



ASD-YOLO: An aircraft surface defects detection method using deformable convolution and attention mechanism

Bin Huang^a, Yan Ding^a, Guoliang Liu^{a,*}, Guohui Tian^a, Shanmei Wang^b

^a School of Control Science and Engineering, Shandong University, Jinan 250061, China

^b Department of Intelligent Harbor & Navigation, Inspur Intelligence Technology Co.,LTD, Jinan 250061, China

ARTICLE INFO

Keywords:

Aircraft surface defects detection

Deep learning

YOLO

Flight safety

ABSTRACT

Aircraft surface defect (ASD) detection is crucial for ensuring flight safety. Addressing challenges such as large-scale variations, irregular shapes, and sample imbalance in ASD detection, this paper proposes a ASD-YOLO network based on YOLOv5. ASD-YOLO incorporates several enhancements to improve recognition capabilities. Firstly, a new deformable convolutional feature extraction module (DCNC3) is designed to better learn defects of different shapes, which is combined with a global attention mechanism (GAM) to pay more attention to defects region information. Secondly, the feature representation of small defects is bolstered by the contextual enhancement module (CEM). Lastly, to alleviate sample imbalance problem, we introduce an exponential sliding average weight function (EMA-Slide). Experimental results on two datasets show improvements in mean Average Precision (mAP) by 5.7% and 3.4%, respectively, surpassing mainstream algorithms and offering a novel approach to ASD detection.

1. Introduction

As aviation technology advances, aircraft are increasingly utilized in both military and civilian domains. The external structures and components of aircraft are essential for ensuring safe, stable, and efficient operations. Over time, influenced by factors such as flight conditions and human interaction, the outer surface is prone to defects like cracks and corrosion. Such defects not only threaten flight safety but also pose a risk of catastrophic consequences. Therefore, regular surface inspection of aircraft is an important means of maintaining the safety and a necessary part of guaranteeing the stability of aircraft operations. Early aircraft structural integrity inspections primarily relied on sensory methods, including visual and percussion detection. The visual approach, the most basic and common, required aircraft crews to identify defects through naked-eye observation, sometimes assisted by magnifying glasses and other assistive tools. Percussion detection involved tapping the aircraft's surface manually or with special tools to analyze changes in the echoes, thereby determining the location and size of defects. Although sensory inspection method are intuitive, easy to operate, and cost-effective, they have several drawbacks: they are time-consuming, expose crews to the risks of working at high altitudes, and depend heavily on the inspectors' expertise, making results susceptible to subjective judgments.

When the traditional manual inspection methods failed to meet the demands of modern aircraft surface inspections, nondestructive

automatic inspection emerged [1]. These technologies evolved from initial methods such as penetrant and magnetic particle testing to more advanced techniques including ultrasonic, radiographic, eddy current, acoustic emission, infrared thermography, laser holography, and microwave testing [2–4]. Although automated nondestructive testing improves accuracy using specialized equipment, it still often relies on manual interpretation of results, which can be inefficient. Furthermore, many techniques require specific conditions to maintain accuracy. For example, eddy current testing must avoid strong electromagnetic fields [5], and infrared thermography must consider thermal conductivity.

In recent years, driven by the rapid advancements in information science and artificial intelligence, computer vision nondestructive testing technology has undergone significant evolution. This technology leverages image-sensing devices like cameras and utilizes advanced image detection algorithms to quickly and accurately identify defects in images. Automatic nondestructive testing technology based on computer vision offers advantages, including real-time online operation, non-contact methods, and high detection accuracy, which are crucial in today's intelligent industrial landscape. Moreover, machine learning and deep learning are increasingly being applied in aircraft visual inspection [6], especially the deep learning-based ASD detection system represented by convolutional neural network (CNN) has become a major application of computer vision in the nondestructive testing field.

* Corresponding author. Correspondence to: Service Robot Laboratory, Shandong University, Innovation Building, 17923 Jingshi Road, Jinan, China.
E-mail address: liuguoliang@sdu.edu.cn (G. Liu).

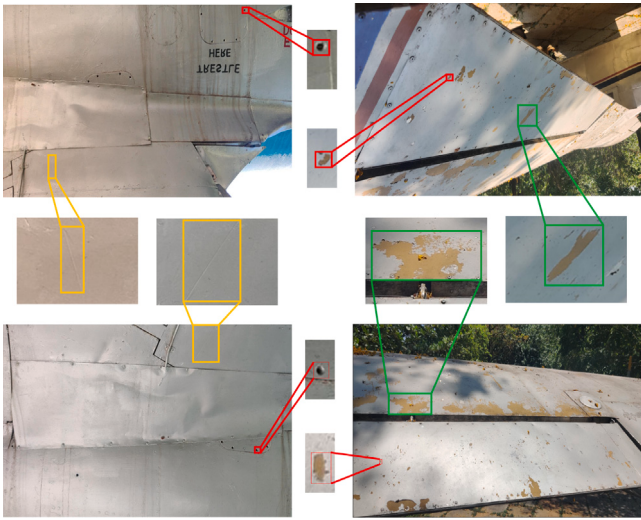


Fig. 1. Characteristics of aircraft surface defects. The defects are visually categorized as follows: Red boxes highlight tiny targets, orange boxes indicate defects that lack clear contrast with the background, and green boxes represent defects with significant intra-class variations such as differences in size and shape.

The ASD detection system typically comprises two main parts: hardware platform and software algorithm. Wall-climbing robots, AGV (Automated Guided Vehicle) and UAVs (Unmanned Aerial Vehicles) are commonly utilized as mobility platforms, carrying computer vision devices to capture clear images of aircraft surface [7–10]. In contrast, software algorithm is the key to decide the performance of inspection system. Currently, computer vision-based defects detection algorithms are categorized into two main types: traditional image processing methods [11,12] and deep learning methods [13–16]. The former mostly manually design and extract features such as texture, shape and color of the image through techniques like filtering, histogram analysis, edge detection. While these methods can be effective in specific contexts, they generally lack flexibility and require manual adjustment of parameters when environmental conditions change. On the other hand, the latter adopt deep neural networks to realize end-to-end learning, automatically extracting multi-layer features at different levels of abstraction from the data. This eliminates the need for manual feature extraction and offers greater adaptability to complex and variable defect scenarios.

Furthermore, ASD detection is a challenging problem. From the perspective of the characteristics of aircraft defects, this is illustrated in Fig. 1. Firstly, the majority of defects are tiny and occupy only a small portion of the entire image, which make feature extraction relatively difficult. Additionally, some defects lack distinct contrast with the background, rendering them nearly indistinguishable. Furthermore, there are notable inter-class disparities among different types of defects, as well as intra-class variations within the same defect category, which include differences in size, shape, and other characteristics. Finally, aircraft defect datasets often suffer from imbalance with significant variation in the quantity of samples across categories.

Consequently, given the characteristics of aircraft surface defects, this paper explores an efficient and accurate deep learning-based method for precise ASD detection. The process of proposed method is illustrated in Fig. 2, encompassing data acquisition, model training, and application. Specifically, we improve the construction of the network architecture from four perspectives: deformable convolution, attention mechanism, context enhancement, and loss function. These modules synergistically interact to significantly enhance the network's ability to recognize defects.

The main contributions encompass the following points:

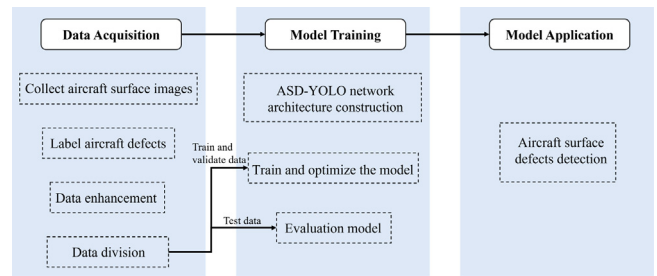


Fig. 2. The algorithm flow chart.

- Aiming at the poor detection ability of the original network framework for small and irregular defects, we propose a novel ASD detection network: ASD-YOLO. This network not only enhances the accuracy of detecting irregularly shaped objects but also selectively learns the more important defect features and considers both global and local information of targets of varying sizes, thereby effectively improving the overall effectiveness of defect detection.
- To alleviate the performance degradation caused by sample imbalance between simple and difficult samples, we design a slide weight function based on exponential sliding average applied to the loss function. This function adaptively smooths the weight assignments for samples near the ambiguous classification boundaries during the training process. By dynamically adjusting through training, it increases the network's focus on difficult samples that are harder to recognize, effectively alleviating the sample imbalance issue.
- To analyze and evaluate the performance of ASD-YOLO, we conduct a series of experiments using publicly available datasets and our own constructed datasets to verify the effectiveness and superiority of the proposed method.

This paper is organized as follows: Section 1 provides an overview of the research development history and background of aircraft surface defects detection. Section 2 briefly introduces related computer vision-based ASD detection methods. Section 3 describes in detail the proposed ASD-YOLO network architecture. Section 4 presents the experimental validation and results. Section 5 discusses and analyzes the performance and feasibility of the proposed method. Finally, Section 6 summarizes the whole paper.

2. Related work

Due to the superior performance of deep learning models, deep learning-based methods have progressively become the mainstream approach in defects detection. Intelligent ASD detection, as an important area of study, has also attracted many scholars to engage in research. Despite the relatively limited volume of literature in this direction, some research findings have been produced in the classification, detection, and segmentation of aircraft surface defects.

