IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2023, accepted 4 June 2023, date of publication 7 June 2023, date of current version 22 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283589

RESEARCH ARTICLE

NDT Method for Weld Defects Based on FMPVit Transformer Model

YANG LIU, KUN YUAN, TIAN LI, AND SHA LI
Department of Computer and Information Technology, Liaoning Normal University, Dalian 116000, China

Corresponding author: Yang Liu (yangliu.0816@lnnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62002146 and Grant 61976109, in part by the Liaoning Revitalization Talents Program under Grant XLYC1807073 and Grant XLYC2006005, in part by the Scientific Research Project of Liaoning Province under Grant LJKZ0963, and in part by the Key Research and Development Projects of Liaoning Provincial Department of Science and Technology.

ABSTRACT The primary NDT method for welding defects is the image-based detection. Currently, the best performance for image-based detection is based on the transformer model. However, with its high accuracy, it has many limitations, such as large model parameters, large data sample requirements, and expensive computer resources. This model has a weaker ability to capture local features compared with global features. In this study, an improved and optimized welding defect detection and identification framework named Fast Multi-Path Vision transformer (FMPVit) is proposed based on the transformer model. This model uses a multilayer parallel architecture and enhances the local information capture ability of the model through advanced multiscale convolution feature aggregation and the addition of a new local convolution module. Finally, a validation test is carried out using an open dataset of weld seams. The model is proven to exhibit an evident performance improvement over the mainstream model baseline.

INDEX TERMS Deep learning, weld detection classification, line laser, one-dimensional sequential time series.

I. INTRODUCTION

Welding inspection or inspection of the quality of welding products is used to ensure the integrity, reliability, safety, and availability of welding product structures [1]. It is widely used in the aerospace, aviation, automobile, machinery, shipbuilding, and other industries. Although industrial production has matured, improper manual operation, environmental instability, and other problems may still lead to various welding defects in industrial products. Common weld defect types in steel plates are shown in Figure 1, including burr, concave, porous, and no defects. Welding defect detection technology can improve the production efficiency of the manufacturing industry, accelerate the production cycle of products, and reduce labor and material costs [2].

Conventional welding inspection methods, which are carried out by experienced professionals with the naked eye and professional tools, not only lead to low inspection efficiency but also have disadvantages such as misjudgment, sparse sampling, visual fatigue, and difficulty in ensuring the quality of inspection results. Modern welding testing methods can be divided into destructive testing (DT) and non-destructive testing (NDT) methods, depending on whether the test method is destructive. NDT is highly efficient and safe; it is the current mainstream welding defect detection method [3]. In welding NDT, most methods use images as the input, and the convolutional neural networks (CNNs) are employed.

Various models of the CNN stand out in ImageNet competitions and have high efficiency for image recognition [4]. The rapid development of CNNs in the field of computer vision has also led to the growth of the NDT of welding defects. AlexNet, ResNet, and other CNNs have been widely used in the field of NDT of welding defects and have achieved good recognition and detection results [5]. Although the CNN has the advantages of high generalization, fast recognition efficiency, and sensitivity to local features, their ability to capture global features is relatively weak [6].

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Kashif Bashir.

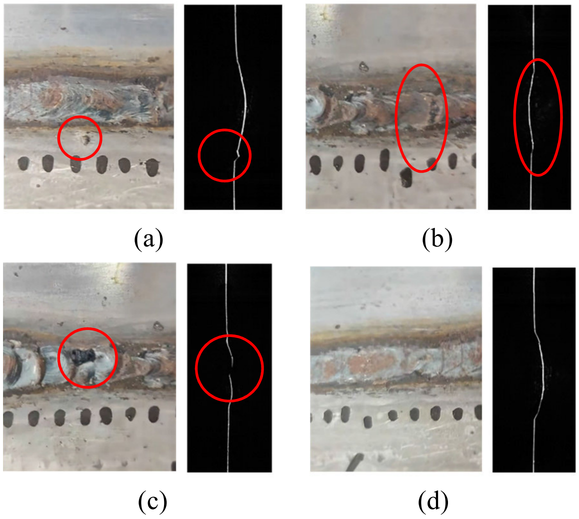

FIGURE 1. Common types of weld defects: (a) burr, (b) concave, (c) porosity, and (d) no defects. The left is an RGB image and right is line laser image.

Transformer model-based attention mechanisms have continuously emerged in the field of natural language processing and have developed into a recognized high-performance model structure [7]. With the first application of the vision transformer in the field of computer vision images, the model has become a competitor to CNNs [8]. It not only outperforms many popular CNN models in terms of computational efficiency and accuracy but also asserts huge development potential in the future. The transformer overcomes the limitation that the RNN model cannot be calculated in parallel. Compared with a CNN, the number of operations required to calculate the correlation between two locations does not increase with distance. However, the transformer can also focus more on global features than CNNs can [9]. Therefore, Multi-Path Vision transformer (MPVit) was proposed, which combines the advantages of convolution and a transformer [44].

In the industrial welding process, the scale of weld defects is relatively small compared with that of other defects, which results in fewer features in the image, making it more difficult for the model to capture such features. When identifying weld defects that are not evident, the effect of relying only on the CNN model is sometimes poor, and the advantage in this respect is not evident. In contrast to ordinary images, weld images are generally more regular and single, with more concentrated areas and fewer features. Although existing models have a higher recognition efficiency for ordinary images, their effect on weld images is unsatisfactory. The Multi-Path Vision transformer (MPVit) model combines the comprehensive advantages of the CNN and transformer, which are excessively redundant and still have deficiencies in small-scale target recognition. However, this model is too complex for weld defect detection, difficult to train, and requires a large number of samples. Therefore, we propose a new method called Fast Multi-Path Vision transformer (FMPVit) for welding defect detection, which is based on the aggregation optimization of local and multi-scale features. The public weld dataset and popular public datasets were compared with the popular neural network model.

The key contributions of this paper are:

1. The proposed method can improve the efficiency of small-scale target identification, such as weld defects. It achieves higher recognition accuracy and lower training costs by reducing the complexity of the model.

2. A fast multiscale convolution feature-priority aggregation module was proposed. The module reduces the redundancy of the three transform structures in the stacking stage of the MPVit model and significantly reduces the complexity of the original model.

