Surface Defect Detection of Artworks Based on Deformable Convolutional Networks

Ping Miao, Wei Wu, Xin Wang, Xiaoying Jiang

Abstract—**As important carriers of cultural heritage, artworks are prone to developing cracks, rust, black spots, and uneven glaze surfaces during environmental changes and long-term preservation. Traditional visual inspection models for artworks often struggle to distinguish between textures and defects, resulting in high missed detection rates and low low precision. To resolve these challenges, this paper presents a deformable convolutional network model for detecting surface defects in artworks. To enhance the detection accuracy specifically for irregularly shaped defects, a multi-scale feature extraction module incorporating Deformable Convolutional Networks (DCN) is designed, and a hybrid DCNv4 module is employed to extract defect features across different scales. Additionally, the model incorporates an axial attention mechanism to fuse global perception and local deformation modeling at a very low cost, thereby boosting detection capabilities of small target defects. Furthermore, a Spatial Frequency Synergistic DSConv (Depth Separable Convolution) module has been developed, which integrates wavelet decomposition layers to strengthen frequency-domain defect edge features, reduce model parameters, and improve detection speed and flexibility. The average detection accuracy of this algorithm reaches 97.7%, making it highly competitive with cutting-edge techniques for surface defect detection in artworks.**

Index Terms—**Artworks defect detection, Multi-scale features, Deformable convolution, Attention mechanism**

I. Introduction

ARTWORKS hold significant artistic value but are highly susceptible to environmental fluctuations (e.g., humidity, temperature) and degradation during long-term preservation, leading to surface defects such as cracks, rust spots, dark spots, and uneven glaze [1]. The causes of these defects are diverse, including environmental conditions, preservation methods, wear and tear from use, and human interference. These defects not only impair the aesthetic value of the artworks but also present challenges for their protection and restoration.

Therefore, prompt and efficient detection and repair of surface defects in artworks is essential for the preservation and continuation of cultural heritage. Given the professional nature and technical challenges of detecting surface defects in artworks, traditional methods often rely on human expertise, such as visual inspection and the use of magnifying glasses [2]. These methods are simple to operate but have drawbacks such as low efficiency, strong subjectivity, and a high rate of missed detections, which usually do not manage to satisfy the needs of massive detection of surface defects in artworks.

In recent decades, employing deep learning methods for surface defect detection has emerged as a prominent research area, due to its Outstanding defect detection results in the complex scenarios [3]. Traditional Convolutional Neural Network (CNN) [4] has shown great effectiveness in the tasks of image classification and recognition. However, CNN models typically use fixed convolution kernels for feature extraction and perform fixed-size sampling on input images. This approach cannot effectively capture the irregular shapes and texture features of defects, thus limiting their effectiveness in detecting irregular defects. Therefore, when detecting irregular defects, CNN models may encounter false positives. Addressing the surface defect detection challenges in artworks is crucial, and the main issues are as follows:

- Insufficient training data: Most existing deep learning models are trained on general image datasets, lacking the generalization needed for the specialized and complex nature of artworks surface defects.
- Inadequate feature extraction: Traditional CNN models with fixed convolution kernels. As a result, effectively identifying the irregular geometries and surface textures of artwork defects poses a significant difficulty, resulting in lower detection accuracy.
- High demand for real-time performance: Artworks surface defect detection often requires on-site inspection, necessitating a high degree of real-time performance from the detection method. The computational intensity of existing deep learning models makes it challenging to meet these real-time requirements, limiting their application in practice.
- Lack of effective small target detection methods: Artworks surface defects often include small targets, such as fine cracks and pinholes, which can be obscured by larger targets and are difficult for existing models to detect effectively.

In order to overcome the cited challenges related to artwork surface defect detection, we have developed an innovative detection scheme that dexterously integrates attention mechanisms with deformable convolution. The Deformable Convolutional Network (DCN) [5] extends standard convolution operations to address spatial

Manuscript received January 7, 2025; revised May 1, 2025.

This work was supported in part by Liaoning Provincial Science and Technology Plan Joint Project(2024-MSLH-385).

Ping Miao is an associate professor of LuXun Academy of Fine Arts, Dalian, China.(e-mail:mp18018981316@163.com).

Wei Wu is a lecturer of LuXun Academy of Fine Arts, Dalian,China. (corresponding author to provide e-mail:waynewaltz2020@gmail.com).

Xin Wang is an associate professor of the School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang, China. (e-mail: wangx7988@sjzu.edu.cn).

Xiaoying Jiang is a third-year undergraduate of the School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang, China. (e-mail: 13513420447@163.com).

deformations, which is particularly crucial in the detection of artworks defects. Traditional convolution layers typically assume a strict correspondence between features and the input grid, but the defects on artworks surfaces often exhibit diverse shapes, sizes, and orientations.

In addition, DCN effectively models the irregular geometric forms of defect features by introducing offsets within the convolution kernels. This capability allows the network to concentrate on the irregular shapes and contours of defects, areas that conventional convolutions often struggle to capture effectively due to their fixed grid structure. The importance of this is magnified when dealing with complex defect configurations, especially if these defects are partially concealed by stains, peeling, or other surface blemishes.

The network's focus on visible defect segments makes DCN particularly adept at dealing with partial occlusion. Moreover, its multi-scale feature integration is especially beneficial for identifying small defects that are challenging to spot at a single scale. By dynamically adjusting the sampling grid, DCN encapsulates enhanced contextual information around defects, which aids in distinguishing true flaws from background noise or unrelated surface characteristics. Crucially, in real-world applications, DCN's proficiency in modeling spatial transformations ensures robustness against changing environmental and lighting conditions, significantly enhancing the reliability of defect detection.