Alberts et al. [17] developed an improved neural network algorithm based on the CIMP robotic system, which enhances the remote detection of normal and cracked areas on aircraft skin. However, this early work struggles with detection in complex environments and has a limited scope of application. Cha et al. [18] implemented an approach combining sliding windows with convolutional neural networks (CNNs) to identify and localize surface defects. While effective in localizing defects, this approach suffers from slow detection speeds and low localization accuracy. Li et al. [19] constructed a lightweight network for aircraft structural crack detection using CNNs, which is capable of detecting various parts and types of aircraft. Nevertheless, the detection accuracy of this network remains relatively low and

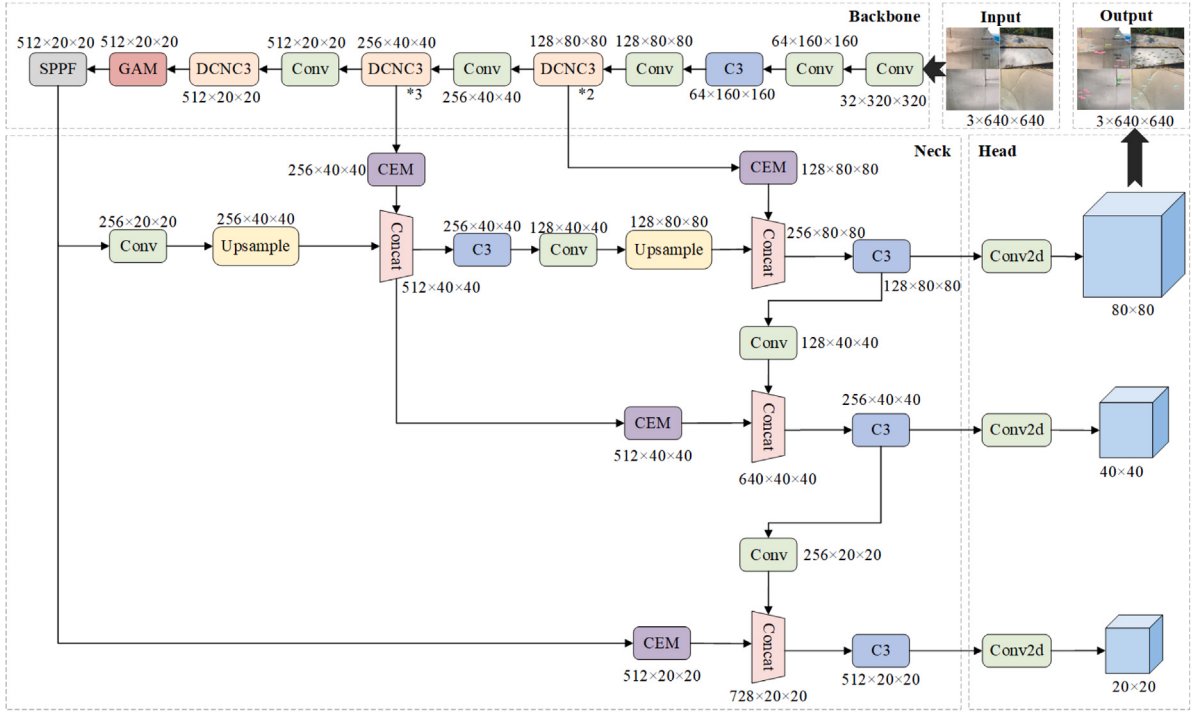


Fig. 3. Overall network architecture. It consists of three parts: the backbone, neck and head. In the backbone, the original C3 module has been replaced with the DCNC3 module, which better enables the network to learn the feature information of irregular defects. Additionally, GAM is integrated to selectively focus on defect features. In the neck, CEM is incorporated to reduce the loss of information from minor defects, thereby enhancing feature representation. Furthermore, the loss function has been refined by utilizing an exponential sliding weight function to improve sample imbalance issue among defects.

requires improvement. Malekzadeh et al. [20], pioneers in using deep neural networks (DNNs) for aircraft defect detection, utilized the SURF keypoint detector and achieved optimal performance with the VGG-F feature extractor and a linear SVM classifier. It is necessary to note that their dataset images were acquired strictly from a straight view, limiting the diversity of detection viewpoints.

Ramalingam et al. [21] enhanced detection of low-contrast stains and flaws on aircraft surfaces by combining SSD and MobileNet algorithms with a reconfigurable absorption robot. They also incorporated a self-filtering based periodic pattern detection filter at the image preprocessing stage to minimize background noise and enhance target information. However, this significantly increased the complexity and time cost of implementation. Zhong [22] introduced an attention module to the YOLOX algorithm for detecting four types of skin defects. This approach performs well with single targets but struggles with misses or overlaps when multiple defects are present on the same image. Hu [23] developed a machine vision system specifically for accurately measuring dimensions and identifying defects on aircraft rivets by analyzing rivet morphology and surface characteristics. This system, however, is limited to specific types of rivets and lacks versatility, operating at a slower speed. Chen [24] constructed a multi-factor comparison experiment using YOLOv3 and Faster R-CNN to evaluate their detection performance in different situations. The experimental results revealed a tendency in defect detection that can be attributed to the models' failure to account for the unequal distribution of sample types and the presence of overlapping defects. Wei and Su [25] tackled defect detection in riveted assembly of aircraft skins using reinforcement learning to design a stochastic data enhancement strategy, which improves the generalization ability at the cost of increased preprocessing time. Zhang [26] proposes a multi-level feature fusion network (DaMFFN) based on a dual-structure attention mechanism for automatic surface defect detection, with limitations including increased computational demand and inadequacy for industrial inspection requirements. Li et al. [27] mainly proposes a defect detection network

specifically for aircraft skin fasteners. It introduces an attention fusion feature pyramid network to solve the challenge of small and fuzzy fasteners defects, and improves the detection of small defects but lacks the consideration for the detection of various defect types. Wang et al. [28] proposed an improved YOLOv8n model for ASD detection that enhances feature fusion, the regression loss function, and incorporates multi-scale feature fusion. However, it did not adequately address the challenges of detecting small targets and classifying difficult samples in specialized scenarios.

With the rapid development of more advanced and novel architectures, segmentation methods for defects regions have emerged. Bouarfa et al. [29] utilized Mask R-CNN with transfer learning to automate the detection of dents in aircraft structures. Building on this foundation, Doğru et al. [30] extended the previous research by integrating various data processing and enhancement techniques to significantly improve the accuracy of dent detection. Ding et al. [31] proposed a pixel-level defect visual detection method on the basis of Mask Scoring R-CNN, which introduces an attention mechanism and a feature fusion module to enhance feature representation, and designs a novel classifier head aimed at alleviating the influence of surrounding information on defective regions. These methods not only enable the classification and localization of different defects but also segment each pixel into different categories to obtain the shape and size of the defects, which is particularly useful when precise contour determination is needed.

Existing studies reveal that target detection algorithms like SSD and YOLO, along with the target segmentation algorithm represented by Mask R-CNN, are commonly employed for ASD detection. While segmentation algorithms can delineate precise contours of defects, detection algorithms are typically faster and less resource-intensive. For most aircraft surface inspection tasks, identifying the location and approximate range of defects sufficiently meets operational requirements. Moreover, most of the above methods use defects images with clean backgrounds and large defects targets, deviating from the practical acquisition viewpoints in real-world applications. They also fail to take into account the multiple scale variations of the targets and the

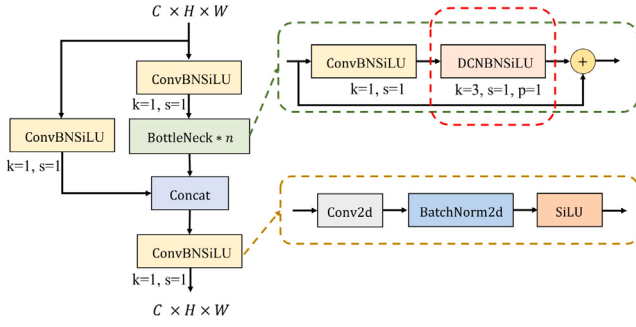


Fig. 4. The DCNC3 module structure, which integrates a deformable convolution in place of one of the standard convolutional blocks in the C3 module (the red dotted box), is specifically designed to enhance the network's accuracy in detecting objects with significant shape variations.

detection of tiny targets at different altitudes. Considering the need for real-time performance and accuracy, this paper adopts YOLO as the basic framework and develops ASD-YOLO, a more efficient and accurate detection algorithm for the small, multi-scale and irregular targets in the ASD detection. This approach is more suitable for the complexity and variability of the aircraft inspection application environment.

3. Method

3.1. Architecture

To enhance the detection performance of the model and better identify defects across various scales and shapes, this paper proposes three structural improvements for the network and one improvement for the loss function. The overall network structure of ASD-YOLO is depicted in Fig. 3, comprising three basic elements: Backbone, Neck, and Head. Within the backbone, a deformable convolution module and an attention module are integrated to enhance feature extraction capabilities. In the neck, a context enhancement module is incorporated to further refine and strengthen the feature fusion process. The input RGB image is first processed through the backbone to extract multi-scale feature information, which is then further fused and amplified in the neck. Finally, the processed three-scale feature maps are used in the detection head for multi-scale target detection.

3.2. Deformable convolution

The C3 module of YOLOv5 uses a standard convolution module to increase the network's depth and receptive field. Although effective for most tasks, it encounters challenges when handling targets that exhibit obvious deformation, such as cracks, scratches, or those with considerable aspect ratio differences in aircraft surface defects. This limitation arises from the fixed shapes of geometric structures in the standard convolutional module, inherently restricting its ability to model geometric transformations. To overcome this limitation and better accommodate geometric deformations, this paper proposes Deformable Convolutional Networks (DCN v2) [32] applied to the feature extraction backbone. Specifically, one of the standard convolutional blocks (Conv) in the C3 module is replaced with a DCN block to form a residual structure formed by sequentially connected ConvBNSiLU and DCNBNSiLU, aiming to enhance the network's capability to detect objects with diverse shapes. The newly adapted C3 module, renamed as the DCNC3 module, is detailed in Fig. 4.