3. We introduce local-to-global feature interaction (LGF) to take advantage of both the local connectivity of the convolutions and global context of the transformer.

In addition, the latest published weld dataset JPEGWD was used in the experiment, which includes 12000 RGB weld images of four different defect types. We also tested our method using LSWD-MTF, which is a two-dimensional time-series image dataset of LSWD encoded by the Markov Transition Field (MTF) method.

II. RELATED WORK

In welding image classification, deep learning methods have performed better than traditional machine learning methods. Many researchers have proposed NDT methods for welding defects based on deep-learning CNN. For example, Je-Kang et al. proposed a CNN-based method that uses a single RGB camera to examine welding defects on the transmission surface of an engine. This method consisted of two steps. In the first step, to extract the welding area from the captured image, a CNN-based method is used to detect the center of the engine transmission in the image. In the second stage, the extracted area is identified by another CNN as having either defects or no defects [11]. Zhang et al. designed an 11-layer CNN classification model based on weld images to identify weld penetration defects. The CNN model makes full use of arcs and combines them in various ways to form complementary features. The test results showed that the designed CNN model performed better than previous models [12]. Dong et al. proposed a multitask deep CNN for defect classification; they built a stack of encoder–decoder autoencoders to learn feature representations from ordinary images. For defect detection, this method can obtain results nearly as good as those of a supervised learning method without any data annotation [13]. Chen et al. focused on establishing an end-to-end automatic detection model for X-ray welding defects based on a deep learning algorithm to improve the accuracy and efficiency of detection. The characteristic information of welding defects is considered in their study, and the method of fast region-based CNNs (R-CNNs) is improved. A residual neural network (ResNet) was used to improve feature extraction ability [14].

The attention mechanism-based transformer model has continuously emerged in the field of natural language processing and developed into a recognized high-performance model structure [15]. With the first application of vision transformer in the field of computer vision images, this model has become a competitor to CNN [16]; it not only outperforms many popular CNN models in terms of computational efficiency and accuracy but also has a huge potential for development in the future. The transformer overcomes the limitation that the RNN model cannot be calculated in parallel. Compared with a CNN, the number of operations required to calculate the correlation between two locations does not increase with distance [17].

As a newly emerging deep-learning model, the transformer is no less advanced than the CNN model despite its short history, and its performance even exceeds that of conventional mainstream CNNs. The transformer model first gained a dominant position in the field of natural language processing owing to its high-performance recognition effect; it gradually rose in the field of computer vision (CV) once it became a strong competitor of CNNs. Some researchers have studied transformers in the field of NDT of welding defects. For example, Wang et al. [18] proposed a deep learning method based on a classic vision transformer to realize welding penetration recognition, constructed an image dataset composed of four different categories, and trained it from scratch to explore its feasibility in welding penetration recognition. Finally, ImageNet was used for pretraining to solve the problems of complex models and insufficient data, and the verification accuracy was improved by 4.45%. To explore the extensibility of the transformer, Gao et al. [19] proposed an improved structure called the variant swin transformer, based on the applicability of the swin transformer (SwinT). A new window shift scheme was designed to further enhance feature conversion between windows and increase the capability of the framework for defect detection. Considering the built private dataset, the overall framework, named the Cas-VSwin transformer, is superior to most existing models. Zhang et al. [20] proposed a novel network structure called the DRCDCT-Net. It was designed as a dual routing structure comprising a characteristic attention deficit diagnosis module (FAD) and cross-domain joint learning deficiency diagnosis module (CJLD). With the transformer as the core design, the FAD is primarily responsible for handling defect classification tasks with sufficient samples and relieving the problem of interdependence among features that are difficult for the CNN to learn. With the designed cross-domain joint learning network as the core, CJLD deals with the task of defect classification with extremely scarce samples and decoupling image domain features. The model achieved accuracy of 99.7 ± 0.2% and 90.0 ± 0.6% in the Northeastern University (NEU)-CLS and SEVERSTAL public datasets, respectively. Although the overall performance of the transformer method was better than that of the CNN method, some problems persisted.

Other NDT methods include magnetic particle testing, eddy current testing, magneto-optical imaging testing, ultrasonic testing, infrared testing, penetrant testing, and phased array ultrasonic testing [10]. Some NDT methods use signals for direct detection; however, some researchers also use two-dimensional images for detection. Because the dimensions of the one-dimensional signal description features are low and description of some defects is unclear, the detection performance is poor. Compared with one-dimensional signals, two-dimensional image detection is more stable, which can be better applied to a depth learning model, taking full advantage of its existing benefits. For example, some researchers [36] coded the one-dimensional structured light centerline of the weld surface into the corresponding two-dimensional time series image, which realized the dimensions of the weld defect and improved the depth learning model for the two-dimensional image.

III. METHODS

The detection of weld defects is a small-sample detection, and the problem of excessive redundancy exists in the transformer model, which leads to difficulty in training the model and a poor effect; moreover, the solution of data enhancement is too cumbersome. Therefore, in the field of welding defect detection, a lightweight model that combines the advantages of CNN and transformers and realizes small-sample training is necessary.

A. ARCHITECTURE

Figure 2 shows the FMPViT architecture. MPVit has an extended multi-path, based on ViT and XCiT [21], as well as an added convolution module. Although the MPVit model has a significantly improved detection performance and accuracy, it has become more complex and requires more training resources. For a small weld-defect detection dataset, the MPVit model is too large and difficult to train. To solve these problems, we use a variety of transformer architectures for reference and are committed to building a single transformer path-stacking framework combined with convolution modules [22], [23]. The goal of building the FMPVit model is to have a faster reasoning speed and lower computing cost, while achieving a higher performance than MPVit. As shown in Figure 2, we construct a four-stage feature hierarchy to generate feature maps at different scales. The characteristics of the MPVit model often require more computation; therefore, we adopted a series of measures to reduce the complexity of the MPVit model. To reduce the linear complexity of the model, only a single transformer structure is used for each stacking stage based on the MPVit model. In addition, a transformer encoder that decompositions self-attention in catCoaT [24] is used, and the convolutional stem block in LeViT [25] is used to improve the present low-level representation to prevent the loss of significant information.