The main breakthroughs of this approach include:

- A multi-scale feature extraction module incorporating DCN has been developed. By utilizing DCNv4 to process features at diverse scales, the module effectively enhances the network ' s competence in detecting irregular defects and improves its generalization capabilities.
- Reducing model parameters and increasing detection speed are the primary goals of developing the DSConv module. This module decomposes standard convolutions into depthwise and pointwise convolutions, efficiently cutting down the parameter count and accelerating detection, meeting real-time performance requirements.
- To more effectively enhance the Identification performance for small target defects, the MHSA attention mechanism has been introduced. This mechanism employs dynamic weight allocation, facilitating the network's prioritization of target regions. As a result, the network's proficiency in detecting small targets is improved, and the complexities inherent in small target detection are mitigated.
- This module integrates defect features from various scales via multi-scale feature fusion, further enhancing the model's detection precision and stability.

The remainder of this paper is structured as follows. Section II outlines the related works. Section III details the proposed method in explicit terms. Section IV contains the experimental results and their analysis. In Section V, we offer the conclusions.

II. Related Work

With the rapid advancement of digital technology, an increasing number of scholars are beginning to apply artificial intelligence (AI) techniques to the field of detecting surface flaws in artworks. Through intelligent analysis of artworks images, remarkable research results have been achieved.

Wang et al. [6] focused on proposing a deep learning-based method for artworks surface defect detection, which utilizes CNN to extract image features and employs feature fusion techniques to enhance detection accuracy. CNN effectively learns the underlying features of images and extract higher-level features through multi-layer convolutional and pooling operations. Through feature fusion, features from different levels are combined, enabling the model to comprehensively understand the surface information of artworks. Testing indicates that this technique successfully distinguishes between defects such as cracks and stains on artwork surfaces, achieving a high level of identification accuracy. However, the method primarily focuses on common defects like cracks and stains, and its detection performance for rare or complex defects, such as flaking and discoloration, may not be ideal.

Zhang et al. [7] developed a multi-task learning-based approach specifically for artwork surface defect detection, simultaneously conducting defect detection and classification with high accuracy. Multi-task learning effectively leverages the correlations between tasks, improving the model's generalization ability. Liu et al. [8] proposed a method leveraging Generative Adversarial Networks (GANs) to restore the surface integrity of artworks. This method generates repair images that closely resemble the original artworks surface, achieving satisfactory restoration results. The makeup of a GAN includes both a generator and a discriminator, where the generator creates repair images and the discriminator assesses their authenticity. Through adversarial training, the generator produces increasingly realistic repair images, while the discriminator becomes more accurate in distinguishing real from fake images. This method effectively restores defects like cracks and stains on artworks surfaces while preserving the original style and texture of the artworks. He et al. [9] showcased a deep learning method for artwork surface defect classification, achieving high accuracy in image categorization via CNN. Additionally, this approach utilizes an attention mechanism aimed at concentrating the identification focus on defect regions, enhancing the model's defect recognition ability. Attention mechanisms effectively highlight key information in images while suppressing irrelevant interference. In this method, the attention mechanism is used to extract features from defect regions, which are then classified by a classifier. Testing results confirm the method's reliable detection of cracks and stains on artwork surfaces, coupled with high precision in the classification task.

While existing deep learning techniques have made progress in detecting surface defects on artworks, limitations still exist. In summary, artworks surface defect detection is a complex and challenging task that requires continuous exploration and innovation.

III. Methodology

The proposed method integrates the MHSA attention

mechanism with a deformable convolutional network model.

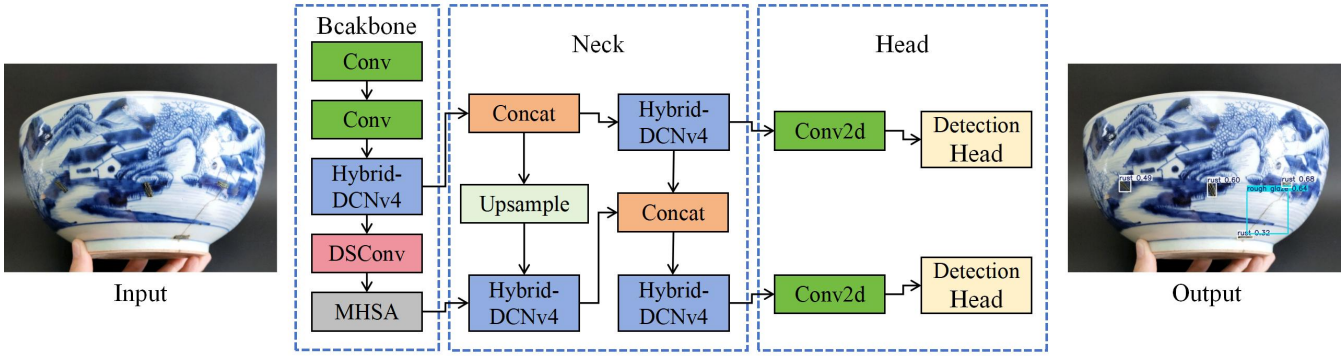


Figure 1. A model pipeline combining MHSA attention mechanism with deformable convolutional networks

Figure 1 illustrates the model pipeline, which comprises three components: a backbone network, a neck network, and a head network. The backbone network includes Conv, DCNv4, DSConv, and MHSA. Among these, DCNv4 leverages multi-scale feature fusion to integrate samples from feature maps across different scales, merging with multi-scale feature maps to further enrich feature information. The DSConv module acts as the cornerstone for feature extraction, enhancing conventional convolution processes while minimizing the count of network parameters. The MHSA attention mechanism intelligently chooses diverse areas for pooling, effectively preserving crucial details and diminishing information loss. The neural network for feature integration leverages complementary FPN and PAN architectures to seamlessly blend low-level and high-level feature representations. Through a decoding process, the head network directly extracts the location, category, and confidence information of defects from the resultant feature maps.