DCN v2 represents an improved convolution operation that dynamically learns the shapes of different convolution kernels based on input data. As illustrated in Fig. 5, DCN v2 introduces 2D offsets to the regular mesh sampling locations in the standard convolution, allowing

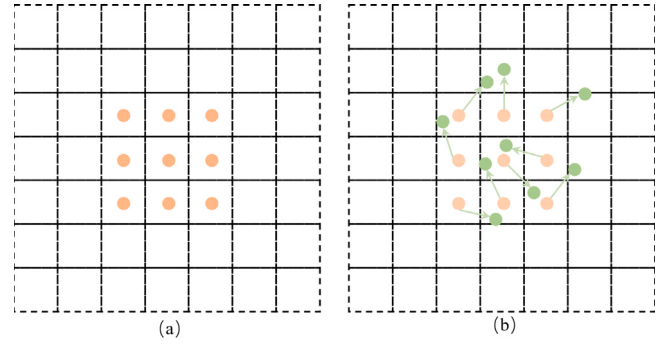


Fig. 5. Deformable convolution kernel. (a) Standard 3×3 convolution kernel, where orange points are standard sampling points; (b) Deformable convolution kernel formed by adding offset to normal convolution, where green points are deformed sampling points, green arrow is the offset direction.

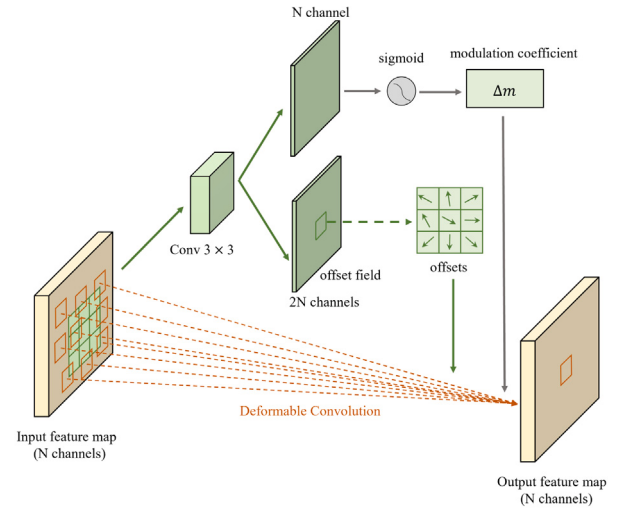


Fig. 6. The deformable convolution process. It obtains offsets and corresponding feature gain coefficients through convolution layers.

the sampled mesh to deform freely, as well as adding feature gain coefficients to control the effective receptive field.

The DCN learns offsets from the previous feature maps, mainly through additional convolutional layers, as illustrated in Fig. 6. Initially, the input feature layer, containing $3N$ channels (where N is the number of channels in the original input feature map), undergoes a standard 3×3 convolution. Here, $2N$ channels are dedicated to calculating XY direction offsets, forming the offset domain, which maintains the same spatial resolution as the input feature map. The outputs from the remaining N channels are processed through a sigmoid layer, which maps their values to a range between 0 and 1, thus generating adjustable gain coefficients. Once the deformed sampling points are determined, the corresponding weights and feature gain coefficients can be dot-multiplied and then summed to obtain the final value of the point on the output feature map, which is as follows:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \cdot \Delta m_n \quad (1)$$

where y denotes the output feature map, p_0 denotes each position on y , R denotes a regular grid over the input feature map x , w stands for weight, $p_n (n = 1, \dots, N, N = |R|)$ enumerates positions in R , representing the relative position of p_0 within the range of the convolution kernel, Δp_n denotes the self-learning offset, typically a small number which is resolved through bilinear interpolation (refer to Eq. (2)), Δm_n denotes the adjustable feature gain of the sampled positions. When the

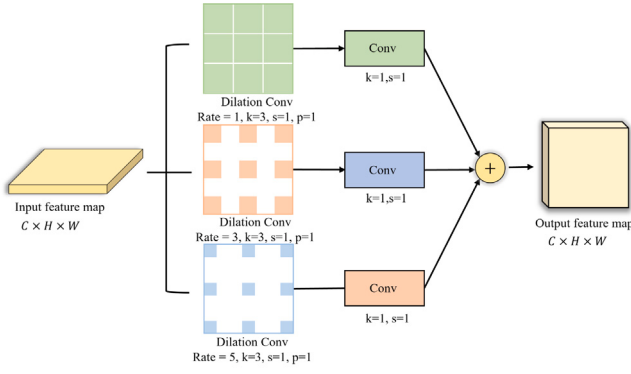


Fig. 7. The CEM module structure. Dilation convolution with three dilation rates is used to obtain contextual information for different receptive fields.

value of Δm_n is 0 it signifies that the corresponding region feature has no impact on the output.

$$x(p) = \sum_q \max(0, 1 - |q_x - p_x|) \cdot \max(0, 1 - |q_y - p_y|) \cdot x(q) \quad (2)$$

where x is the input feature map, $p = p_0 + p_n + \Delta p_n$, q enumerates all integral space locations in x , and q_x, p_x, q_y, p_y denotes two-dimensional interpolation.

3.3. Context enhancement module

The original YOLOv5 network is difficult to recognize the small defects on aircraft surface. As the network deepens, the receptive field gradually expands, leading to the loss or merging of some information from the original feature maps during multiple downsampling processes. Additionally, the spatial resolution of the feature maps gradually decreases, which is particularly unfavorable for small targets. To enhance the network's capability to detect small targets, this paper incorporates the Context Enhancement Module (CEM) [33] to aggregate contextual information from feature maps of different scales, thereby improving the overall feature representation. This enhancement allows for more precise identification and localization of small defects by compensating for the loss of detail in lower-resolution feature maps.

Fig. 7 illustrates the structure of CEM, which utilizes dilation convolution with different expansion rates to capture contextual information from various receptive fields. In the network, feature maps of different scales are divided into three branches during the feature map fusion splicing operation. Each branch undergoes dilation convolutions with convolution kernel sizes of 3×3 and dilation rates of 1, 3, and 5. This configuration enlarges the effective receptive field of each convolution kernel, thereby increasing the contextual information at each location on the feature map. This process strengthens the capability of the model to perceive global and local information while maintaining the spatial size of the feature map, and contributes to the extraction and preservation of features associated with small targets. Subsequently, the information from the three branches is dimensionally adjusted individually through 1×1 convolutions. Finally, the outputs are then aggregated by direct summation to produce an enhanced feature map. This final map is seamlessly integrated with feature maps from other scales, ensuring a robust and detailed feature representation across the network.

3.4. Attention mechanism

To enhance focus on defect features while suppressing irrelevant background information, the Global Attention Mechanism (GAM) [34]

is introduced into the backbone network to improve its feature extraction capability. Compared to previous attention methods that ignore the importance of spatial and channel information for enhancing cross-dimensional interactions, GAM operates across these dimensions. It effectively captures key features from all three dimensions, thus preserving channel and spatial information integrity while amplifying global spatial-channel interactions. The structure of GAM is shown in Fig. 8.

The channel attention submodule uses a 3D permutation to rearrange the channel dimensions, preserving information spanning three dimensions, while using a two-layer MLP (multilayer perceptron) to amplify the cross-dimensional channel-space dependencies. Regarding the input feature map $F_1 \in \mathbb{R}^{C \times H \times W}$, it is the first to perform the dimension transformation and then fed to MLP for processing, after which the feature map output by the MLP is re-transformed to the initial dimension, and sigmoid operation is applied to get the output channel attention feature vector $M_c(F_1)$, defined as:

$$M_c(F_1) = \sigma(\text{permute}(\text{MLP}(\text{permute}(F_1)))) \quad (3)$$

where σ is the sigmoid function.

The channel attention feature vector is elementwise multiplied with the input features to obtain the intermediate state F_2 , which is served as the input feature for the spatial attention submodule, denoted as:

$$F_2 = M_c(F_1) \otimes F_1 \quad (4)$$

To focus on spatial information, the spatial attention sub-module then uses two convolutional layers for spatial information fusion. The number of channels is first reduced by convolution with convolution kernel of 7 (C to C/r , $r = 4$, denoting the reduction factor) to shrink the computation, then the number of channels is increased by a convolution operation with convolution kernel of 7 to keep the number of channels the same, and finally, the spatial attention feature $M_s(F_2)$ is outputted by the Sigmoid, which is defined as follow:

$$M_s(F_2) = \sigma(f^{7 \times 7}(f^{7 \times 7}(F_2))) \quad (5)$$

The spatial attention feature vector is elementwise multiplied with the input features of the spatial attention sub-module to obtain the output feature F_3 of the GAM attention module, using the formula expressed as:

$$F_3 = M_s(F_2) \otimes F_2 \quad (6)$$

where M_c and M_s are channel and spatial attention feature maps, respectively, and $\otimes$ denotes element-based multiplication.

3.5. Loss function

Given the significant variations in sample sizes across different defect types in aircraft defect images, along with the uneven difficulty of samples within each category, these imbalances considerably affect the training model's performance. Hence, this paper proposes a sliding weight function EMA-Slide based on exponential moving average. Different samples will obtain different loss weight coefficients through the EMA-Slide function, which is applied to calculate the binary cross entropy of confidence and classification. It can effectively focus on difficult samples to improve the sample imbalance problem. The specific calculation process is as follows:

$$L_{BCE} = -y \log(\hat{p}) - (1 - y) \log(1 - \hat{p}) \quad (7)$$

$$L_{ESL} = L_{BCE} \times ES(x, \mu) \quad (8)$$

where L_{BCE} represents the binary cross-entropy loss function; $\hat{p}$ represents the predicted probability of each sample; y represents the true label of the sample, which is 1 when it belongs to a certain category, otherwise it is 0; L_{ESL} represents the new EMA-Slide loss function; $ES(x, \mu)$ denotes the sample loss weight coefficient obtained through