As shown in Figure 3, we propose a new NDT framework for welding defects, i.e., FMPVit, based on the MPVit model, which has a faster training speed and higher accuracy

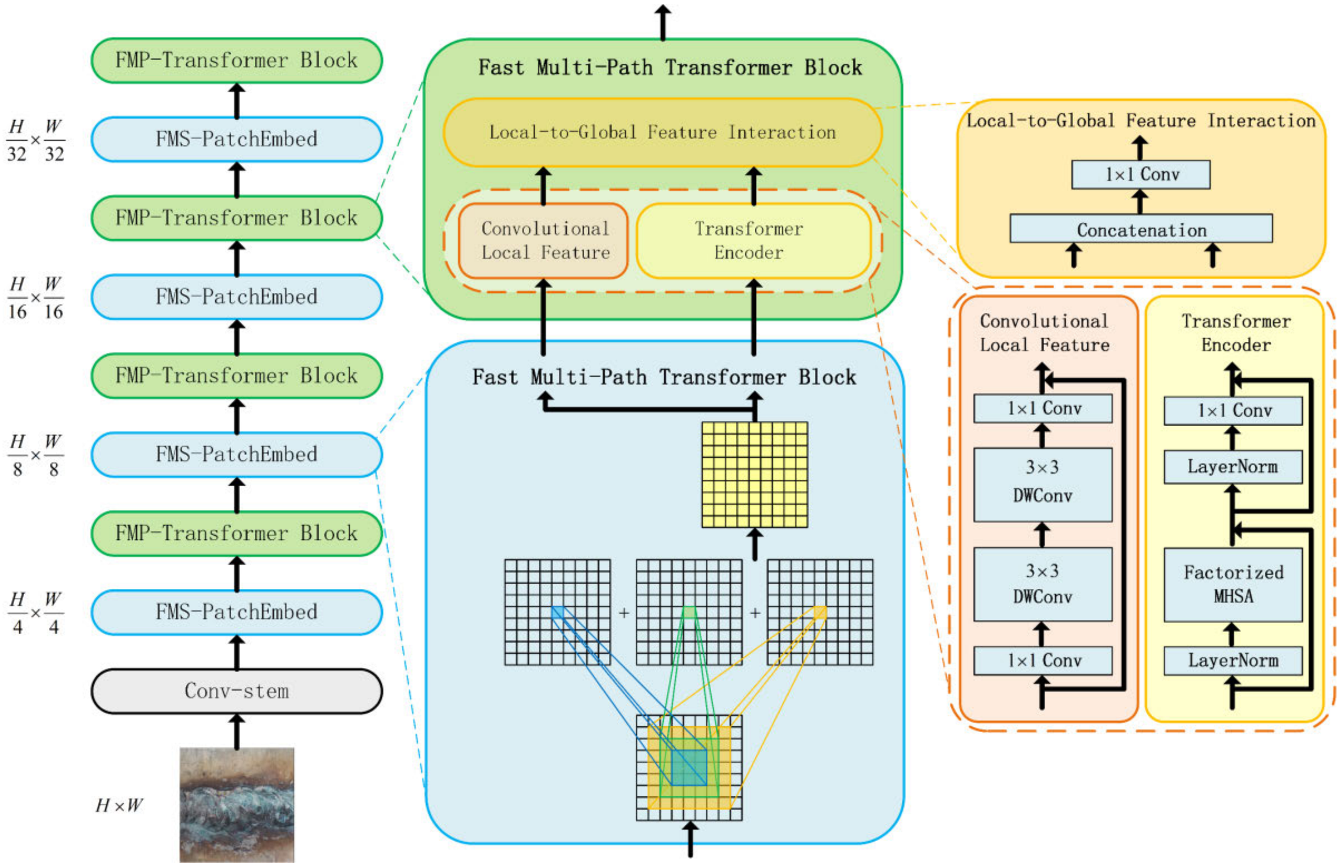
FMP-Transformer Block
FMS-PatchEmbed
$\frac{H}{32}\times\frac{W}{32}$
FMP-Transformer Block
FMS-PatchEmbed
$\frac{H}{16}\times\frac{W}{16}$
FMP-Transformer Block
FMS-PatchEmbed
$\frac{H}{8}\times\frac{W}{8}$
FMP-Transformer Block
FMS-PatchEmbed
$\frac{H}{4}\times\frac{W}{4}$
Conv-stem
$H\times W$
Fast Multi-Path Transformer Block
Local-to-Global Feature Interaction
Convolutional Local Feature
Transformer Encoder
Fast Multi-Path Transformer Block
+
+
Local-to-Global Feature Interaction
1×1 Conv
Concatenation
Convolutional Local Feature
1×1 Conv
3×3 DWConv
3×3 DWConv
1×1 Conv
Transformer Encoder
1×1 Conv
LayerNorm
Factorized MHSA
LayerNorm

FIGURE 2. Fast multi-path vision transformer (FMPViT) architecture.

while considering local and multi-scale features. In FMPVit, a fast multistage transformer structure is constructed, and a new 3 × 3 convolution is embedded to enhance the model's convolutional local feature capture ability. This solves the problem of the transformer lacking local features compared with the convolutional network. In addition, by aggregating the multiscale features of each stage into a transformer, the excess transformer path structure of the original framework was reduced, which significantly reduced the model complexity and improved the overall model performance efficiency. Notably, this is different from the existing vision transformers.

The excessive complexity of the model is not desirable in the field of NDT for weld defects. Because the sample size of a welding dataset is usually small, data enhancement is time-consuming and energy-intensive during the training process, and transfer learning for small samples makes the training process more complex. Overfitting and unstable training often occur during the large-scale model training. Therefore, it is necessary to reduce the path length and complexity of the model. It is also important to reduce the complexity of the model while considering its accuracy. Based on the multistage transformer architecture design with smaller complexity, we merged all redundant transformer structures in the MPVit model and proposed a fast multiscale patch embedding module.