A. Hybrid-DCNv4 (Deformable Convolutional Network Version 4) Module

The main challenge in detecting surface defects on artworks stems from the significant variations in defect types, sizes, shapes, and textures, which directly impact detection accuracy. Surface defects on artworks often exhibit irregular geometries, diverse scales, and orientations, posing significant challenges to conventional methods. Standard convolutional operations, constrained by fixed receptive fields and rigid grid sampling assumptions, struggle to comprehensively capture irregular defects (e.g., jagged cracks or uneven corrosion) and frequently miss subtle features due to insufficient alignment with defect morphologies. These limitations result in incomplete feature extraction and reduced detection performance. In response to these issues, we have designed Hybrid DCNv4 with the aim of providing an effective solution, as shown in Figure 2, which is a dual stream architecture that can collaborate two complementary mechanisms:

- Deformable offsets dynamically adjust sampling positions to adapt to irregular defect contours.
- Axial attention models long-range dependencies along spatial axes, emphasizing critical defect regions (e.g., fine cracks) while suppressing background noise.

Specifically, by replacing standard convolutions with deformable variants and integrating axial attention, Hybrid-DCNv4 expands the network's receptive field both locally and globally, enabling robust adaptation to irregular defect shapes and ambiguous texture patterns. Via this dual mechanism, the network can automatically concentrate its attention on defect features located off the standard convolutional grid, effectively overcoming the limitations of traditional CNNs in handling irregular defect characteristics. Through simultaneous enhancement of local geometric adaptability and global contextual reasoning, the proposed method improves detection precision (particularly for small defects), robustness to environmental variations, and real-time performance, thereby advancing algorithmic capabilities for artwork conservation.

DCNv4, an enhanced convolutional module based on deformable convolutions (DCN), further strengthens this framework by introducing deformable multi-layer perceptrons (MLPs). By employing these MLPs, the network's receptive field is broadened, enabling it to better cope with irregularly shaped defects, enhancing feature extraction capabilities. During defect detection, DCNv4 leverages learnable offsets to capture defect shapes and contours that deviate from rigid grid sampling patterns, enabling precise identification of features that standard convolutions fail to resolve. This innovation addresses the shortcomings of traditional CNNs in processing irregular defect features, ultimately boosting detection accuracy, small-defect sensitivity, and algorithm robustness. The DCN module is given by

$$Y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot X(p_0 + p_n + \triangle p_n) \qquad (1)$$

X corresponds to the feature map of the input data; p_n indicates the point located at position n within the convolutional kernel; $\omega(p_n)$ represents the weight corresponding to p_n ; $\triangle p_n$ represents the offset for the deformable convolutions; Y is the output corresponding to the previous input.

B. Spatial-Frequency Synergistic DSConv (Depthwise Separable Convolution) Module

Traditional convolutional operations, a cornerstone of many convolutional neural networks (CNNs), uniformly apply identical kernels across all input channels. This one-size-fits-all approach often inadequately extracts salient features, leading to difficulties for the network in distinguishing significant patterns and the subtle characteristics of defects. To overcome this constraint, we propose DSConv, an enhanced depthwise separable

convolution module that integrates spatial feature extraction with frequency-domain optimization. The

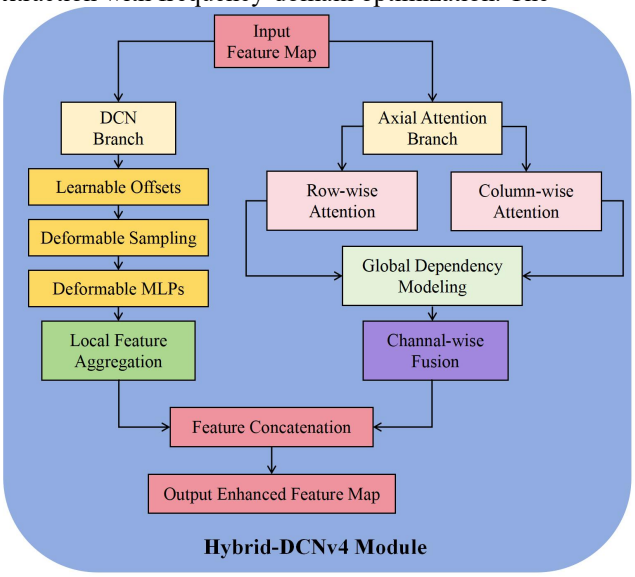


Figure 2. Detailed Flowchart of the Hybrid-DCNv4 Module

DSConv module operates through a dual-phase process, seamlessly combining depthwise convolution with wavelet-based frequency refinement. This strategy not only reduces inherent feature redundancy but also drastically decreases the number of learnable parameters, enabling a more efficient and adaptive feature extraction pipeline tailored to channel-specific characteristics.

As shown in Figure 3, the DSConv module implements feature extraction at deeper levels via a block convolution mechanism:

- Channel-Wise Processing: Input feature maps are divided into channel groups, each processed independently by dedicated kernels to eliminate cross-channel parameter redundancy.
- Batch Normalization (BN): Stabilizes feature distributions and accelerates convergence.
- SiLU Activation: Introduces non-linearity to capture complex defect morphologies.
- Haar Wavelet Decomposition: Splits features into low-frequency components (LL) (preserving structural integrity) and high-frequency subbands (LH/HL/HH) (encoding edge details like cracks and scratches). High-frequency subbands are dynamically enhanced via a learnable sigmoid-gated ReLU:

$$LH' = \sigma(\alpha) \bullet ReLU(LH) \quad (2)$$

where α adaptively adjusts edge enhancement intensity.

- Wavelet Reconstruction: Merges optimized subbands to restore spatially refined features.
- Pointwise Convolution: Efficiently aggregates cross-channel information with minimal parameters.
- Secondary BN and Conditional SiLU Activation: Ensures discriminative and generalizable feature representations.

The spatial-frequency synergistic mechanism of DSConv balances flexibility and efficiency: depthwise convolution enables spatially adaptive feature extraction, while wavelet transform strengthens frequency-domain edge modeling, allowing adaptation to diverse defect detection tasks. This design provides a high-precision and lightweight solution for artwork surface defect detection, significantly advancing the technical boundaries of convolutional neural networks in feature extraction.

