aerospace

MDPI

Article

DEE-Net: A Multi-Scale Discriminative Edge Enhancement Network for Aircraft Surface Defect Detection

Xin Wang [1], Mingxu Lu [1], Yi Liu [2,*] and Jide Qian [3,*]

[1] School of Computer Science, Civil Aviation Flight University of China, Guanghan 618307, China; wangxin@cafuc.edu.cn (X.W.); lumxyy@cafuc.edu.cn (M.L.)
[2] Key Laboratory for Civil Aviation Data Governance and Decision Optimization, Civil Aviation Management Institute of China, Beijing 100102, China
[3] Key Laboratory of Flight Techniques and Flight Safety, Civil Aviation Flight University of China, Guanghan 618307, China
* Correspondence: liuyi@camic.cn (Y.L.); qianjide@cafuc.edu.cn (J.Q.); Tel.: +86-13811299061 (Y.L.); +86-13882161442 (J.Q.)

Abstract

Efficient detection of aircraft surface defects (ASD) is a cornerstone of aviation safety. However, ASD detection is challenged by microscopic defect scales, extremely low contrast, and severe background interference. This paper proposes the Multi-Scale Discriminative Edge Enhancement Network (DEE-Net) based on an improved YOLO11. First, to mitigate feature dissipation of tiny defects, a lossless reassembly mechanism using space-to-depth convolution (SPD-Conv) is introduced, safeguarding sub-pixel topological information through space-to-depth conversion. Second, an adaptive selective edge-enhancement (ASE) module, integrating a dual-domain selection mechanism (DSM), is designed to suppress non-target redundant information on the fuselage skin. Finally, a Wise-CIoU loss function with a non-monotonic focusing mechanism is introduced to enhance localization stability under stringent IoU thresholds. Experimental results demonstrate that DEE-Net outperforms the baseline, improving mAP50 by 7.15% and mAP50-95 by 2.43%. To provide a more reliable evaluation, a 5-fold cross-validation experiment is further conducted on the original non-augmented images, and the results are reported as mean ± standard deviation. The cross-validation results provide a more conservative estimate and indicate that the proposed method achieves competitive performance across different data partitions.

Keywords: visual inspection; defect detection; aircraft maintenance; YOLO

Check for updates

Academic Editor: Jules Simo

Received: 1 April 2026
Revised: 2 May 2026
Accepted: 5 May 2026
Published: 23 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As the bedrock of ensuring operational reliability and continuous airworthiness, the structural integrity of the aircraft skin remains a focal point in civil aviation maintenance and inspection. Under complex and volatile operational conditions, fuselage surfaces are inevitably subjected to the synergistic effects of cyclic aerodynamic stresses, environmental corrosion, and foreign object damage (FOD). These factors subsequently induce various types of degradation, including crack propagation, localized indentations, and coating delamination [1,2]. Research indicates that if such incipient micro-scale damages are not accurately identified during early-stage preventive maintenance, they can escalate into catastrophic structural failures [3]. Nevertheless, contemporary maintenance frameworks remain heavily reliant on conventional non-destructive testing (NDT) techniques, primarily manual visual inspection (MVI) and acoustic-based tap testing. While these

methods offer operational flexibility in engineering practice, their diagnostic outcomes are profoundly influenced by the subjective expertise of inspectors, leading to significant inter-operator variability. Furthermore, when tasked with large-scale inspections, these manual approaches encounter severe efficiency bottlenecks and expose personnel to the inherent hazards of high-altitude operations [4,5].

Modern aviation industry's pursuit of inspection efficiency and objectivity has catalyzed the widespread application of various NDT technologies. However, the performance of existing techniques in practical hangar environments remains constrained by inherent limitations. For instance, ultrasonic testing (UT) exhibits high sensitivity to surface roughness and the quality of coupling agents; even minute interfacial air gaps can cause severe acoustic energy attenuation, thereby compromising the accuracy of defect characterization [6]. Although digital shearography can capture sub-micron surface displacement gradients, its imaging quality is highly susceptible to interference from rigid-body motion and environmental instabilities when inspecting large-scale skin structures [7–9]. More critically, the interpretation of results from these techniques still largely depends on expert heuristics, making a truly closed-loop, fully automated system difficult to achieve [10,11]. Consequently, developing a real-time, precise, and robust automated surface inspection system—such as deep learning-driven visual detection—has emerged as a pivotal research focus in the field of intelligent aviation maintenance [12–14].