B. FAST MULTI-SCALE PATCH EMBEDDING

To make better use of the fine-grained and coarse-grained visual tokens, a convolution operation with overlapping patches was used, similar to CNNs [26] and CvT [27]. By changing the size and filling amount of the convolution kernel, a same-size feature map with different feature information can be obtained. As shown in Figure 2, visual tokens of different sizes with the same sequence length can be generated with patch sizes of 3 × 3, 5 × 5, and 7 × 7. During implementation, because the continuous convolution operations of the same channel and filter size enlarge the receptive field (e.g., two 3 × 3 equal 5 × 5, and three 3 × 3 equal 7 × 7), the use of 3 × 3 convolution kernel substitution requires fewer parameters, thereby reducing complexity. We used three consecutive 3 × 3 convolutions with the same channel size; the fill was 1 and step length was s, where s was 2 when the spatial resolution was reduced; otherwise, it was 1. Because MPViT has more embedding layers owing to its multi-path structure, we reduce the parameters and computation of the convolutional local feature overhead by adopting 3 × 3 depthwise separable convolutions [28], which consist of 3 × 3 depthwise convolution followed by 1 × 1 pointwise convolution in the embedding layers.

For fast multiscale patch embedding, we proposed a different multiscale patch aggregation method. The polymerization process is illustrated in Figure 4. In this process, the size of

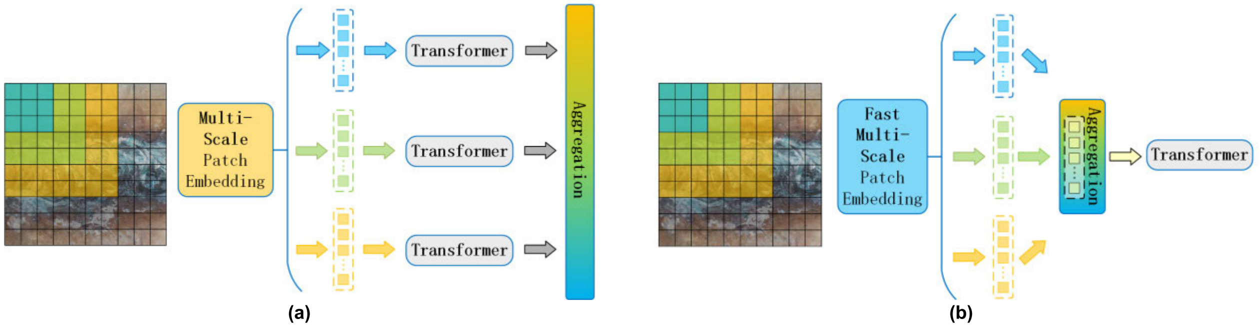

FIGURE 3. (a) Most advanced MPVIT models [44] use multi-scale patches and multi-path transformer encoders. (b) Our FMPViT uses multi-scale patch embedding, each multi-path embedded patch using only one independent transformer encoder.

the output feature matrices of the three convolution kernels (3 × 3, 5 × 5, and 7 × 7) was unified by padding to zero, and the three output features were superimposed by matrix addition. Finally, the superimposed features are input into a single transformer module to realize the early aggregation of the token, which can minimize the calculation amount of this module while ensuring that the multi-path and multi-scale convolution features are not lost.

Because the convolution results of different sized convolution kernels are different for different images and input of a single transformer structure is generally the token of a single image, it is necessary to unify the same size when merging three different feature maps. The process of zero padding is described in detail as follows.

We assume that, before padding, the input size is (H, W), filter size is (F_H, F_w), output size is (O_H, O_w), padding is P, and step length is S. The output size after padding is obtained using Equations (1) and (2).

$$O_H = \frac{H + 2P - F_H}{S} + 1 \tag{1}$$

$$O_W = \frac{H + 2P - F_W}{S} + 1 \tag{2}$$

After the output matrices A, B, and C of the three different paths (3 × 3, 5 × 5, and 7 × 7) were patched with zero to unify their sizes, the final aggregate matrix D was obtained by summing their matrices, as shown in Equation (3).

$$D = A + B + C \tag{3}$$

Because FMPViT has more embedding layers owing to its multi-path structure, a 3 × 3 deep separation convolution and 1 × 1 point convolution are adopted to reduce the model parameters and computational overhead. A 3 × 3 separable convolution improves the efficiency of the model, whereas a 1 × 1 point convolution reduces the dimensions, increases the depth of the model, and improves its nonlinear expression ability.

C. CONVOLUTION LOCAL FEATURE

As shown in Figure 5, to enable the model to effectively capture local features, we added a new 3 × 3 convolution kernel to the local convolution module. Two 3 × 3 convolution kernels are equivalent to a 5 × 5 convolution kernel. Although the two 3 × 3 convolution kernels must be convolved twice, the actual convolution operation efficiency of the convolution kernel is higher, there are fewer parameters, and the computer processing speed is faster. This optimization method appeared in early VGG networks [29]. In addition, replacing a 5 × 5 convolution kernel with two 3 × 3 convolution kernels increases the depth (number of layers) of the network, and the nonlinear expression of features is enhanced, which was also proven in later experiments [30]. We reduce the number of model parameters and computational overhead by adopting 3 × 3 depthwise separable convolutions, which consist of 3 × 3 depthwise convolutions followed by 1 × 1 pointwise convolutions in the embedding layers.

D. LOCAL-TO-GLOBAL FEATURE INTERACTION

The self-attention mechanism in the transformer can better capture long-term dependencies, i.e., the global context information; however, the capture ability of structural features and local relationship features is weak [31], [32], which can be compensated by local convolution. The CNN uses the same weight to process each patch in the image in terms of translation invariance and local connectivity [33]. This inductive bias encourages the CNN to exhibit a stronger dependence on texture when classifying visual objects [34]. Combining the advantages of the CNN's local feature capture with the advantages of the transformer's global feature capture can enhance the image feature information acquisition of the model. Therefore, a local-to-global feature interaction module was proposed for FMPVit. We used a deep residual bottleneck block, which comprises a 1 × 1 convolution, two 3 × 3 depth convolutions, and 1 × 1 convolution composition, with the same channel size and residual connection [35]. The local and global features are aggregated by concatenation as follows:

$$U_i = Concat([R_i, L_{i,0}, L_{i,1}, \ldots, L_{i,j}]) \tag{4}$$

$$X_{i+1} = P(U_i) \tag{5}$$

The 2D-reshaped global features from each transformer $L_{i,j} \in \mathbb{R}^{H_i \times W_i \times C_i}$, $R_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ represent local feature, where j is the index of the path, i is the stage number, U_i is the aggregated feature, and $P(\cdot)$ is a function which