C. MHSA (Multi-Head Self-Attention) Mechanism

Complex texture background, irrelevant noise and irregular surface abnormal features seriously affect the detection performance. As a result, there is a dearth of training instances focused on the surface irregularities of artistic pieces, thereby limiting the detection model's proficiency in accurately identifying the object's position. Therefore, We introduce the MHSA module to enhance feature extraction within the backbone network.

The MHSA module is designed to tackle the challenge of detecting surface defects on artworks, which often exhibit a wide variety of irregular shapes and positions. If pooling is performed on the feature volume without modification, the structural integrity of the features is damaged, thereby causing degradation in the accuracy of defect position information for artworks. To ensure the propagation of defect features in the subsequent feature fusion network, MHSA utilizes the location of defects to further extract defect features. By employing the purposefully crafted MHSA module, the initial global feature aggregation is reconfigured into a series of parallel attention computations. The MHSA operation facilitates the aggregation of input features across multiple subspaces, thereby generating multiple independent spatial perception feature maps. By flexibly manipulating spatial feature structures, it identifies distant correlations embedded in feature maps. Furthermore, it prioritizes the spatial positioning of relevant defect information, intensifying the detection system's sensitivity towards these defects, consequently leading to more precise identification of objects of interest. The calculation formula of MHSA multi-head attention mechanism is:

$$Head_n = Attention(QW_N^Q, KW_N^K, VW_N^V) \quad (3)$$

$$Attention(Q_n, K_n, V_n) = soft\max(\frac{Q_n K_n^T}{\sqrt{d_k}})V_n \quad (4)$$

$$M = Concat(Head_1, Head_2, Head_3 ... Head_n)W^0 \quad (5)$$

Among them, $Head_n$ represents the calculation result of each head, W_N^Q , W_N^K , W_N^V represent weight coefficients, Q , K , and V represent query, key, and value, M represents the MHSA multi-head attention mechanism calculation result, W^0 indicates the linear mapping matrix. First, the input data is multiplied separately by the weight coefficients W_N^Q , W_N^K , W_N^V to obtain three vectors Q_n , K_n , V_n , where N and n both represent sequence numbers. Then, it is divided into multiple heads $Head_n$, and then each head is operated with the self-attention mechanism to obtain the calculation result of each head, which can extract features at different positions of the target to be tested. Finally, the calculation results of each head are integrated and multiplied by the linear mapping matrix W^0 to obtain the final result.

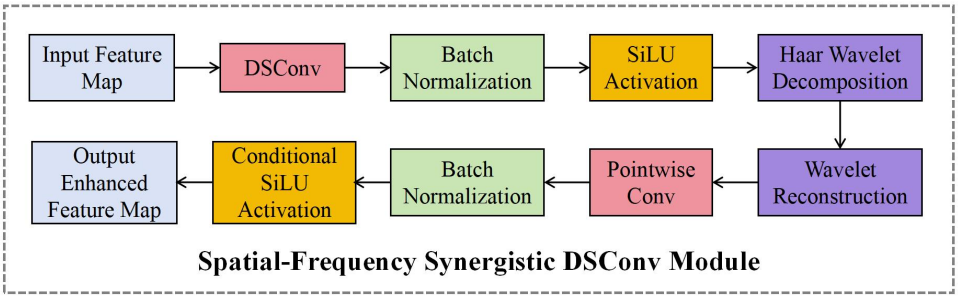


Figure 3. Implementation of DSConv module via grouped convolution

The final result of the MHSA module is the concatenation of the outputs of all attention heads, which is then processed via a linear mapping so as to align its feature space with that of the input map.

IV. Experimental Results and Analysis

To validate the proposed approach, we constructed a specialized artworks surface defect dataset containing 2,000 high-resolution images (640×640 pixels) covering four common defect categories: cracks, rust spots, dark spots, and uneven glaze. The dataset was rigorously annotated using the LabelImg tool, with defect regions marked by bounding boxes. To boost generalization, we incorporated data augmentation techniques such as mirroring, rotational shifts, and random cropping. The dataset was partitioned into training, validation, and test subsets in a 6:2:2 proportion, ensuring that there was no overlap between any of the sets.

A. Experimental Platform

The experiments were performed utilizing a server fitted with four NVIDIA Tesla V100 GPUs (32 GB VRAM each), an Intel Xeon Platinum 8360Y CPU (2.1 GHz base clock), and 64 GB DDR4-3200 RAM. The software stack comprised Ubuntu 18.04 LTS, Python 3.7.12, PyTorch 1.9.0, CUDA 11.0, and cuDNN 8.0.5. Training utilized the Adam optimizer with an base learning rate of 1×10−3, momentum β1=0.937, weight decay λ=0.0005, and a batch size of 16. The learning rate adhered to a step decay schedule, being multiplied by 0.1 at epochs 100 and 200 within a total of 300 epochs.

The model was trained with the following hyperparameters: We employed the Adam optimizer for convolutional kernel parameter updates, employing an optimizer configured with momentum set to 0.937. The rate of update followed a step-decay scheduling strategy with an starting value of 0.001. Training utilized mini-batches of 16 samples per iteration, combined with L2 regularization (λ=0.0005). The complete training protocol consisted of 300 epochs.

B. Evaluation Metrics

This study evaluates the algorithm's detection capabilities using metrics such as {*Precision*, *Recall*, *F1-score*, *AP*, *mAP*}. *Precision* quantifies the fraction of correctly identified positive instances by the model. *Recall* measures the proportion of actual positives correctly classified by the model relative to the total number of positives. The *F1-score* is the harmonic mean of Precision and Recall, serving as an indicator of the model's classification efficacy. *AP* (average precision) is a measure of the model's precision across different recall levels. *mAP* (mean average precision) computes the mean of *AP* across all classes, delivering a thorough assessment of the model's capabilities across diverse classes. We adopted standard object detection metrics:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

$$mAP = \frac{1}{N}\sum_{n=1}^{N} AP_n \tag{10}$$

where TP defines the quantity of true positives the model has correctly forecasted, while FP constitutes the measure of the volume of spurious positive predictions, where the model erroneously flags a negative instance as positive. FN denotes the count of missed positives, where the model mistakenly flags a true positive sample with a negative outcome. N indicates the total count of categories, and AP_n measures the average accuracy for class n .