Over the past five years, the evolution of artificial intelligence has spearheaded a paradigm shift in computer vision (CV) within the realm of industrial quality inspection, propelling automated NDT into a new stage of intelligence [15]. By integrating high-resolution imaging terminals with advanced algorithms, modern inspection systems have achieved a leap from coarse-grained monitoring to precision-driven analysis. In the landscape of modern intelligent manufacturing, deep learning-based visual inspection solutions—leveraging their distinct advantages of non-contact operation, high precision, and real-time online processing—are progressively superseding traditional handcrafted feature extraction methods. Consequently, these technologies have emerged as the core decision-making engine for aircraft Maintenance, Repair, and Overhaul (MRO) systems [16–19], as illustrated in Figure 1.

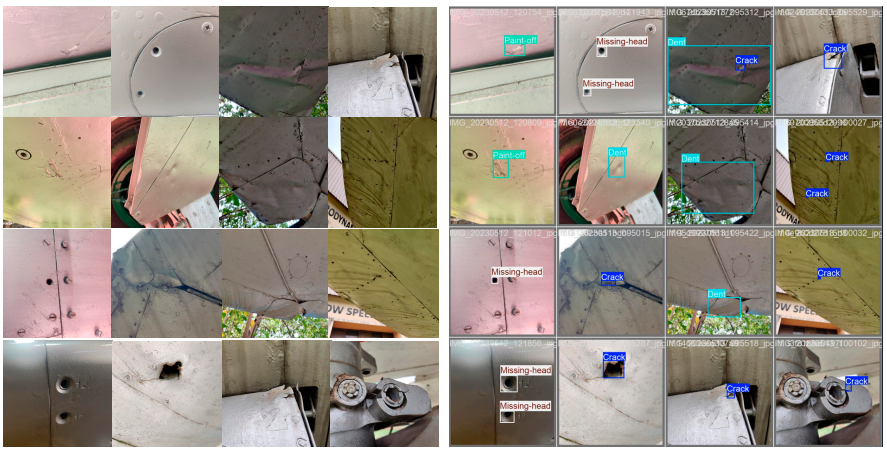


Figure 1. Application of deep learning in aircraft surface inspection.

The technical evolution of aircraft skin inspection systems is essentially a paradigm shift in feature representation. To surmount imaging challenges in high-altitude and occluded regions, a hardware matrix centered around unmanned aerial vehicles (UAVs) and intelligent guidance platforms has emerged as the standardized acquisition infrastructure [20,21]. Nevertheless, the core efficacy of the inspection system ultimately hinges

upon the backend algorithm's capacity to deconstruct and interpret massive volumes of visual data. Consequently, the research nexus has shifted from manual, rule-based feature engineering to data-driven, hierarchical representation learning. Conventional detection paradigms are often constrained by the limited adaptability of predefined filters; when confronted with low-contrast defects such as subtle scratches or indentations, rigid parameterization frequently results in high miss rates. In stark contrast, deep learning architectures leverage multi-scale feature fusion strategies to spontaneously learn semantic-rich feature descriptors directly from pixel sequences [22]. This end-to-end mechanism facilitates the adaptive decoupling of non-linear noise interference under complex operational conditions. Consequently, in real-world hangar scenarios characterized by variable illumination and severe background clutters, it demonstrates generalization resilience and recognition accuracy that are significantly superior to those of traditional heuristic algorithms.

Despite substantial advancements in visual inspection, achieving fully automated and precise identification within the domain of ASD still encounters significant bottlenecks. For instance, the majority of surface impairments, such as microscopic cracks, occupy an extremely low pixel ratio within a macroscopic field of view; their attenuated feature responses are frequently eclipsed during successive downsampling operations. Moreover, exacerbated by the high specular reflectivity of aluminum alloy skins, volatile ambient illumination, and complex fuselage curvatures, defects tend to amalgamate with background textures—such as rivet boundaries and coating particulates—resulting in exceptionally low contrast [23].