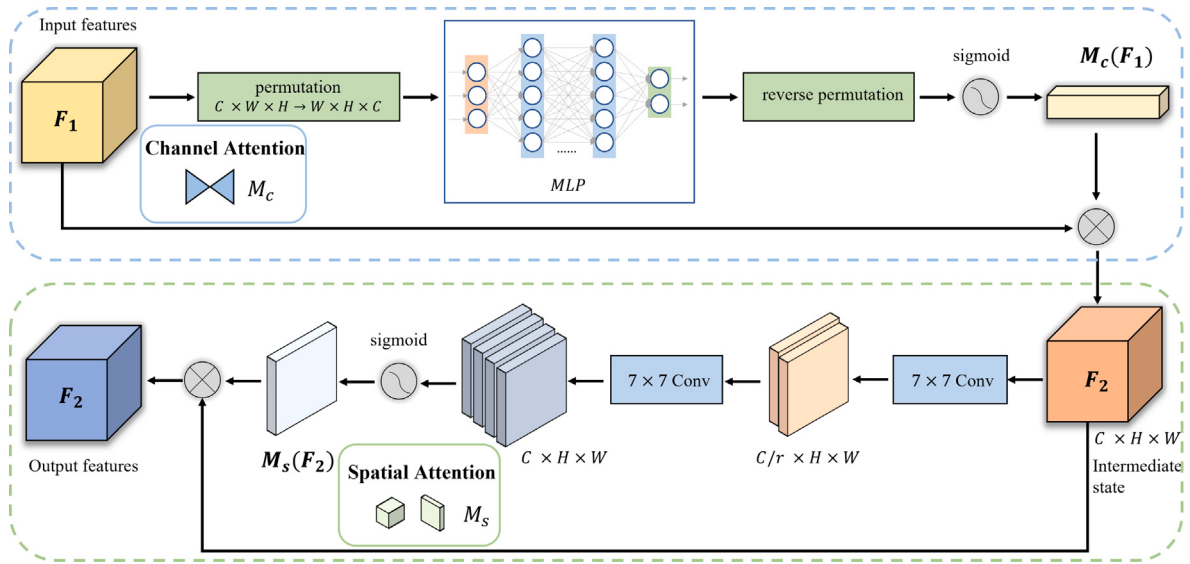


Fig. 8. The GAM module structure. It uses a sequential channel-space attention mechanism, including two sub-modules: the channel and the spatial.

the EMA-Slide function, the principle and calculation are described below.

EMA-Slide introduces an automatically updatable IoU mean value μ based on the exponential movement method as a threshold to judge the difficulty of a sample, thus assigning different loss weights to different samples. The judgment rule is that samples larger than μ are regarded as positive samples and samples smaller than μ are regarded as difficult to identify. Typically, samples well below μ are considered almost impossible to have a significant impact through optimization, while samples in the vicinity of μ may have fuzzy boundaries that can be further identified and optimized to improve performance. In order to more fully utilize this portion of samples to optimize the learning capability of the network, we designs the EMA-Slide function to smoothly assign greater weights to samples near μ to increase the relative loss of confusingly difficult samples, so that the network pays more attention to these samples, and at the same time makes the network more focused on the most recent trend of the IoU data under the current training, so that the model can be adjusted based on the latest optimization results in each iteration. The EMA-Slide function is formulated as follows:

$$ES(x, \mu) = \begin{cases} 1 & x \leq \mu - 0.1 \\ e^{1-\mu} & \mu - 0.1 < x < \mu + 0.1 \\ e^{1-x} & x \geq \mu + 0.1 \end{cases} \quad (9)$$

where x denotes the IoU value of the current prediction sample; μ is the average IoU value of all samples, which is automatically updated for a new round by exponential movement each time it is invoked, and the update calculation formula is as follows:

$$\begin{cases} \mu_n = \beta \cdot \mu_{n-1} + (1 - \beta) \cdot \mu_n \\ \beta = \beta_0 \cdot \left(1 - e^{-\frac{n}{\tau}}\right) \end{cases} \quad (10)$$

where μ_n is the current value of the IoU mean of all samples and its updated value is automatically calculated when the EMA-Slide function is called; μ_{n-1} is the IoU mean after the last update; β_0 is the initial exponential moving weight decay rate; β is the exponential moving weight decay coefficient; n represents the number of calculation updates of the loss; τ is the time constant used to control the decay rate.

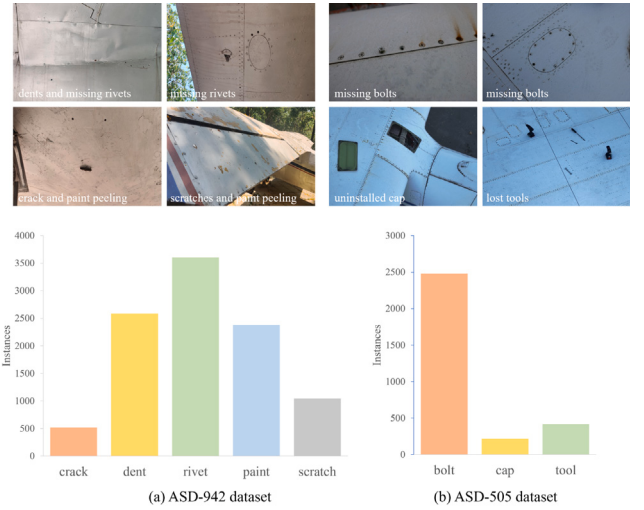


Fig. 9. Sample images and distribution of defects instance numbers by category from the datasets used.

4. Experimental results analysis

4.1. Dataset

Aircraft surface defects datasets for training and testing is the basis of related research work. To verify the effectiveness and generality of the algorithm, we used two datasets: one from a public source and the other that we constructed. The sample images and defect instance distribution are illustrated in Fig. 9, with more details on the datasets provided below. Moreover, during the training process, each dataset is divided into a training set, a validation set and a test set in the ratio of 8:1:1.

The ASD-942 dataset was developed by preprocessing a public dataset that include 314 original images sourced from the website (https://universe.roboflow.com/ddisc/aircraft_skin_defects). This dataset primarily contains five types of defects: cracks, dents, missing rivets, paint peeling, and scratches. Among these, 123 images have a size of 1920×1080 , and 191 images have a size of 4000×1824 .

Table 1
Ablation experiments on ASD-942 dataset.

Method						Metrics		
Case	Baseline	DCNC3	CEM	GAM	EMA-Slide Loss	P (%)	R (%)	mAP@0.5 (%)
(a)	✓					73.2	47.1	50.3
(b)	✓	✓				74.7	49.6	52.5
(c)	✓		✓			79.2	48.7	52.6
(d)	✓			✓		78.1	48.2	51.6
(e)	✓				✓	75.7	47.7	50.7
(f)	✓	✓	✓			79.8	51.7	54.3
(g)	✓	✓		✓		78.5	49.0	53.1
(h)	✓		✓	✓		74.3	48.7	52.2
(i)	✓	✓	✓	✓		79.1	52.0	55.6
(j)	✓	✓	✓	✓	✓	80.7(7.5 ↑)	50.5(3.4 ↑)	56.0(5.7 ↑)

Note: The bold value indicates that best result among all models.

Due to the small size of the original dataset, we implemented five data augmentation strategies to expand the sample size and prevent model overfitting. These strategies include brightness adjustment, cropping, rotation, translation, and mirror flipping. Following these enhancements, the dataset comprises a total of 942 sample images.

The ASD-505 dataset, specifically constructed for this study, comprises 505 images captured under varied lighting conditions, distances, and viewing angles. It includes three types of defects: missing screws, missing caps, and lost tools. Of these, 312 images were acquired from long-distance and multi-viewpoints by drones equipped with cameras, each with an image resolution of 4000×3000 . The remaining 193 images were collected by handheld cameras simulating ultra-close drone shooting environments, with each image having a resolution of 4928×3288 .

4.2. Experimental condition

In the experiments, the model was implemented using PyTorch 1.12 and Python 3.7 as the deep learning framework. The hardware used was a PC equipped with a 12th Gen Intel(R) Core(TM) i7-12700KF CPU@3.60 GHz, 64 GB RAM, and an NVIDIA GeForce RTX 3090Ti (24G). The operating system environment was Windows 10.

The parameter settings used during the experimental training are summarized as follows: In training, the epoch is set to 300, the batch size is set to 16, the input image size is uniformly 640×640 , pre-training weights are not used, the SGD optimizer is employed to update the network weights, cosine cooling strategy is used for the learning rate update, the initial learning rate lr_0 is 0.01, the periodical learning rate lrf is 0.01, and the weight decay coefficient is 0.0005 and momentum is 0.937. In addition, HSV, multi-scale, translation, flip, and mosaic are also applied for data enhancement.

4.3. Evaluation metrics

The main performance evaluation metrics in this study include P (Precision), R (Recall), and mAP (mean Average Precision). The computation of these metrics is based on the confusion matrix, which are defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

$$\begin{cases} AP = \int_0^1 P(R) dR \\ mAP = \frac{\sum_{i=1}^N AP_i}{N} \end{cases} \quad (13)$$

where N is the number of defects categories, i refers to the i th category, TP represents the count of correctly predicted positive samples by the model, TN represents the count of correctly predicted negative samples, FP represents the count of incorrectly predicted positive samples, and FN represents the count of incorrectly predicted negative samples.

The mAP@0.5 chosen in this paper represents the average accuracy value at IoU threshold of 0.5. When the IoU of the predicted and labeled boxes is greater than 0.5, it is considered as a correct prediction of the object. Under this premise, the mAP value is further calculated.

4.4. Ablation experiment

To validate the effectiveness of each module in the proposed improvement method, we conducted different ablation experiments on two datasets.

(1) ASD-942 dataset

Ten ablation experiments were designed on the ASD-942 dataset. As shown in Table 1, Case (a) uses the original YOLOv5s as the baseline model. Case (b) to (e) introduce four improved measures based on Case (a) to assess the contribution of each individual module to model performance. Case (f) to (j) aim to verify the synergy and complementarity between modules by sequentially stacking multiple improvement modules on top of the baseline.

According to the results in Table 1, the first five rows demonstrate improvements in the values of P, R, and mAP@0.5 after individually adding deformable convolution, contextual enhancement, attentional mechanism, and exponential moving averaging slide function by utilizing the control variable method. Specifically, the improvements in mAP are 2.3%, 2.4%, 1.4%, and 0.5%, respectively, indicating that each module provides a positive effect on the defects detection performance. Notably, the addition of the context enhancement module gives better results than the other improvements. This is attributed to the presence of a larger number of small targets in the dataset, and the fact that the CEM utilizes the dilatation convolution with different dilatation rates to obtain the contextual information of the different receptive fields of the small targets, which enriches the input of the small-target feature information into the feature fusion structure. Further analysis of other combination experiments from Case (f) to Case (j) clearly show that the detection performance of stacked modules surpasses that of individual modules, with mAP improvements of 4.1%, 2.9%, 2.0%, 5.4%, and 5.7%, respectively. These results suggest that the modules not only enhance detection effectiveness individually but also exhibit significant synergistic and complementary interactions, thereby boosting overall performance.