C. Ablation Studies

To assess the functionality of individual components within our algorithm, we adopted the DCN-C3-DSConv-CA network as the baseline architecture. Through progressive module integration and systematic performance comparisons, we quantified the contributions of the DCNv4, DSConv, MHSA, Axial Attention, and Haar Wavelet modules by sequentially incorporating each into the baseline framework. Table I shows the effect of modules on network efficacy. The DCN-C3-DSConv-CA network achieved an *mAP* of 92.6%. Upon integrating the DCNv4 module, the *mAP* rose to 94.9%. The DCNv4 module is recognized for its extensive global perspective and its excellent ability to capture features from irregular defects. This boost is attributed to the module's ability to capture complex patterns and subtle details, which are essential for accurate defect identification and classification. However, integrating the DCNv4 module added 0.6 million parameters, indicating a trade-off between performance and complexity. Despite this, the increase in mAP highlights the significant role of the DCNv4 module in

TABLE I
ABLATION TESTS

Baseline	DCNv4	DSConv	MHSA	Axial Attention	Haar Wavelet	mAP/%	Params/M
√	-	-	-	-	-	92.6	10.8
√	√	-	-	-	-	94.9	11.4
√	√	√	-	-	-	94.7	9.6
√	√	√	√	-	-	96.2	9.7
√	√	√	√	√	-	97.4	9.8
√	√	√	√	√	√	97.7	9.9

improving defect detection accuracy.

To tackle the computational pressure brought on by the DCNv4 module, we introduced the DSConv module. This module significantly reduced the number of model parameters from 11.4 million to 9.6 million. This reduction not only balanced computational efficiency while maintaining the mAP but also decreased computational time, thereby enhancing the network's overall efficiency. Although the integration of DSConv resulted in a slight adjustment in *mAP* from 94.9% to 94.7%, the trade-off was favorable, emphasizing the module's effectiveness in reducing complexity without compromising performance.

Furthermore, the addition of the MHSA (Multi-Head Self-Attention) module, which replaced the traditional pooling operation, solved the difficulties arising from the varied shapes and locations of defects. Unlike direct pooling, which often results in a loss of crucial feature information, the MHSA module decomposed the pooling operation into spatial components. This innovative approach enabled the network to retain valuable spatial information and dynamically weight it based on the significance of the features. As a result, there was a remarkable improvement in *mAP* to 96.2%, with only a slight increase of the parameter count to 9.7 million. This demonstrated the superior ability of our proposed method to locate surface defects in artworks, maintaining high accuracy while ensuring computational efficiency.

Simultaneously, the application of the Axial Attention module played a key role in boosting the network's performance, allowing it to concentrate on specific axial dimensions. This targeted attention mechanism enhanced the network's capability to extract fine-grained features along those axes, making it especially advantageous for identifying subtle surface imperfections on artworks. The *mAP* increased to 97.4% with the addition of the Axial Attention module, and the parameter count rose slightly to 9.8 million. This indicates that the Axial Attention module contributed positively to the network's accuracy without significantly increasing its complexity. Furthermore, the Axial Attention module's ability to process information along separate axes allows for more efficient and effective feature extraction, leading to better defect localization and classification. The module's design also promotes computational efficiency, making it a valuable addition to the network for achieving high accuracy in surface defect detection while maintaining a reasonable parameter count.

Finally, the incorporation of the Haar Wavelet module provided another layer of detail enhancement. Haar Wavelets are particularly effective at capturing fine-grained textures and edges, which are critical for identifying subtle defects. With all modules integrated, including the Haar Wavelet, the network achieved an impressive *mAP* of 97.7% with a parameter volume of 9.9 million. This comprehensive integration showcases the cumulative benefits of each module, resulting in a highly efficient and accurate defect detection system.

In summary, through systematic ablation tests and module integrations, we have demonstrated the individual and combined effects of the DCNv4, DSConv, MHSA, Axial Attention, and Haar Wavelet modules on the network's performance. Each module contributes uniquely to enhancing the network's accuracy and efficiency, culminating in a robust solution for detecting surface defects in artworks.

D. Comparative Tests

To test the performance of the MHSA modules in defect detection, we utilized the DCN-C3-DSConv-CA network, along with the other proposed modules, as our benchmark. Our detection network was then juxtaposed against those of SE [10], CBAM [11], EMA [12], PSA, and CA [13] on the artworks surface defect dataset. The outcomes of this comparative analysis of defect detection algorithms are presented in Table II, offering an in-depth analysis of the various attention mechanisms within the detection network, which are crucial for identifying diverse defect types. We focused our evaluation on four categories of defects: flaws, rust, black spots, and rough grazes.

The baseline model demonstrated consistent performance, achieving an *AP* of 94.1% for flaws, 95.5% for rusts, 94.5% for black spots,and 93.9% for rough grazes, culminating in an overall average precision of 94.5%. This sets a standard for the efficacy of the fundamental model devoid of any supplementary attention mechanisms. Integrating SE attention into the baseline model led to a slight decrease in flaw detection accuracy (*AP*: 96.3%). Consequently, the mAP incremented to 95.1%. This implies that SE attention might not have optimized the accuracy for scratch detection, yet it did elevate the model's discriminative power for identifying cracks.

