy while reducing computational effort. Therefore, both the inverted residual attention module and the low-rank MobileViT module are effective in improving the model performance and can make the model more lightweight to some extent.

**TABLE 2.** Comparison of ablation of models.

| Network model | F1 | Accuracy | parameters | FLOPs |
|---|---|---|---|---|
| Baseline | 93.14% | 93.12% | 1.27 M | 547.34 M |
| Baseline-I | 94.33% | 94.27% | 1.08 M | 519.21 M |
| Baseline-II | 94.12% | 94.07% | 1.28 M | 498.69 M |
| ECA-ViT | 95.68% | 95.66% | 1.09 M | 470.56 M |

The ablation experiment was also performed on the selection of the size of the middle layer convolution kernel in the inverted residual attention module. We replace the inverted residual module in MobileViT with the inverted residual attention module as the baseline model. We use several different sizes, which are 3 × 3,5 × 5,7 × 7,9 × 9, respectively. The experimental results are shown in Table 3. As the convolution kernel increases, the number of parameters and calculations will also increase, and the accuracy rate reaches saturation when using a 7 × 7 convolution kernel. It shows that the large-scale convolution layer can effectively improve the receptive field of the model, to capture more subtle differences in the image and improve the accuracy of the model. However, the cost is that more computation is required.

**TABLE 3.** Comparison of ablation with different convolution kernel sizes.

| Kernel size | F1 | Accuracy | parameters | FLOPs |
|---|---|---|---|---|
| 3×3 | 93.11% | 93.09% | 1.06 M | 449.97 M |
| 5×5 | 93.72% | 93.68% | 1.07 M | 477.67 M |
| 7×7 | 94.33% | 94.27% | 1.08 M | 519.21 M |
| 9×9 | 92.85% | 92.81% | 1.10 M | 574.10 M |

To demonstrate the interpretability of the proposed models, we visualize the last feature layer of all models. The attention of the neural network for the whole image is determined by generating a heat map of this feature layer. As shown in Figure 7, four images are selected for heat map visualization, with the first column being the tested image and the rest of the columns corresponding to the results of different detection models. As seen in Figure 7, the models present darker marker colors if they have stronger attention weights for the information in the image. On the contrary, a weaker weight of attention for the information in the image presents a lighter marker color. However, the metal pendant regions in the images are all small and the features are easily ignored. In the first three rows of Figure 7, the Mobilenetv2, Efficientnet-b0, and MobileViT models do not focus on the metal pendant region in the image very accurately, and some of them deviate from the region of interest. The ECA-VIT is much more sensitive to capture the details in the image, and the attention weights of the image are more concentrated on the imaging region of the metal pendant. In addition, the results in the last row give a good indication of the robustness of the proposed method. The image used for detection is spiked with Gaussian white noise, and there is not much difference between the foreground and background information in the whole image, and ECA-VIT can still accurately focus on the corresponding foreground information. The results show that the proposed method is able to expand the receptive field and thus effectively aggregate the contextual information of the features.

## V. DISCUSSION

Traditional detection methods for curtain wall metal pendants are mainly based on manual visual inspection and external equipment detection techniques, which have some limitations that are particularly evident in large building projects. These limitations include inefficiencies in inspection, significant time and human resources, limited field of view, and inadequate coverage of hard-to-reach areas, which can lead to project delays and increased costs.

Millimeter-wave non-destructive testing technology for metal pendants offers a more advanced and effective alternative. The technology is based on millimeter-wave radiation, which is capable of penetrating non-metallic materials, thus enabling high-quality inspection of metal pendants. Millimeter-wave technology is characterized by high accuracy, wide spectral coverage, the ability to monitor in real-time and automatic data recording, which have significant potential to improve construction quality, reduce safety risks and reduce labor costs.

Further, the combination of millimeter-wave technology with a specially designed wall-climbing robot provides an innovative solution for the inspection and maintenance of metal hangings. This integration enables efficient and highly accurate inspection while reducing the risk of working at height, ensuring the efficiency and quality of construction projects. The promotion and development of this technology application is of significant value to construction projects.

## VI. CONCLUSION

In this paper, a lightweight model of efficient channel attention Vision Transformer (ECA-ViT) is proposed for classification of millimeter-wave images of curtain wall metal hangings, which is able to improve the accuracy of classification of fine-grained images. The network is a hybrid architecture model of CNN and Transformer. An inverse residual attention module (IRAM) and a low-rank MobileViT module (LR-ViT) are proposed. The IRAM is beneficial to improve the receptive field of the model and improve the ability of the model to capture details. The LR-ViT reduces the computation of the self-attention layer in the Transformer module and improves the classification accuracy to a certain extent. The ablation experiments show that the proposed IRAM and LR-ViT module can effectively improve the accuracy.

## REFERENCES

[1] Z. Huang, M. Xie, H. Song, and Y. Du, "Modal analysis related safety-state evaluation of hidden frame supported glass curtain wall," *J. Building Eng.*, vol. 20, pp. 671–678, Nov. 2018.

[2] J. Lin, X. Hong, Z. Ren, and J. Chen, "Scanning laser in-depth heating infrared thermography for deep debonding of glass curtain walls structural adhesive," *Measurement*, vol. 192, Mar. 2022, Art. no. 110902.