Additionally, it is evident from Figs. 10(a) and 10(b) that ASD-YOLO exhibits promotion compared to the baseline curve. The AP values of the five defects (crack, dent, rivet, paint, and scratch) have increased by 14.0%, 3.7%, 0.3%, 0.6%, and 10.7%, respectively, further validating the effectiveness of the improved method. Among them, the most significant improvement is achieved for cracks and scratches with large differences in length and width, indicating that the improved deformable convolution module DCNC3 plays an important role here.

Analyzing the color of the diagonal region in the confusion curves from Figs. 10(c) and 10(d), the proposed method shows a darker color compared to the baseline, while the color of the two sides is lighter. This indicates that the ability of the proposed method to correctly

Table 2
Ablation experiments on ASD-505 dataset.

Method						Metrics		
Case	Baseline	DCNC3	CEM	GAM	EMA-Slide Loss	P (%)	R (%)	mAP@0.5 (%)
(a)	✓					84.8	84.3	84.7
(b)	✓	✓				87.4	83.4	88.0
(c)	✓		✓			88.0	81.0	86.2
(d)	✓			✓		87.4	84.9	87.6
(e)	✓				✓	83.7	87.1	86.7
(f)	✓	✓	✓	✓	✓	85.9(1.1 ↑)	86.5(2.2 ↑)	88.1(3.4 ↑)

Note: The bold value indicates that best result among all models.

Table 3
Comparison experiments on ASD-942 dataset.

Method	Backbone	Input size	AP (%)					mAP@0.5 (%)	Speed (ms)
			Crack	Dent	Rivet	Paint	Scratch		
Faster R-CNN [35]	ResNet50	640 × 640	9.4	24.3	0.08	6.3	12.3	10.5	38.7
YOLOv5	—	640 × 640	40.0	64.9	42.6	67.4	36.4	50.3	9.4
YOLOv6 [36]	—	640 × 640	13.6	47.4	18.8	30.6	9.8	24.0	32.5
YOLOv7 [37]	—	640 × 640	10.8	43.6	31.5	33.9	7.39	25.4	29.9
YOLOv8	—	640 × 640	32.5	77.0	26.0	52.8	40.9	45.8	21.1
YOLOv9 [38]	—	640 × 640	27.9	61.7	24.4	45.9	27.7	37.5	39.5
SSD [39]	VGG	640 × 640	16.4	40.5	6.8	25.9	26.6	23.2	16.6
Ours	—	640 × 640	53.0	69.0	42.7	68.0	47.2	56.0	15.8

Note: (1) The bold value indicates that best result among all models. (2) The Algorithm speed is obtained by the average inference time of three experiments per 640 × 640 image at batch-size 1.

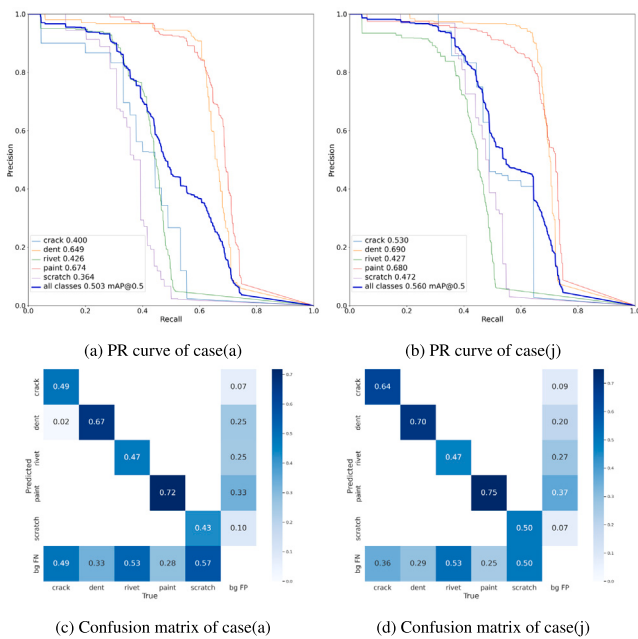


Fig. 10. Comparison of PR curve and confusion matrix between baseline and our method on ASD-942 dataset.

predict object categories is enhanced and misidentification is reduced. However, for objects such as missing screws, scratches, the percentage of objects judged as background is high, implying that a portion of these categories are missed in the detection process. Although the improved model reduces these types of missed detections, they still constitute a significant portion of the total, attributed to the slight nature of scratches and the small size of screws, making them challenging to detect in complex environments with similar background.

(2) ASD-505 dataset

The experimental results of the ASD-505 dataset are listed in Table 2, after the improvement strategies are sequentially introduced to the baseline model yolov5s, the mAP of each module is improved by 3.3%, 1.5%, 2.9% and 2.0%, respectively, and the ASD-YOLO model

integrating all the improvements improves the P, R, and mAP by 1.1%, 2.2%, and 3.4%, respectively, compared to the baseline. In addition, observing Fig. 11 shows that the PR curves of the proposed method are smoother and closer to the upper right side, and the AP values of all three defects (missing screws, missing caps and lost tools) are also increased by 2.9%, 1.5%, and 5.7%, respectively, which indicates that the performance of the model has been improved to some extent. Meanwhile, the confusion matrix shows an increase in the values representing the correct identification of the missing screws, and a decrease in values representing the misidentification of the missing screws as background and the misidentification of the background as a lost tool. This indicates that the identification capabilities of small targets like missing screws in ASD-YOLO is enhanced, and the misidentification rate is further reduced due to the greater focus on the defect information.

4.5. Comparison of detection performance

In this section, the proposed method is compared with various object detection methods (Faster R-CNN, SSD, YOLOv 5 YOLOv9) [35–39]. Each algorithm was selected based on its influence, advancement, representativeness in the field of object detection, and relevance to the proposed algorithm. This ensures a comprehensive comparison with state-of-the-art methods across various metrics, thereby emphasizing the advantages of our proposed method.

Broadly speaking, Faster R-CNN, SSD, and the YOLO series are the most typical and widely applied algorithms in both industrial and research domains, making them standard benchmarks for evaluating new methods. Specifically, Faster R-CNN was chosen for its high standards in object detection accuracy, particularly through its Region Proposal Network (RPN) mechanism. Comparing with Faster R-CNN allows us to demonstrate how ASD-YOLO maintains detection speed while improving accuracy. SSD was selected because it strikes a good balance between detection speed and accuracy, aligning with the goals of ASD-YOLO. Comparing with SSD helps illustrate improvements in this trade-off. The YOLO series was chosen for its real-time capabilities, which are critical in practical applications. YOLOv5 YOLOv9 represent incremental versions, with YOLOv9 being the latest in this series. Comparing ASD-YOLO with these versions highlights its advantages in real-time performance under similar conditions. Next, qualitative and quantitative experiments will validate these comparisons.

Table 4
Comparison experiments on ASD-505 dataset.

Method	Backbone	Input size	AP (%)			mAP@0.5 (%)	Speed (ms)
			Bolt	Cap	Tool		
Faster R-CNN [35]	ResNet50	640 × 640	0.47	95.2	39.9	45.2	34.1
YOLOv5	—	640 × 640	72.0	97.4	84.7	84.7	12.4
YOLOv6 [36]	—	640 × 640	45.5	99.5	90.3	78.4	38.4
YOLOv7 [37]	—	640 × 640	70.3	99.3	86.8	85.5	46.0
YOLOv8	—	640 × 640	49.3	97.8	91.3	79.5	29.5
YOLOv9 [38]	—	640 × 640	39.1	99.2	88.8	75.7	68.6
SSD [39]	VGG	640 × 640	34.3	98.2	80.5	71.0	26.6
Ours	—	640 × 640	74.9	98.9	90.4	88.1	21.6

Note: (1) The bold value indicates that best result among all models. (2) The Algorithm speed is obtained by the average inference time of three experiments per 640 × 640 image at batch-size 1.

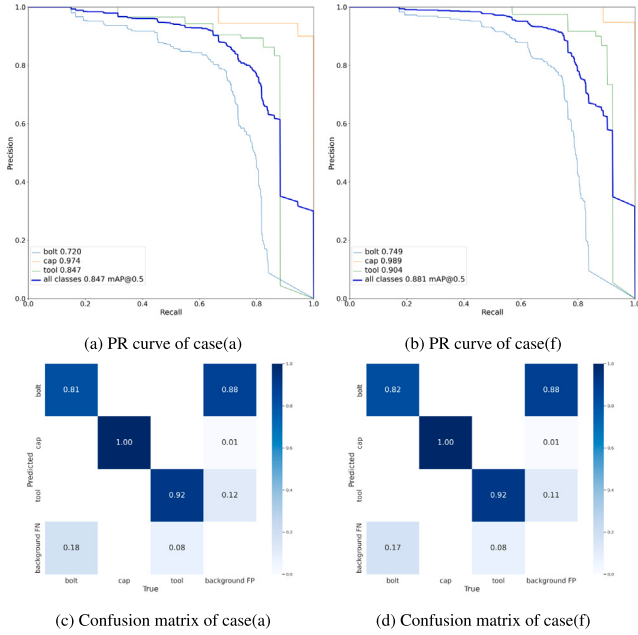


Fig. 11. Comparison of PR curve and confusion matrix between baseline and our method on ASD-505 dataset.