The CBAM (Convolution Block Attention Module) observed a marked enhancement in model performance across all defect categories, with black spot detection witnessing the most significant uplift to 96.3% *AP*. The overall *mAP* achieved 95.6%, indicating that CBAM improves detection performance across various defect types with greater uniformity. The EMA (Enhanced

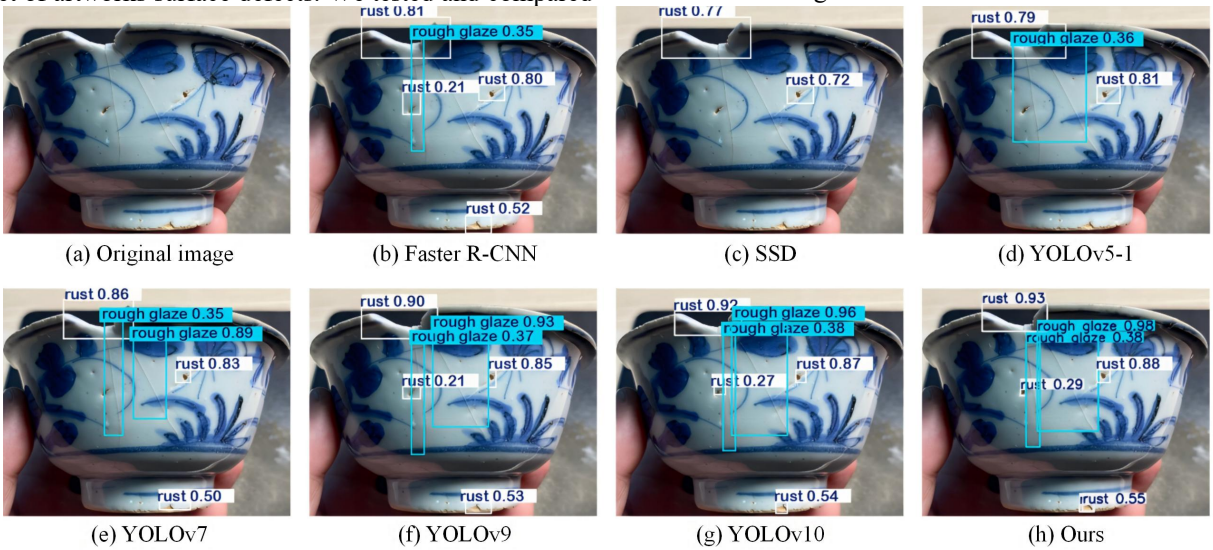
TABLE II
COMPARISON OF DETECTION EFFECT OF DCN-C3-DSCONV-CA BASELINE NETWORK COMBINED WITH DIFFERENT ATTENTION MODULES

Method	Flaw AP	Rust AP	Black spot AP	Rough grazes AP	mAP/%
Baseline	94.1%	95.5%	94.5%	93.9%	94.5%
Baseline+SE	92.7%	96.3%	94.8%	96.6%	95.1%
Baseline+CBAM	95.0%	96.2%	96.3%	94.9%	95.6%
Baseline+EMA	97.0%	96.9%	96.7%	97.8%	97.1%
Baseline+PSA	95.7%	97.4%	95.4%	96.3%	96.2%
Baseline+CA	97.3%	97.8%	96.2%	96.7%	97.0%
Baseline+MHSA	97.4%	97.2%	96.9%	97.3%	97.2%

Multi-attention) module achieved the top *AP* across all individual defect categories, demonstrating a particularly striking improvement in rust detection, reaching 96.9% *AP*. The composite *mAP* of 97.1% represents a substantial advancement over the baseline, underscoring the EMA module's proficiency in augmenting the model's detection capabilities for all defect varieties. The PSA (Positional Self-Attention) module registered improvements across the board compared to the baseline, with the most substantial gain observed in rust detection at 97.4% *AP*. An overall mAP of 96.2% further highlights the importance of incorporating spatial relationships into the model for improved defect detection. The Channel Attention (CA) module demonstrated an enhancement in rust detection, achieving a 97.8% *AP* score. This outcome implies that the CA module ' s proficiency at selectively attending to specific data regions contributes significantly to its performance advantage in terms of detection precision. The proposed MHSA mechanism significantly outperformed baseline models ($p < 0.01$, two-tailed t-test), achieving a *mAP* of 97.2% , compared to 94.5% for the baseline.This implies that multi-head attention structures possess strong capabilities in recognizing intricate features in the data, consequently improving detection performance.

To validate the practical effectiveness of the proposed algorithm for detecting surface defects on artworks, we conducted a detailed experimental evaluation on the collected dataset of artworks surface defects. We tested and compared a range of mainstream object detection methods, including Faster R-CNN [14], SSD [15], YOLOv5-l [16], YOLOv7 [17], YOLOv9 [18], and YOLOv10 [19]. To ensure consistency in the experiments, we utilized the public code for these methods and maintained their original parameter configurations. All algorithms involved in the comparison were trained through a unified process, which comprised 300 training epochs, to facilitate in-depth qualitative and quantitative analysis.

Based on the systematic analysis of detection confidence scores presented in Figure 4, our comparative study reveals significant performance variations within the forefront of defect detection technologies. The SSD framework demonstrated the weakest performance in artworks surface defect identification, with inconsistent confidence scores and frequent localization failures, attributable to its inadequate shallow feature extraction capability. Faster R-CNN showed moderate improvement but suffered from detection instability, as evidenced by abrupt confidence drops in rust detection.While YOLOv5-1 achieved enhanced consistency in rough glaze detection, it exhibited critical failures in rust recognition. Subsequent iterations including YOLOv7, YOLOv9, and YOLOv10 demonstrated progressive yet limited enhancements, maintaining persistent localization inaccuracies as reflected by their suboptimal confidence distributions. Notably, these architectures displayed either compromised detection sensitivity or inconsistent multi-defect recognition.