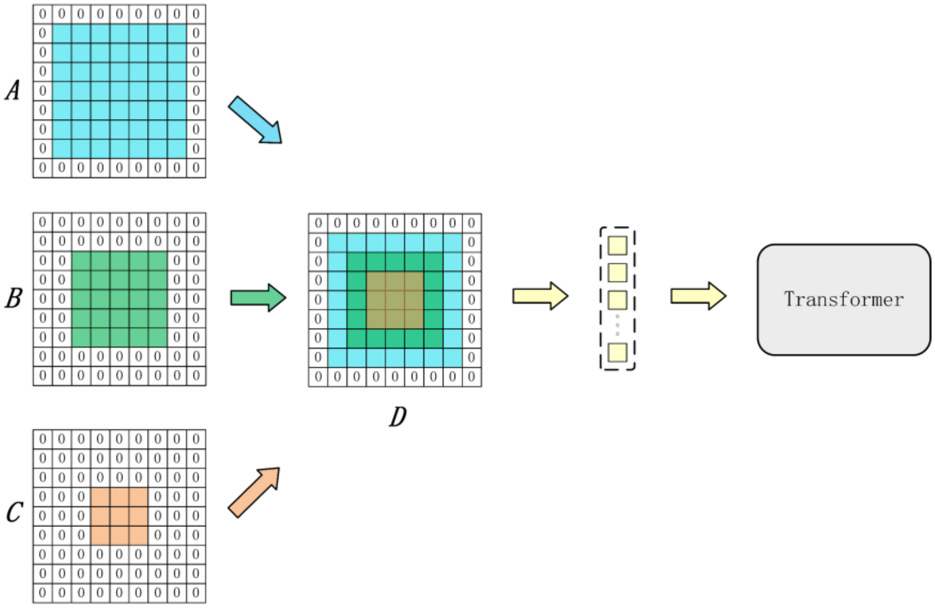

FIGURE 4. Fast multi-scale patch embedding process.

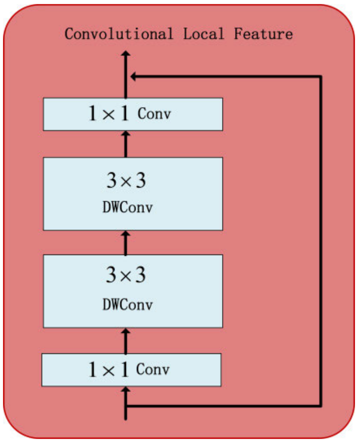

FIGURE 5. Convolutional local feature.

learns to interact with features, yielding the final feature $X_{i+1} \in \mathbb{R}^{H_i \times W_i \times C_{i+1}}$ with the size of next stage channel dimension C_{i+1}.

E. MODEL CONFIGURATION

To reduce computational burden, the effective factored self-attention proposed in CoaT was used as follows:

$$FactorAtt\,(Q, K, V) = \frac{Q}{\sqrt{C}}(softmax\,(K)^T V), \tag{6}$$

where $Q, K, V \in \mathbb{R}^{N \times C}$ are linearly projected queries, keys, values, respectively; N, C denote the number of tokens and embedding dimension, respectively. The factor self-attention method reduces the FMPVit model parameters and FLOPs and improves the overall efficiency of the model. We did not use the traditional multi-path structure in FMPVit, but reduced it to a transformer path through the method described in Section III-B, which greatly reduced the resource expenditure of the model. The application of the factor self-attention method to the FMPVit model maximized the overall efficiency of the model.

In addition, we found that after reducing the original three-path transformer, the multiscale aggregated single-path FMPVit showed better performance in classification, with a faster training speed and higher accuracy. This demonstrates that aggregating multiscale features into a single transformer path in advance is an effective approach. We built three different versions of FMPVit: the original basic scale FMPViT-Base (*M), expansion of two layers of FMP-Transformer Block and medium-scale FMPVit-Base+(*M) of MS-PatchEmbed, and expansion of four layers of FMP-Transformer Block and large-scale FMPVit-Base++(*M) of MS-PatchEmbed. All FMPVit models used eight transformer encoder heads. Table 1 shows the details of the FMPVit models.

TABLE 1. FMPViT configurations.

FMPVit	#Layers	Param.	GFLOPs
FMPVit-Base	[1,2,8,1]	4.5	2.2
FMPVit-Base+	[1,2,12,1]	6.2	3.1
FMPVit-Base++	[1,2,16,1]	7.6	4.4

IV. EXPERIMENT

A. EXPERIMENT SETTING

The Python programming language based on Python3.7 environment was used in the experiment, and the mainstream

TensorFlow framework was built on PyCharm. The framework environments used were Keras 2.2.4, PyTorch-GPU 2.2.0, CUDA 10.1, and CuDNN 7.6. The hardware experimental environment was a single all-in-one NVIDIA RTX 3090 GPU and an Intel Core i9 CPU. Table 2 lists the detailed parameter settings for the experimental environment. All network models had the same setup parameters and dataset.

B. EXPERIMENT DATASET

1) JPEGWD DATASET

In the field of weld defect detection, only a few large-scale weld image datasets are currently available. The published weld datasets are limited; therefore, to better guarantee the experimental results, two datasets that are currently available are used. One is Joint Photographic Experts Group Welding (JPEGWD), a JPEG format industrial weld image dataset newly published by Chen et al. [36], from the artificial intelligence laboratory of Beijing ByteDance Technology Co., Ltd. The other is the Linear Structured Light Welding (LSWD) and Linear Structured Light Welding Markov Transfer Field (LSWD-MTF), which was newly published by Liu et al. [37].

As shown in Table 2, the JPEGWD weld dataset consists of 12000 images; is the only RGB image dataset containing common weld defect types. Based on the image of the defect type in the dataset, the dataset was subdivided into four weld types: burr, concave, hole, and no-defect. This results is two versions of the dataset, each comprising 4000 images, i.e., the weld dataset JPEGWD for four image types: burr, concave, porous, and no-defect. The size of each image in the welding dataset was 500 × 500, and the image format was JPEG.