(1) ASD-942 dataset

The main challenges of this dataset are multi-scale defects, small defects and cluttered background information in the images. The quantitative results, detailed in Table 3, clearly demonstrate that ASD-YOLO achieves the best performance in terms of mAP. Compared to Faster R-CNN, YOLOv5, YOLOv6, YOLOv7, YOLOv8, YOLOv9 and SSD, the mAP is improved by 45.5%, 5.7%, 32.0%, 30.6%, 10.2%, 18.5%, and 32.8%, respectively. Upon closer examination and analysis of the AP values for five different categories, it is observed that, in the specific defect category of missing screws, Faster R-CNN and SSD exhibit lower metrics, indicating their relatively insufficient capability to recognize small targets. For defects categories such as scratches and cracks, YOLOv7 shows lower AP values, suggesting limitations in its recognition ability when dealing with targets with significant aspect ratio differences. It is noteworthy that ASD-YOLO achieves the highest AP values for the categories of cracks, paint loss, missing screws, and scratches, highlighting its outstanding performance in the comprehensive detection task of various types of aircraft surface defects.

As shown in Fig. 12, four representative aircraft surface defects images were selected as test samples, encompassing all defects types, for a qualitative comparison of detection performance among different methods. The results for Image1 and Image2 indicate that Faster R-CNN and SSD essentially fail to detect missing screws, while YOLOv6, YOLOv7, YOLOv8 and YOLOv9 exhibit missed detections for missing screws. In

the results for Image3, YOLOv7 and SSD miss the presence of dents. Observing Image4, when dealing with dense paint peeling defects involving multiple targets, Faster R-CNN and SSD exhibit significant missed detections, especially for smaller paint peeling areas, and Faster R-CNN's target localization is inaccurate. YOLOv5 show high overlap of detection boxes. YOLOv6 excels at identifying small targets but fails to recognize large-area normal-sized targets. YOLOv7 performs well on multi-scale targets but still has occasional missed detections of normal-sized targets. YOLOv8 and YOLOv9 performs well in recognizing most paint loss areas, with only a few instances of difficulty in accurately identifying tiny defects. In comparison to these methods, ASD-YOLO demonstrates more accurate and comprehensive detection of various defects, with a lower missed detection rate, more precise target localization, and higher confidence, exhibiting greater effectiveness and adaptability in real-world aircraft surface inspection scenarios. Overall, the qualitative results further validate the superiority of ASD-YOLO.

(2) ASD-505 dataset

The experimental results of quantitative comparison between ASD-YOLO and other methods on ASD-505 dataset are displayed in Table 4. Compared with Faster R-CNN, YOLOv5, YOLOv6, YOLOv7, YOLOv8, YOLOv9 and SSD, the mAP is improved by 42.9%, 3.4%, 9.7%, 2.6%, 8.6%, 12.4% and 17.1%, respectively. A detailed analysis of the AP values across various defect categories reveals that ASD-YOLO leads in overall metrics, although it did not achieve the highest metrics on every category. It is noticeable that Faster R-CNN has extremely low metrics in tiny screw missing defect recognition, which may be due to the fact that the receptive field of the network is too large and does not capture tiny targets when generating candidate frames using RPN. In addition, the single image inference time of ASD-YOLO is slightly increased compared to the baseline network, but still meets the real-time requirements.

The biggest challenge of this dataset is that there are many tiny defects in the images and these images are from different lighting conditions, so we selected four very typical images for qualitative comparison of ASD-YOLO to show the recognition effect of the different methods, as shown in Fig. 13. Image1 is normal lighting with many tiny defects, Image2 is weak lighting with defective objects of different scales, Image3 is dark illumination and defective objects with large differences in scale, and Image4 is strong illumination and has many defective objects. From the detection results, we can clearly find that the Faster R-CNN cannot detect small defects, and overlapping detection frames often appear in the detection of small and medium-sized targets, and the tools that are similar to the background in Image4 cannot be detected; YOLOv5 has a better effect on the recognition of small defects under different lighting, but overlapping frames appear in the detection of cross-targets in a close distance, and the background is easily recognized as a defective target in Image4; YOLOv6, YOLOv7, YOLOv8, and YOLOv9 all have different degrees of missed recognition, especially in strong illumination; SSD also fails to detect tiny missing screw defects altogether, but detects normal-sized defective targets well; ASD-YOLO has no missed recognition although there are a few misdetections of tiny defects in Image1, and under other conditions all defects are correctly recognized. All defects correctly recognized, with the best combined detection results.

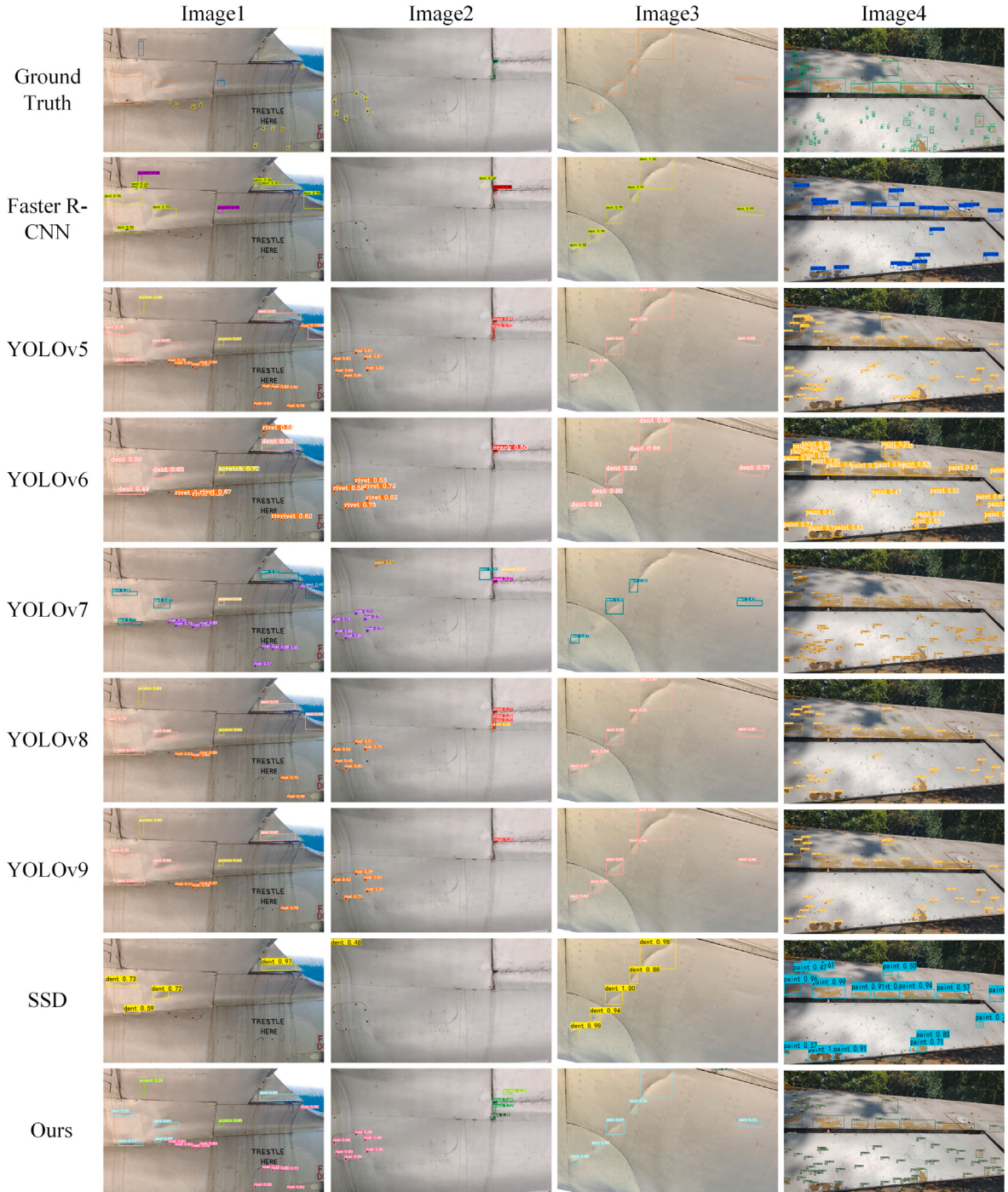


Fig. 12. Detection results of qualitative comparison on ASD-942 dataset.

4.6. Comparison of computational performance

Compared to the proposed ASD-YOLO, the parameters and computational complexity of other commonly used models are listed in Table 5. Although ASD-YOLO does not have the lowest computational complexity — being approximately four times higher than that of the baseline YOLOv5, it is evident from the preceding detection performance metrics that ASD-YOLO outperforms these models. Furthermore, it meets the real-time detection needs. In industrial application scenarios where high precision and rapid detection are critical, achieving an optimal balance between speed and accuracy remains a key focus for future development.

5. Discussion

5.1. Analysis of small defects detection performance

In the MS COCO dataset, objects with an area smaller than 32×32 pixels are considered small targets. According to this criterion, we analyzed the two datasets used in this paper, and the distribution of small defects is shown in Fig. 14. It is obvious that small targets occupy an important position in the datasets. The challenge of small target detection lies in the limited number of pixels and weak features, making effective feature extraction difficult. Moreover, as the network deepens,