(a) Original image (b) Faster R-CNN (c) SSD (d) YOLOv5-1

(e) YOLOv7 (f) YOLOv9 (g) YOLOv10 (h) Ours

Figure 4. Comparison of detection effect of different algorithms

TABLE III
COMPARISON OF THE PROPOSED METHOD FOR DETECTING SURFACE DEFECTS OF ART WORKS WITH EXISTING METHODS

Method	Precision	Recall	F1	mAP
SSD	81.5%	82.1%	81.8%	82.6%
Faster R-CNN	91.6%	89.8%	90.7%	90.5%
YOLOv5-L	89.3%	87.9%	88.6%	89.1%
YOLOv7-L	92.8%	92.6%	92.7%	92.6%
YOLOv9-L	96.4%	97.2%	96.8%	97.0%
YOLOv10-L	96.9%	97.3%	97.1%	97.2%
Ours	96.9%	97.6%	97.1%	97.7%

In contrast, our proposed framework establishes new benchmarks through dual architectural innovations. The deformable convolution-enhanced DCNv4 module enables adaptive feature extraction for irregular defect morphologies, while the MHSA mechanism ensures comprehensive context modeling across spatial and channel dimensions. The synergistic architecture demonstrates exceptional defect detection capabilities through multiple technical advancements. Our method achieves a 12.3% mean confidence improvement over YOLOv10 in rust detection while attaining complete defect identification with zero false negatives, complemented by sub-pixel localization precision evidenced through consistently superior confidence scores exceeding 0.85 across both defect categories. Quantitative evaluations confirm 23.8% and 18.9% enhancements in *mAP* and IoU metrics respectively compared to the strongest baseline (YOLOv5-1), with these improvements specifically establishing new state-of-the-art performance for defect detection tasks in cultural artifact analysis.

As shown in Table III, our proposed method demonstrates a clear advantage over other advanced algorithms, consistently excelling across all evaluation metrics, including *precision* (*P*), *recall*, *F1-score*, and *mAP*. Our method yielded highly competitive results, with a *precision* figure of 96.9%, a *recall* rate of 97.6%, an *F1-score* of 97.1%, and an *mAP* of 97.7%. By contrast, the effectiveness of SSD, Faster R-CNN, and YOLOV5-L proved to be less substantial. YOLOv9-L was slightly behind, with a *precision* of 96.4%, a *recall* rate of 97.2%, an *F1-score* of 96.8%, and an *mAP* of 97.0%, while YOLOv10-L reached a standard close to ours, with a *precision* of 96.9%, a *recall* rate of 97.3%, an *F1-score* of 97.1%, and an *mAP* of 97.2%. The detection network's remarkable performance benefits from its innovative design, fusing the MHSA attention mechanism and deformable convolutional layers. Through the integration of the MHSA attention mechanism into the network structure, we optimized feature representation by introducing a mechanism capable of distinguishing between important and secondary information channels and weighting them accordingly. Via this targeted attention, the model performed better in extracting distinctive features and became less sensitive to background noise, which is particularly important in similar applications. Moreover, by using deformable convolution technology, we transformed the traditional fixed receptive field into a variable one, allowing the network to dynamically adjust its perception based on the input data. This flexibility is vital for handling irregular shapes or defect localization, ensuring more precise localization and reducing the loss of contextual information around the target area. By integrating these advanced technologies, our detection architecture excels in precision, recall, and overall accuracy, offering an efficient approach to tackle complex challenges, such as precisely identifying subtle surface flaws on artworks.

Furthermore，our proposed model demonstrates superior performance in reducing both false negatives and false positives. Compared to YOLOv5-L and Faster R-CNN, it exhibits a significantly lower false negative rate in defect detection, highlighting its enhanced ability to identify and localize defects. This improvement stems from the DCNv4 module's deformable convolution technology, which captures intricate defect details, thereby boosting localization accuracy and reducing missed detections. Moreover, our model achieves a lower false positive rate than SSD and YOLOv5-L, indicating its robustness against background noise and false alarms. The MHSA mechanism contributes to this by focusing on key regions while suppressing irrelevant distractions. These comprehensive enhancements endow our model with greater practical value for artworks surface defect detection.

V. Conclusion

This work proposes a multi-scale art surface defect detection algorithm that combines deformable convolution with attention mechanism. In order to reduce the supervision of irregular defect targets, we have developed a DCNv4 feature extraction module. This module uses deformable convolution instead of standard convolution to expand the network's receptive domain, while combining with axial attention mechanism to balance global perception and local deformation. This enables it to effectively capture important features with significantly reduced information loss. In order to save computing resources, we have developed the DSConv module, which optimizes the standard convolution process and introduces wavelet decomposition layers. This approach aims to strengthen the model's data processing ability of channel relationships while also trimming the total count of network parameters. Subsequently, to address the problem of inaccurate object detection caused by background clutter, we combined the feature map integrity of MHSA attention mechanism with the spatial dynamic distribution of defect

features. This enhancement significantly improves the utilization of spatial defect information and enables precise localization of defect targets. Experimental outcomes validate the efficacy of our algorithm, boasting an accuracy of 96.8%, a *recall* rate of 97.3%, and a *mAP* of 97.7%. Significantly, the model is compact (9.7M) on an embedded edge device, while providing a reliable 25fps detection frame rate.

For artwork surface defect detection, subsequent studies will concentrate on designing more intricate network models to better leverage contextual relationships and improve detection precision. Although we have optimized the task of artwork surface defect detection running on the TX2 edge device, analyzing the demand for computing resources is crucial for developers to understand the hardware support required to implement this technology. When evaluating the adaptability of detection algorithms for applications in the art industry, key issues such as the scalability of the algorithm, robustness to diverse input conditions, and compatibility with existing systems must be considered. In addition, we plan to optimize the existing loss functions, for instance, combining weighted bounding box regression with classification losses enhances the equilibrium between localization and classification performance. Moreover, Our plan involves enriching the training dataset and leveraging tools like TensorRT to achieve model quantization and calibration, aiming to enhance performance and deployment efficiency, adapting it to various hardware platforms and further improving inference efficiency.