TABLE 2. JPEGWD dataset settings.

Defect type	Number
	JPEGWD
burr	4000
concave	4000
porosity	4000
no-defect	4000

2) LSWD DATASET

The LSWD dataset was collected by the line-structured light equipment described in this section; 1680 original line-structured light images were sorted out under short time, including burrs (798), depressions (421), holes (108), and no defects (353). In the experiment, the steel plate welds were marked at 0.5-cm intervals, and the image data of the weld line structured light were collected twice; the first time, by the way of the line structured light and weld line perpendicular, and the second time, the original structured light images of the weld line were acquired with an angle of 30° between the line structured light and weld line. There were 1680 original structured light images of the weld line obtained at 0.5-cm intervals. The four types of defects were manually classified and labeled. The size of the original line-structured light image was 1280 × 520 pixels. Because the original experiment sample is small, in the following experiments, the unified use of scaling, rotation, and other data expansion methods was adopted to expand the dataset and obtain better experimental results. This resulted in a total of 6720 structured light images, and Table 3 presents the dataset setup.

LSWD-MTF is a two-dimensional color time-series image obtained using the MTF coding method from one-dimensional weld height information data in the LSWD dataset, and the dataset images correspond to the LSWD dataset one-by-one. Equation (7) describes the coding principle.

$$M = \begin{pmatrix} W_{ij}|x_1 \in q_i, x_1 q_j & \cdots & W_{ij}|x_1 \in q_i, x_n \in q_j \\ W_{ij}|x_2 \in q_i, x_1 \in q_j & \cdots & W_{ij}|x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ W_{ij}|x_n \in q_i, x_1 \in q_j & \cdots & W_{ij}|x_n \in q_i, x_n \in q_j \end{pmatrix} \tag{7}$$

When a time series X is given to define the packet number box Q of the time series and each X_i in the time series is assigned to the respective storage box $q_j (j \in [1, Q])$, the weighted adjacent matrix W, which can be constructed as $Q \times Q$, is converted from the first-order Markov chain count-point box. W is not sensitive to the distribution of X, which overcomes the disadvantage of insensitive sequence time dependence. The LSWD-MTF dataset corresponding to the LSWD was obtained by encoding the one-dimensional weld information into the MTF two-dimensional time-series image, and the size and quantity parameters of the two datasets were consistent. The one-dimensional weld height information of the original line-structured light is encoded into the corresponding two-dimensional information using Equation (7), and the corresponding MTF two-dimensional color time-series image can be generated using the Python pseudo-color library. Table 3 presents the number, configuration, and generation effects of the datasets.

TABLE 3. LSWD dataset settings.

Defect type	Number	
	LSWD	LSWD-MTF
burr	1680	1680
concave	1680	1680
porosity	1680	1680
no-defect	1680	1680

C. EXPERIMENT PROCESS

The performance of FMPVit was evaluated based on the JPEGWED and LSWD weld datasets. The two datasets have the same weld defect category, including burr, concave, hole and no-defect. The comparison results of JPEGWED and LSWD weld datasets are shown in Tables 4, 5, and 6. To highlight the advantages of the model, we used the mainstream CNN model and mainstream visual transformer model in the comparative experiment. Mainstream CNN models include VGG-16 [38], ResNet50 [39], GoogleNet [40], DenseNet [41], and MobileNet [42], and mainstream vision transform models include Vit [43], Swin [44], and MPVit [45]. In FMPVit, the Adam optimizer [46] was used to train 300 iterations; the batch size was 64, and the initial learning rate was 0.001. This was scaled using the cosine attenuation learning rate scheduler, and each image was cropped to 224 × 224 pixels, which is consistent with Table 3.

TABLE 4. Experimental comparison results of different models in JPEGWD weld dataset.

Route	Model	Param .(M)	GFL OPs	ACC
CNN	VGG-16	6.4	1.2	75.08%
	ResNet50	5.9	1.7	78.53%
	GoogleNet	6.5	2.0	81.39%
	DenseNet	7.3	1.5	81.26%
	MoblileNet	8.6	2.2	82.78%
Transfor mer	Vit	6.1	3.4	83.92%
	Swin	7.2	2.6	84.42%
	MPVit	5.6	1.7	84.49%
	FMPVit-base	4.5	2.8	86.60%

Table 4 presents the experimental comparison results of different models in the JPEGWD weld dataset. In Table 4, two experimental routes can be seen: the CNN and transformer. In the mainstream CNN model route, the accuracy rate of the JPEGWD dataset gradually increased with an increase in model complexity and parameter quantity. The highest accuracy rate for this route was 82.78% for the MobileNet network. In the mainstream transformer model route, the JPEGWD dataset also presents the same trend; however, in FMPVit, not only the accuracy and GFLOPs improved, but the complexity and parameters of its model are also greatly reduced.

To effectively highlight the performance of the model, we divided the JPEGWD weld dataset into two categories according to the presence or absence of defects and named the weld dataset JPEGWD-2CLASS. The weld data samples were all obtained from JPEGWD, changing only the four categories into two categories: with or without defects. Owing to the small number of defect-free samples in the dataset, dataset enhancement methods such as zooming, rotating, and cropping are used for data enhancement [47]. There were 6000 images with and without defects in the enhanced dataset. The same experimental environment parameters and models were used in the experiments, and the results are listed in Table 4.

TABLE 5. Experimental comparison results of different models in LSWD weld dataset.

Route	Model	Param.	GFL OPs	ACC
CNN	VGG-16	5.5	1.1	93.48%
	ResNet50	5.4	1.5	94.94%
	GoogleNet	6.2	1.6	95.50%
	DenseNet	5.5	2.2	95.25%
	MoblileNet	7.6	2.3	97.72%
Transfor-mer	Vit	5.8	2.2	96.71%
	Swin	6.2	2.1	98.33%
	MPVit	5.7	1.6	98.61%
	FMPVit-base	**4.2**	**2.5**	**99.72%**

Table 5 presents the experimental comparison results of the different models for the LSWD weld dataset; it lists the two experimental routes: CNN and transformer. As the LSWD weld dataset had better image quality and recognition, it performed very well in the overall experiment. Among the mainstream CNN model routes, the highest accuracy rate of the LSWD dataset exceeds that of the transformer model route by 97.72%. Compared with other models in Table 5, it is worth noting that in FMPVit, not only the accuracy and GFLOPs improved but also the complexity and parameters of its model reduced significantly.