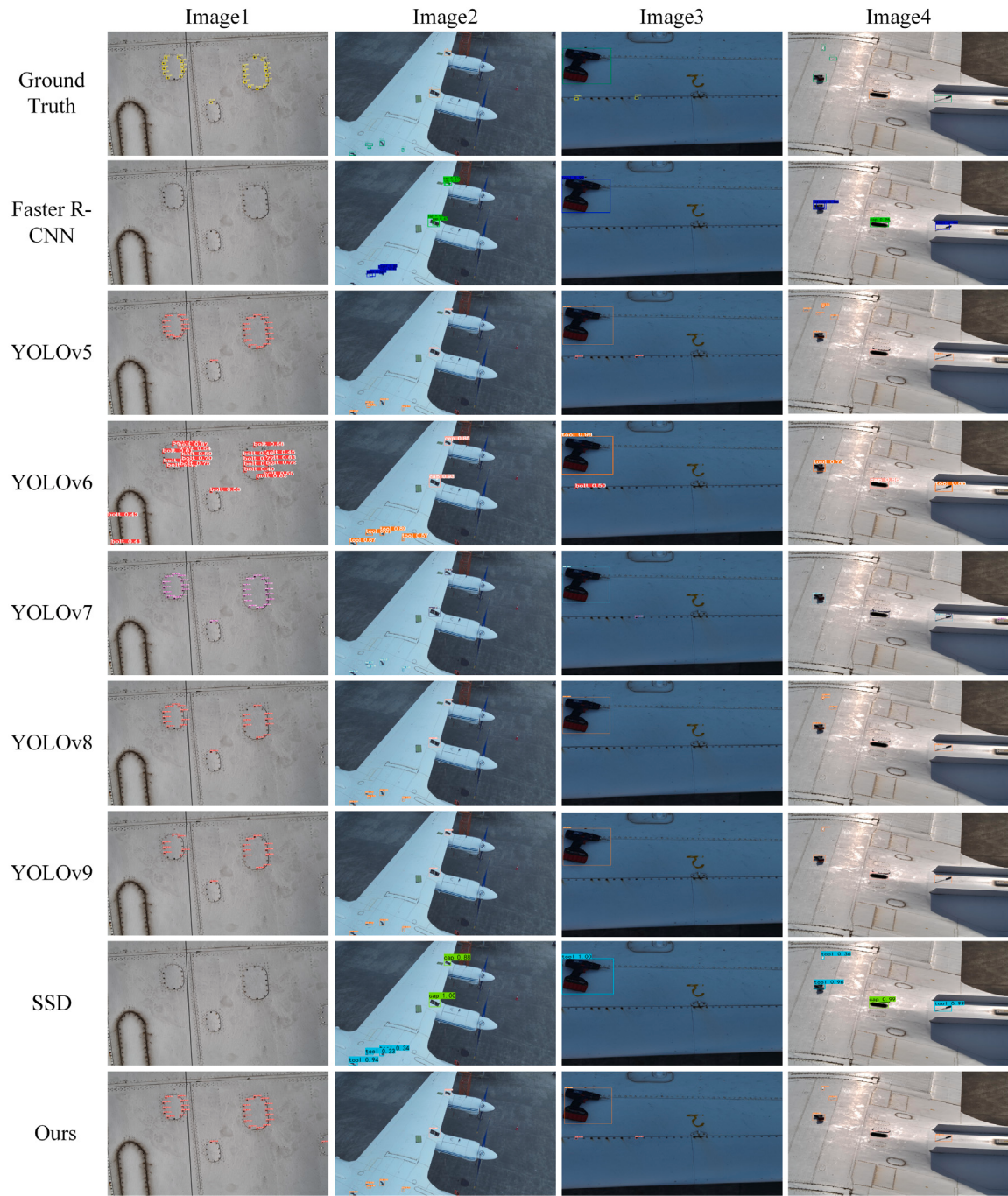


Fig. 13. Detection results of qualitative comparison on ASD-505 dataset.

Table 5

Comparison in Total GFlops (G) and Total Params (M)

Model	GFlops (G)	Params (M)
Faster R-CNN [35]	401.79	136.77
YOLOv5	15.92	7.03
YOLOv6 [36]	45.17	18.50
YOLOv7 [37]	105.10	37.21
YOLOv8	28.50	11.13
YOLOv9 [38]	266.10	60.77
SSD [39]	274.93	24.15
Ours	59.76	27.39

Note: Input shape is [640, 640].

the receptive field continues to expand, and the microscopic information of small targets will also be lost. Therefore, the proposed algorithm alleviates those by designing a context enhancement module in the shallow layers of the network. This module uses dilated convolutions with different rates to capture contextual information from various receptive fields, which is then fused in the deeper network layers to enhance the representation of small target features. Additionally, the introduction of a global attention mechanism suppresses irrelevant background features, allowing the network to focus more on learning the features of small target defects.

In the ASD-942 dataset, small target defects account for 40.4% of the total number of detected targets. Since small targets are distributed across five types of defects, we analyzed the detection metrics for these five defects to reflect the performance of the proposed algorithm in

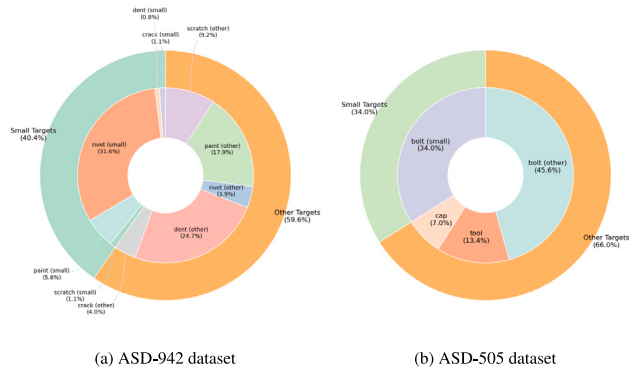


Fig. 14. Distribution of small defects on the datasets.



Fig. 15. Example of small defects detection results. The red arrows in the images indicate the missing of the screws.

small target detection in this dataset. As shown in Table 3, compared to the other algorithms, ASD-YOLO achieves the highest AP values for four types of defects. Specifically, compared to the baseline model, ASD-YOLO improves the AP values by 13.0%, 4.1%, 0.1%, 0.6%, and 10.8% for the five defects, respectively. In the ASD-505 dataset, small target defects constitute 34.0% of the total defects, and only one type of defect (screw missing) includes small targets. Therefore, we assume that the detection performance of this defect (missing screws) represents the overall small-defect detection performance of this dataset. As shown in Table 4, ASD-YOLO achieves superior AP values for this type of defect compared to other algorithms, with an improvement of 2.9% over the baseline model.

We also selected images containing only small targets for testing (indicated by red arrows in the images, representing missing screws that need to be correctly detected) in practical scenarios. The detection results are shown in Fig. 15. Most of the small targets can be recognized correctly, but there are inevitably some misdetections, which are due to the fact that the features of the small targets are not obvious, and the small targets to be detected are similar to the information of other normal screws in the background, which is what makes it difficult for the algorithm to distinguish between targets and background accurately. False detections show the limitations of ASD-YOLO, indicating that its small target detection performance still needs to be continuously improved and perfected to meet higher detection requirements.

5.2. Analysis of practical system feasibility

We have conducted a preliminary application of the proposed method by establishing an aircraft surface detection scenario on the



Fig. 16. Side and top view of algorithm application scenarios.



Fig. 17. The actual scene detection results of the visual interactive platform embedded with the proposed algorithm.

DJI M30 drone platform to verify the deployment effectiveness of the algorithm in actual systems. The specific application scenario is illustrated in Fig. 16. The premise of the implementation is the generation of the automatic inspection path of the drone. First, a full coverage viewpoint planning algorithm is designed based on the aircraft model, and then the inspection path used to execute automated drone inspections is generated according to the viewpoint planning results, completing the automatic inspection function of the drone. On this basis, we developed a visual interactive platform embedded with the proposed algorithm to perform defect detection tests on defect images and real-time video streams obtained by the drone inspection. The results showed that our algorithm is fundamentally feasible in terms of real-time performance and accuracy, addressing the practical needs of automated inspection tasks. The actual scene detection results of the visual interactive platform is shown in Fig. 17.

During the implementation process, we faced technical challenges in the recognition of small targets and multi-scale irregular surface defects. By the improvements in the feature extraction and fusion design of the proposed algorithm, we effectively enhanced the detection performance. These enhancements are critical for the practical deployment of our method, ensuring that it can handle the complexity of real-world inspections. Meanwhile, the test of the actual system also exposed some problems to be solved, which need to be further optimized to promote demonstration and promotion in actual application scenarios.

In short, there are three problems to be optimized. Firstly, one of the factors affecting the generation of inspection viewpoints and paths is the focal length of the camera. If we want to reduce the number of inspection viewpoints and thereby decrease inspection time, we

have to set a smaller zoom factor, which will result in an increased field of view in the inspection images, and relatively smaller defect targets such as screws, significantly complicating detection and posing greater challenges for the algorithm's small target detection performance. Secondly, in the process of real-time detection of video streams through streaming media, the frame rate limit of data transmission leads to insufficient clarity of the returned video images, which are blurrier than the images returned by a single drone, thereby increasing the risk of missed detection and false detection. How to reconstruct the image frame clarity during real-time video stream detection is an important direction for our future research. Secondly, during real-time video stream detection via streaming media, limitations in data transmission frame rate resulted in the returned video frame having insufficient clarity and being more blurred compared to images directly returned by the drone, thereby increasing the risk of missed and false detections. Lastly, although the current algorithm has been deployed on a laptop and has achieved the real-time inference speed, if there is a need to deploy the proposed algorithm on resource-constrained edge computing devices in the future, the demands on model size and processing speed will become more rigorous. Therefore, it is necessary to consider further reducing model size and accelerating the inference process through methods such as model pruning and distillation to enhance the real-time performance of the algorithm, making it more suitable for deployment in actual systems.

5.3. Advantages and limitations

In the experimental study, we utilized three datasets to assess the performance of the proposed method. Following analysis, we summarized the advantages and limitations of ASD-YOLO as follows:

(1) Advantages

- We propose an effective deformable convolution-based feature extraction module to capture irregularly shaped defects with large differences in aspect ratio deformation, and utilize a context-enhanced module to aggregate feature information at different scales to retain more tiny defective features, while employing a global attention mechanism to suppress the background information, which significantly improves detection performance.
- We design an exponential sliding average weight assignment function to improve the loss function, which adaptively adjusts the loss weights of poorly categorized boundary samples through dynamic changes in the training process, so as to improve the network's attention to difficult-to-recognize samples and alleviate the imbalance problem between simple and difficult samples.

(2) Limitations

- Despite improvements in detection accuracy on the datasets, the defect detection performance on the public datasets (ASD-924) remains below expectations for industrial applications, constrained by the complexity of the data sets and the nature of aircraft surface defects.
- Compared with the baseline model, the proposed method enhances detection accuracy through improved strategies, inevitably at the expense of increased model parameters and computational complexity. Although this method still meets the real-time requirements, it also leads to longer inference times and complicates model deployment.