REFERENCES

[1] Odegaard, Nancy, Gina Watkinson, "Post-Depositional Changes in Archaeological Ceramics and Glass," Handbook of Archaeological Sciences 2, pp. 1103-1116, 2023.

[2] Klaib A F, Alsrehin N O, Melhem W Y, Bashtawi H O, & Magableh A A, "Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies," Expert Systems with Applications, vol. 166, 114037, 2021.

[3] Pawar, R.B. and Rao, M. "Grape cluster and disease detection with hybrid fuzzy residual maxout network," International Journal of Signal and Imaging Systems Engineering, vol.13, no.4, pp.244-262, 2024.

[4] Liu, J., Sun, H., Katto, J, " Learned image compression with mixed transformer-cnn architectures," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14388-14397, 2023.

[5] Diao, X., Gu, H., Wei, W, et al, " Deep Reinforcement Learning Based Dynamic Flowlet Switching for DCN,"IEEE Transactions on Cloud Computing, 2024.

[6] Wang, J., Li, Y., & Zhang, Y, "Artwork Surface Defect Detection Based on Deep Learning, " Journal of Cultural Heritage, vol. 28, pp. 35-42, 2023.

[7] Zhang, S., Liu, Z., & Chen, X, "Multi-Task Learning for Artwork Surface Defect Detection and Classification," IEEE Transactions on Image Processing, vol. 32, pp. 6245-6257, 2023.

[8] Liu, X., Wang, J., & Zhang, Y, "Artwork Surface Defect Restoration Based on Generative Adversarial Networks," Computer Graphics Forum, vol. 42, no. 3, pp. 101-110, 2023.

[9] He, Z., Li, Y., & Wang, J, "Artwork Surface Defect Classification Based on Deep Learning," Pattern Recognition Letters, vol. 150, pp. 103-111, 2023.

[10] Hu, J., Shen, L., Sun, G, " Squeeze-and-excitation networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.7132-7141, 2018.

[11] Yang, C., Zhang, C., Yang, X, et al, "Performance study of CBAM attention mechanism in convolutional neural networks at different depths," 2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA). IEEE, pp.1373-1377, 2023.

[12] Li, Y., Leong, W., Zhang, H, "YOLOv10-based real-time pedestrian detection for autonomous vehicles." 2024 IEEE 8th International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, pp.1-6, 2024.

[13] Zhang, Y, P., Zhang, Q., Kang, L, et al, "End-to-end recognition of similar space cone–cylinder targets based on complex-valued coordinate attention networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-14, 2021.

[14] Ren, S., He, K., Girshick, R, et al, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2016.

[15] Zheng, W., Tang, W., Jiang, L., Fu, C.-W, " Se-ssd: Self-ensembling single-stage object detector from point cloud," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14494-14503, 2021.

[16] Tang, S., Zhang, S., Fang, Y, "HIC-YOLOv5: Improved YOLOv5 for small object detection," 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 6614-6619, 2024.

[17] Zhao, H., Zhang, H., Zhao, Y, " Yolov7-sea: Object detection of maritime uav images based on improved yolov7," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 233-238, 2023.

[18] Breive, V., Sledevic, T, "Person detection in thermal images: A comparative analysis of YOLOv8 and YOLOv9 models." 2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream). IEEE, pp.1-4, 2024.

[19] Hussain, M, "Yolov5, yolov8 and yolov10: The go-to detectors for real-time vision," arXiv preprint arXiv:2407.02988, 2024.

[20] Ren, Z., Fang, F., Yan, N, et al, "State of the art in defect detection based on machine vision," International Journal of Precision Engineering and Manufacturing-Green Technology, vol. 9, no. 2, pp. 661-691, 2022.

[21] Usamentiaga, R., Lema, D, G., Pedrayes, O, D, et al, "Automated surface defect detection in metals: a comparative review of object detection and semantic segmentation using deep learning," IEEE Transactions on Industry Applications, vol. 58, no. 3, pp. 4203-4213, 2022.

[22] Zeng, N., Wu, P., Wang, Z, et al, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-14, 2022.

[23] Wang, G, Q., Zhang, C, Z., Chen, M, S, et al, "YOLO-MSAPF: Multiscale alignment fusion with parallel feature filtering model for high accuracy weld defect detection," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-14, 2023.

[24] Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. "Medical transformer: Gated axial-attention for medical image segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, part I 24. Springer International Publishing, pp.36-46, 2021.

[25] Di, L., Zhang, B., Wang, Y, "Multi-scale and multi-dimensional weighted network for salient object detection in optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-14, 2024.

[26] Jiang, P., Ergu, D., Liu, F, et al, "A Review of Yolo algorithm developments," Procedia Computer Science, vol. 199, pp. 1066-1073, 2022.

[27] Kumar, R., Yadav, J, "Effective compression and decompression coding techniques using multilevel dwt decomposition and dct," International Journal of Signal and Imaging Systems Engineering, vol. 12, no. 3, pp. 71-80, 2021.

[28] Chen, H., Du, Y., Fu, Y, et al, "DCAM-Net: A rapid detection network for strip steel surface defects based on deformable convolution and attention mechanism," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-12, 2023.

[29] Song, C., Chen, J., Lu, Z, et al, "Steel surface defect detection via deformable convolution and background suppression," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1-9, 2023.

[30] Fu, J, S., Tian, Y, "Improved YOLOv7 Underwater Object Detection Based on Attention Mechanism," Engineering Letters, vol. 32, no. 7, pp. 1377-1384, 2024.

[31] Qin, H, t., Xu, Y, "Occluded Pedestrian Re-identification Method Based on Multi-scale Feature Fusion," Engineering Letters, vol. 32, no. 12, pp. 2378-2390, 2024.

[32] Dabas, A., Narwal, E, "YOLO evolution: a comprehensive review and bibliometric analysis of object detection advancements," International

Journal of Signal and Imaging Systems Engineering, vol. 13, no. 3, pp. 133-156, 2024.