TABLE 6. Experimental comparison results of different models in LSWD-MTF weld dataset.

Route	Model	Param.	GFL OPs	ACC
CNN	VGG-16	5.1	1.5	98.23%
	ResNet50	6.0	1.7	97.83%
	GoogleNet	5.7	1.6	98.59%
	DenseNet	5.3	2.0	98.79%
	MoblileNet	6.7	2.1	98.46%
Transfor -mer	Vit	6.5	2.8	98.89%
	Swin	7.1	2.4	99.54%
	MPVit	5.5	1.9	99.33%
	FMPVit-base	**4.8**	**2.7**	**99.89%**

In the LSWD weld dataset, the author provided a version of the two-dimensional color time series image dataset encoded by the proposed MTF method [48], which was also used in our comparative experiments. The dataset and sample numbers individually correspond to the original LSWD weld dataset samples. We call this version of the weld dataset LSWD-MTF. The same experimental environment and network model parameters were used in this experiment. The final experimental results are listed in Table 6. The parameters of the transformer series model are more complex than those of the CNN series model, but the accuracy and GFLOPs

TABLE 7. Ablation experiment analysis of FMPVit model based on JPEGWD weld dataset.

Model	Param.	GFLOPs	Time	ACC
MPVit	5.6	1.7	7 h 24 min	84.49%
FMPVit-Base	4.5	2.2	4 h 21 min	86.60% (+2.11%)
FMPVit-Base+	6.2	3.1	5 h 37 min	87.16% (+2.97%)
FMPVit-Base++	7.6	4.4	6 h 28 min	87.79% (+3.30%)

TABLE 8. Ablation experiment analysis of FMPVit model based on JPEGWD weld dataset.

Dataset	MPVit	FMPVit model improvement results		
		All (Token+Conv)	Only Token	Only Conv
JPEGWD	84.49%	86.60% **(+2.11%)**	85.33% **(+0.84%)**	85.43% **(+0.94%)**
LSWD	98.61%	99.72% **(+1.11%)**	99.17% **(+0.56%)**	99.44% **(+0.72%)**
LSWD-MTF	99.33%	99.89% **(+0.56%)**	99.64% **(+0.31%)**	99.75% **(+0.42%)**

are significantly higher than those of the CNN series model. Moreover, in the transformer series models, FMPVIT has excellent performance as well as the highest accuracy and GFLOPs, while maintaining minimum parameter complexity.

D. ABLATION STUDY

For the FMPVit model, we performed ablation experiments with different layers; the specific configurations are listed in Table 1. We used the JPEGWD weld dataset for comparative experimental research, as listed in Table 7. The FMPVit-Base++ model version with the largest number of layers is more accurate; however, it exhibits greater model complexity and number of parameters. The basic version of FMPVit can have fewer model parameters and lesser complexity while maintaining a small gap with the high-stack version, which is beneficial for reducing the model training time. In Table 7, we can see that the accuracy of the FMPVit and MPVit models improved by 2–3% on JPEGWD. While maintaining the improvement, all versions of FMPVit have a smaller model size and faster reasoning speed, which is commendable.

We analyzed the ablation experimental results of different configurations of the FMPVit model. Two configurations stand out as the two improved parts of the proposed model; one is the newly added 3 × 3 convolution module, and the other is the multi-scale multi-path convolution, which is pre-aggregated into a single transformer module. The results of the ablation experiments for all the configurations are listed in Table 8; the two configurations improve the accuracy and efficiency of the model relative to MPVit. The addition of the convolution module has a greater impact on model improvement but consumes more model parameters. The improved method of early aggregation not only improves the performance of the model but also greatly reduces the number of model parameters and computational resources.

V. CONCLUSION

Based on the transformer model, this study proposed an improved and optimized welding defect detection and recognition framework, namely, a Fast Multi-path Vision Transformer (FMPVit). The performance of the transformer network model in weld defect detection was studied, although the CNNs are widely used in the field of NDT weld defect detection. The experiment shows that the new CNN module and single transformer module in the model could be combined with higher accuracy and smaller model size, while reducing the path. Compared with the mainstream network model, the FMPVit proposed in this study can more effectively capture the global and local feature information of the weld image, exhibiting better performance. This can improve the accuracy while ensuring that the model is sufficiently simplified. The model adopts a multilayer parallel architecture and combines the advanced multiscale convolution feature priority aggregation with a new local convolution module to enhance its local information capture ability. Finally, the LPEGWD and LSWD universal weld datasets prove that the model exhibits an evident performance improvement at the baseline of mainstream models.

REFERENCES

[1] R. Saha and P. Biswas, "Current status and development of external energy-assisted friction stir welding processes: A review," *Welding in the World*, vol. 66, pp. 577–609, Jan. 2022.

[2] A. S. Madhvacharyula, A. V. S. Pavan, S. Gorthi, S. Chitral, N. Venkaiah, and D. V. Kiran, "In situ detection of welding defects: A review," *Welding World*, vol. 66, pp. 1–18, Jan. 2022.

[3] F. Zhang, B. Zhang, and X. Zhang, "Automatic forgery detection for X-ray non-destructive testing of welding," *Weld. World*, vol. 66, no. 4, pp. 673–684, Apr. 2022.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[5] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar. 2021.

[6] Y. Liu, J. Hu, X. Kang, J. Luo, and S. Fan, "Interactformer: Interactive transformer and CNN for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5531715.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[8] T. Kattenborn, J. Leitloff, F. Schiefer, and S. Hinz, "Review on convolutional neural networks (CNN) in vegetation remote sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 24–49, Mar. 2021.

[9] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[10] A. Chabot, N. Laroche, E. Carcreff, M. Rauch, and J.-Y. Hascoët, "Towards defect monitoring for metallic additive manufacturing components using phased array ultrasonic testing," *J. Intell. Manuf.*, vol. 31, no. 5, pp. 1191–1201, Jun. 2020.