6. Conclusion

In this study, we propose ASD-YOLO, an efficient and accurate method for detecting defects of aircraft surface, in response to the real-world needs of aircraft structural integrity and flight safety in complex and changing environments. This approach firstly designs the DCNC3 module for network feature extraction and fusion to improve

the detection accuracy of irregularly shaped defects. Secondly, a global attention mechanism is introduced to suppress redundant background information so that the network is more focused on the damage features. Then, the context enhancement module is utilized to consider both local and global information to enhance the defect feature representation at different scales. Finally, a sliding weight function based on exponential sliding average is proposed and applied to the loss function of the network to alleviate the problem of sample imbalance. A series of experimental results show that the method outperforms the original framework and other mainstream algorithms, significantly improves the aircraft surface defects recognition, and is of great significance in the field of aircraft maintenance and protection.

However, scarcity and quality issues in available datasets lead to leakage and misdetection during UAV inspections in industrial scenarios. Subsequent research will therefore focus on the following aspects: first, expanding the diversity of aircraft defects to produce high-quality aircraft skin datasets; second, exploring low-sample training strategies to address sample limitations; and third, designing and optimizing lightweight algorithms for easy deployment on airborne equipment at the application scene. In conclusion, we are committed to continuously optimizing the aircraft surface detection algorithm to improve the defect detection performance, make the algorithm efficient and accurate enough for application, so as to reduce labor costs, enhance the maintenance safety and elevate overall maintenance efficiency in the field of aircraft surface detection.

CRediT authorship contribution statement

Bin Huang: Data curation, Validation, Writing – original draft. **Yan Ding:** Writing – original draft, Methodology. **Guoliang Liu:** Writing – review & editing, Supervision, Funding acquisition. **Guohui Tian:** Project administration, Conceptualization. **Shanmei Wang:** Visualization, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships with other people or organizations that could inappropriately influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is partially supported by the Jinan Science and Technology Bureau, China (2021GXRC026), Young Scholars Program of Shandong University (2018WLJH71), the Fundamental Research Funds of Shandong University, China, and the Taishan Scholars Program of Shandong Province, China.

References

- [1] A. Katunin, K. Lis, K. Jozsko, P. Žak, K. Dragan, Quantification of hidden corrosion in aircraft structures using enhanced D-sight NDT technique, *Measurement* 216 (2023) 112977.
- [2] A. Katunin, K. Dragan, M. Dziendzikowski, Damage identification in aircraft composite structures: A case study using various non-destructive testing techniques, *Compos. Struct.* 127 (2015) 1–9.
- [3] V.Y. Senyurek, Detection of cuts and impact damage at the aircraft wing slat by using Lamb wave method, *Measurement* 67 (2015) 10–23.
- [4] Y. Choi, S.H. Abbas, J.-R. Lee, Aircraft integrated structural health monitoring using lasers, piezoelectricity, and fiber optics, *Measurement* 125 (2018) 294–302.
- [5] M.R. Bato, A. Hor, A. Rautureau, C. Bes, Impact of human and environmental factors on the probability of detection during NDT control by eddy currents, *Measurement* 133 (2019) 222–232.
- [6] Y.D.V. Yasuda, F.A.M. Cappabianco, L.E.G. Martins, J.A.B. Gripp, Aircraft visual inspection: A systematic literature review, *Comput. Ind.* 141 (2022) 103695.

- [7] M.W. Siegel, W.M. Kaufman, C.J. Alberts, Mobile robots for difficult measurements in difficult environments: Application to aging aircraft inspection, *Robot. Auton. Syst.* 11 (3–4) (1993) 187–194.
- [8] J.R. Leiva, T. Villemot, G. Danguomeau, M.-A. Bauda, S. Larnier, Automatic visual detection and verification of exterior aircraft elements, in: 2017 IEEE International Workshop of Electronics, Control, Measurement, Signals and their Application to Mechatronics, ECMSM, IEEE, 2017, pp. 1–5.
- [9] N.P. Avdelidis, A. Tsourdos, P. Lafiosca, R. Plaster, A. Plaster, M. Droznika, Defects recognition algorithm development from visual UAV inspections, *Sensors* 22 (13) (2022) 4682.
- [10] M. Tappe, D. Dose, M. Oelsch, M. Karimi, L. Hösch, L. Heller, J. Knufinke, M. Alpen, J. Horn, C. Bachmeir, UAS based autonomous visual inspection of airplane surface defects, in: NDE 4.0, Predictive Maintenance, and Communication and Energy Systems in a Globally Networked World, vol. 12049, SPIE, 2022, pp. 8–21.
- [11] R. Mumtaz, M. Mumtaz, A.B. Mansoor, H. Masood, Computer aided visual inspection of aircraft surfaces, *Int. J. Image Process.* 6 (1) (2012) 38–53.
- [12] P. Nie, X. Yue, J. Yu, H. Zhang, Scratch detection method of aircraft skin based on machine vision, in: 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information, ICETCI, IEEE, 2023, pp. 839–843.
- [13] Y. Li, H. Huang, Q. Xie, L. Yao, Q. Chen, Research on a surface defect detection algorithm based on MobileNet-SSD, *Appl. Sci.* 8 (9) (2018) 1678.
- [14] X. Huang, Z. Liu, X. Zhang, J. Kang, M. Zhang, Y. Guo, Surface damage detection for steel wire ropes using deep learning and computer vision techniques, *Measurement* 161 (2020) 107843.
- [15] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, O. Büyükoztürk, Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types, *Comput.-Aided Civ. Infrastruct. Eng.* 33 (9) (2018) 731–747.
- [16] Y. Liu, J. Dong, Y. Li, X. Gong, J. Wang, A UAV-based aircraft surface defect inspection system via external constraints and deep learning, *IEEE Trans. Instrum. Meas.* 71 (2022) 1–15.
- [17] C.J. Alberts, C.W. Carroll, W.M. Kaufman, C.J. Perlee, M.W. Siegel, et al., Automated Inspection of Aircraft, Tech. Rep., William J. Hughes Technical Center (US), 1998.
- [18] Y.-J. Cha, W. Choi, O. Büyükoztürk, Deep learning-based crack damage detection using convolutional neural networks, *Comput.-Aided Civ. Infrastruct. Eng.* 32 (5) (2017) 361–378.
- [19] Y. Li, Z. Han, H. Xu, L. Liu, X. Li, K. Zhang, YOLOv3-lite: A lightweight crack detection network for aircraft structure based on depthwise separable convolutions, *Appl. Sci.* 9 (18) (2019) 3781.
- [20] T. Malekzadeh, M. Abdollahzadeh, H. Nejati, N.-M. Cheung, Aircraft fuselage defect detection using deep neural networks, 2017, arXiv preprint arXiv:1712.09213.
- [21] B. Ramalingam, V.-H. Manuel, M.R. Elara, A. Vengadesh, A.K. Lakshmanan, M. Ilyas, T.J.Y. James, et al., Visual inspection of the aircraft surface using a tele-operated reconfigurable climbing robot and enhanced deep learning technique, *Int. J. Aerosp. Eng.* 2019 (2019) 1–14.
- [22] R. Zhong, Research on aircraft skin defect identification method based on YOLOX algorithm, *J. Civ. Aviat. Flight Univ. China* 34 (57–59) (2023).
- [23] J. Hu, Research on Machine Vision-Based Dimensional Measurement and Defect Detection System for Aircraft Rivets (Master's thesis), Shanxi University of Science & Technology, 2017.
- [24] C. Chen, Application of Deep Neural Network Based on YOLOv3 in Aircraft Surface Defect Recognition (Master's thesis), Civil Aviation Flight University of China, 2020.
- [25] Y. Wei, M. Su, Defect detection method for aircraft skin riveting in framework of YOLOv5, *Modul. Mach. Tool Autom. Manuf. Tech.* (106–109) (2023).
- [26] X. Zhang, J. Zhang, J. Chen, R. Guo, J. Wu, A dual-structure attention-based multi-level feature fusion network for automatic surface defect detection, *Vis. Comput.* (2023) 1–20.
- [27] H. Li, C. Wang, Y. Liu, YOLO-FDD: Efficient defect detection network of aircraft skin fastener, *Signal Image Video Process.* (2024) 1–15.
- [28] H. Wang, L. Fu, L. Wang, Detection algorithm of aircraft skin defects based on improved YOLOv8n, *Signal Image Video Process.* (2024) 1–15.
- [29] S. Bouarfa, A. Doğru, R. Arizar, R. Aydoğan, J. Serafico, Towards automated aircraft maintenance inspection. a use case of detecting aircraft dents using mask R-CNN, in: AIAA Scitech 2020 Forum, 2020, p. 0389.
- [30] A. Doğru, S. Bouarfa, R. Arizar, R. Aydoğan, Using convolutional neural networks to automate aircraft maintenance visual inspection, *Aerospace* 7 (12) (2020) 171.
- [31] D. Meng, W. Boer, X. Juan, A.N. Kasule, Z. Hongfu, Visual inspection of aircraft skin: Automated pixel-level defect detection by instance segmentation, *Chin. J. Aeronaut.* 35 (10) (2022) 254–264.
- [32] X. Zhu, H. Hu, S. Lin, J. Dai, Deformable convnets v2: More deformable, better results, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.
- [33] J. Xiao, H. Guo, J. Zhou, T. Zhao, Q. Yu, Y. Chen, Z. Wang, Tiny object detection with context enhancement and feature purification, *Expert Syst. Appl.* 211 (2023) 118665.
- [34] Y. Liu, Z. Shao, N. Hoffmann, Global attention mechanism: Retain information to enhance channel-spatial interactions, 2021, arXiv preprint arXiv:2112.05561.
- [35] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
- [36] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., YOLOv6: A single-stage object detection framework for industrial applications, 2022, arXiv preprint arXiv:2209.02976.
- [37] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.
- [38] C.-Y. Wang, I.-H. Yeh, H.-Y.M. Liao, YOLOv9: Learning what you want to learn using programmable gradient information, 2024, arXiv preprint arXiv:2402.13616.
- [39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single shot multibox detector, in: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.