[3] D. Xu, Y. Wang, and J. Xie, "Monitoring and Analysis of building curtain wall deformation based on optical fiber sensing technology," *Iranian J. Sci. Technol., Trans. Civil Eng.*, vol. 46, pp. 3081–3091, Sep. 2022.

[4] C. D. Haworth, Y. R. Petillot, and E. Trucco, "Image processing techniques for metallic object detection with millimetre-wave images," *Pattern Recognit. Lett.*, vol. 27, no. 15, pp. 1843–1851, Nov. 2006.

[5] K. Brinker, M. Dvorsky, M. T. Al Qaseer, and R. Zoughi, "Review of advances in microwave and millimetre-wave NDT&E: Principles and applications," *Philos. Trans. Roy. Soc. A*, vol. 378, no. 2182, 2020, Art. no. 20190585.

[6] D. Zhang, J. Liu, J. Yao, Z. Zhang, B. Chen, Z. Lin, J. Cao, and X. Wang, "Enhanced sub-terahertz microscopy based on broadband airy beam," *Adv. Mater. Technol.*, vol. 7, no. 5, May 2022, Art. no. 2100985.

[7] D. Meier, C. Zech, B. Baumann, B. Gashi, M. Malzacher, M. Schlechtweg, J. Kühn, M. Rösch, and L. M. Reindl, "Millimeter-wave radar sensor for automated tomographic imaging of composite materials in a manufacturing environment," *IEEE Sensors Lett.*, vol. 5, no. 3, Mar. 2021, Art. no. 3500104.

[8] N. Wingren and D. Sjöberg, "Nondestructive testing using mm-wave sparse imaging verified for singly curved composite panels," *IEEE Trans. Antennas Propag.*, vol. 71, no. 1, pp. 1185–1189, Jan. 2023.

[9] N. Vidhya, L. C. Ong, M. Y. Siyal, and M. F. Karim, "A 2-D radon transformation for enhancing the detection and imaging of embedded defects in layered composite structures using millimeter-wave system," *IEEE Sensors J.*, vol. 20, no. 14, pp. 7750–7760, Jul. 2020.

[10] J. Wu and N. Dahnoun, "A health monitoring system with posture estimation and heart rate detection based on millimeter-wave radar," *Microprocessors Microsyst.*, vol. 94, Oct. 2022, Art. no. 104670.

[11] S. Wagner, A. Alkasimi, and A.-V. Pham, "Detecting the presence of intrusive drilling in secure transport containers using non-contact millimeter-wave radar," *Sensors*, vol. 22, no. 7, p. 2718, Apr. 2022.

[12] I. Iliopoulos, S. D. Meo, M. Pasian, M. Zhadobov, P. Pouliguen, P. Potier, L. Perregrini, R. Sauleau, and M. Ettorre, "Enhancement of penetration of millimeter waves by field focusing applied to breast cancer detection," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 3, pp. 959–966, Mar. 2021.

[13] H. Yang, D. Zhang, A. Hu, C. Liu, T. J. Cui, and J. Miao, "Transformer-based anchor-free detection of concealed objects in passive millimeter wave images," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5012216.

[14] B. Zhang, B. Wang, X. Wu, L. Zhang, M. Yang, and X. Sun, "Domain adaptive detection system for concealed objects using millimeter wave images," *Neural Comput. Appl.*, vol. 33, pp. 11573–11588, Mar. 2021.

[15] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, ''MobileNetV2: Inverted residuals and linear bottlenecks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[16] M. Tan and Q. Le, ''EfficientNet: Rethinking model scaling for convolutional neural networks,'' in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[17] S. Mehta and M. Rastegari, ''MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer,'' 2021, *arXiv:2110.02178*.

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, ''Swin transformer: Hierarchical vision transformer using shifted windows,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

**CHONGLIANG GAO** received the B.S. degree from Northeast Petroleum University, in 2011.

He is currently an Engineer with China Construction Shenzhen Decoration Company Ltd., Shenzhen, China. His research interests include curtain wall design and development. He received the National First-Class Architect, in 2020.

**JIE GAO** received the B.S. degree in measurement and control technology and instrumentation from Nanchang Hangkong University, in 2019.

He is currently an Assistant Engineer with China Construction Shenzhen Decoration Company Ltd., Shenzhen, China. His research interest includes curtain wall inspection technology.

**YAJUN CAO** received the B.S. degree in civil engineering from Wuhan University of Science and Technology.

He is currently a Senior Engineer with China Construction Shenzhen Decoration Company Ltd., Shenzhen, China. His research interest includes curtain wall construction technology. He has been professionally engaged in technical work for 14 years, presided over two projects with results appraised to reach the international leading level, obtained more than 30 patents, published 16 professional scientific and technological papers, and six provincial and ministerial level work methods.

Mr. Cao received the ''bridge building a glass curtain wall coupling vibration technology results'' and other results or the national building decoration industry, ''Top Ten Scientific and Technological Innovation Achievement Award,'' and ''Top Ten Scientific and Technological Paper Award.''