[11] J.-K. Park, W.-H. An, and D.-J. Kang, "Convolutional neural network based surface inspection system for non-patterned welding defects," *Int. J. Precis. Eng. Manuf.*, vol. 20, no. 3, pp. 363–374, Mar. 2019.

[12] Z. Zhang, G. Wen, and S. Chen, "Weld image deep learning-based online defects detection using convolutional neural networks for AL alloy in robotic arc welding," *J. Manuf. Processes*, vol. 45, pp. 208–216, Sep. 2019.

[13] X. Dong, C. J. Taylor, and T. F. Cootes, "Defect classification and detection using a multitask deep one-class CNN," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 3, pp. 1719–1730, Jul. 2022.

[14] Y. Chen, J. Wang, and G. Wang, "Intelligent welding defect detection model on improved R-CNN," *IETE J. Res.*, pp. 1–10, Mar. 2022.

[15] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 15908–15919.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.

[18] Z. Wang, H. Chen, Q. Zhong, S. Lin, J. Wu, M. Xu, and Q. Zhang, "Recognition of penetration state in GTAW based on vision transformer using weld pool image," *Int. J. Adv. Manuf. Technol.*, vol. 119, nos. 7–8, pp. 5439–5452, Apr. 2022.

[19] L. Gao, J. Zhang, C. Yang, and Y. Zhou, "Cas-VSwin transformer: A variant Swin transformer for surface-defect detection," *Comput. Ind.*, vol. 140, Sep. 2022, Art. no. 103689.

[20] J. Wang, Q. Zhang, and G. Liu, "DRCDCT-Net: A steel surface defect diagnosis method based on dual-route cross-domain convolution-transformer network," *Meas. Sci. Technol.*, vol. 33, no. 9, 2022, Art. no. 095404.

[21] A. El-Nouby, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, and G. Synnaeve, "Xcit: Cross-covariance image transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 20014–20027.

[22] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 548–558.

[23] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, "Focal self-attention for local-global interactions in vision transformers," 2021, *arXiv:2107.00641*.

[24] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9961–9970.

[25] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "LeViT: A vision transformer in ConvNet's clothing for faster inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12239–12249.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[27] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.

[28] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[29] S. Ghosh, A. Chaki, and K. Santosh, "Improved U-Net architecture with VGG-16 for brain tumor segmentation," *Phys. Eng. Sci. Med.*, vol. 44, no. 3, pp. 703–712, Sep. 2021.

[30] M. Agarwal, A. Singh, S. Arjaria, A. Sinha, and S. Gupta, "ToLeD: Tomato leaf disease detection using convolution neural network," *Proc. Comput. Sci.*, vol. 167, pp. 293–301, 2020.

[31] M. Amirul Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" 2020, *arXiv:2001.08248*.

[32] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Dec. 1999, pp. 1150–1157.

[33] O. S. Kayhan and J. C. van Gemert, "On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14262–14273.

[34] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLOS Comput. Biol.*, vol. 14, no. 12, Dec. 2018, Art. no. e1006613.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] *Weld Quality Inspection Items (JPEGWD Weld Data Set)*. [Online]. Available: https://github.com/ppogg/YOLOv5-Lite

[37] Y. Liu, K. Yuan, T. Li, S. Li, and Y. Ren, "NDT method for line laser welding based on deep learning and one-dimensional time-series data," *Appl. Sci.*, vol. 12, no. 15, p. 7837, Aug. 2022.

[38] R. M. Nazarov, Z. M. Gizatullin, and E. S. Konstantinov, "Classification of defects in welds using a convolution neural network," in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (ElConRus)*, Jan. 2021, pp. 1641–1644.

[39] W. Dai, D. Li, D. Tang, H. Wang, and Y. Peng, "Deep learning approach for defective spot welds classification using small and class-imbalanced datasets," *Neurocomputing*, vol. 477, pp. 46–60, Mar. 2022.

[40] R. Anand, T. Shanthi, M. S. Nithish, and S. Lakshman, "Face recognition and classification using GoogleNET architecture," in *Soft Computing for Problem Solving*. Singapore: Springer, 2020, pp. 261–269.

[41] Y. Zhu and S. Newsam, "DenseNet for dense flow," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 790–794.

[42] Y. Chen, X. Dai, D. Chen, M. Liu, X. Dong, L. Yuan, and Z. Liu, "Mobile-former: Bridging MobileNet and transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5260–5269.

[43] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.

[44] 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.

[45] Y. Lee, J. Kim, J. Willette, and S. J. Hwang, "MPViT: Multi-path vision transformer for dense prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7277–7286.

[46] S. Mehta, C. Paunwala, and B. Vaidya, "CNN based traffic sign classification using Adam optimizer," in *Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS)*, May 2019, pp. 1293–1298.

[47] C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, and D. Tao, "An underwater image enhancement benchmark dataset and beyond," *IEEE Trans. Image Process.*, vol. 29, pp. 4376–4389, 2020.

[48] M. Bugueño, G. Molina, F. Mena, P. Olivares, and M. Araya, "Harnessing the power of CNNs for unevenly-sampled light-curves using Markov transition field," *Astron. Comput.*, vol. 35, Apr. 2021, Art. no. 100461.

YANG LIU received the B.S. degree in mechanical engineering from Dalian Jiaotong University, in 2010, and the M.S. and Ph.D. degrees in computer science from Pukyong National University, in 2012 and 2016, respectively. Since 2016, he has been a Lecturer with Liaoning Normal University. His research interests include computer vision, machine learning, and SAR image applications.

KUN YUAN received the B.S. degree in engineering and in computer science and technology from the University of Jinan, Shandong, China, in 2020. He is currently pursuing the M.S. degree in computer science with Liaoning Normal University, Dalian, Liaoning, China.

TIAN LI received the B.S. degree in engineering and in computer science and technology from Liaoning Normal University, Dalian, Liaoning, China, in 2020, where she is currently pursuing the M.S. degree in computer science.

SHA LI received the B.S. degree in engineering and in computer science and technology from Liaoning Normal University, Dalian, Liaoning, China, in 2021, where she is currently pursuing the M.S. degree in computer science.