Critically, ASD samples exhibit pronounced intra-class heterogeneity (diverse morphologies within a single defect category) and inter-class similarity (e.g., superficial scratches resembling background textures), posing a rigorous challenge to feature decoupling and the demarcation of precise classification boundaries. Furthermore, the extreme scarcity of severe defect samples in real-world datasets creates a distribution imbalance that predisposes models to overfit majority classes, thereby compromising the sensitivity toward critical yet rare anomalies [24].

To address the aforementioned challenges, this study proposes DEE-Net, a high-precision defect detection framework based on an improved YOLO11. This framework adheres to the core design principles of information fidelity and selective enhancement, aiming to solve the problem of accurately identifying minute damages in complex industrial settings. The primary contributions of this paper are summarized as follows:

- We incorporate a feature reassembly strategy based on the existing SPD-Conv mechanism to mitigate information loss during downsampling for tiny defects. By replacing traditional strided convolutions with space-to-depth transformations, this approach helps preserve fine-grained spatial information and maintains sub-pixel structural details. As a result, it provides a more informative feature representation for subsequent stages of the network.
- We designed a dual-domain coordinated ASE module for feature enhancement. The module combines multi-scale perception paths with an edge enhancement component to better capture fine-grained defect features. By incorporating the existing DSM [25], the module jointly considers spatial localization and frequency-domain information. This design helps the model distinguish defect-related features from redundant background textures under complex conditions, such as metallic specular reflections.
- We incorporate a quality-aware regression optimization strategy based on the existing Wise-CIoU loss function. By integrating Wise-CIoU into the proposed framework, the model retains the comprehensive penalty terms of CIoU (including overlap area, center distance, and aspect ratio), ensuring stable localization performance. In addition, the non-monotonic focusing mechanism of Wise-CIoU is utilized to adaptively adjust the gradient contributions of samples with different qualities. This design helps mitigate

the influence of low-quality samples and reduces training instability in complex industrial scenarios. As a result, it contributes to improved bounding box regression performance, particularly for subtle defects.

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of the related work in the field of ASD detection, highlighting the evolution of methodologies. Section 3 elaborates on the architectural design and technical mechanisms of the proposed DEE-Net, detailing the synergy between its core modules. Section 4 presents the experimental setup, results, and a rigorous performance analysis, including comparative studies with state-of-the-art models. Finally, Section 5 summarizes the research findings and outlines potential avenues for future work.

2. Related Works

In the early stages, aircraft skin defect detection predominantly hinged upon MVI and fundamental NDT methodologies. While MVI is inherently straightforward, it is highly susceptible to inspector fatigue, subjective expertise, and the rigorous constraints of high-altitude working environments, often leading to elevated missed detection rates [26]. To bolster objectivity and reliability, techniques such as eddy current testing, UT, and digital shearography have been extensively deployed for identifying metallic surface impairments. However, these methods exhibit significant practical constraints: UT demonstrates poor adaptability to complex curvilinear geometries and relies heavily on the quality of the acoustic coupling agent; meanwhile, shearography remains extremely sensitive to ambient vibrations [27]. Most critically, these conventional technologies struggle to facilitate large-scale, real-time autonomous operations.

With the rise of convolutional neural networks (CNNs), deep learning-based object detection algorithms have achieved milestone advancements in the field of industrial quality inspection. Currently, mainstream architectures are categorized into two primary paradigms: two-stage algorithms, epitomized by Faster R-CNN, and one-stage algorithms, represented by YOLO and SSD. Although two-stage algorithms offer advantages in localization precision, their inference speeds often fall short of engineering standards in aircraft inspection scenarios with stringent real-time requirements (e.g., UAV-based hangar inspections). In contrast, the YOLO series achieves an effective balance between detection velocity and accuracy by reformulating the detection task into a unified regression problem.

However, general-purpose detection models often suffer from domain shift when directly migrated to ASD tasks. The presence of intense spectral reflections on the skin surface, irregular rivet textures, and microscopic crack features poses a significant challenge. Specifically, during successive downsampling operations, universal models tend to lose critical, subtle signals, leading to both missed detections and false alarms [23], as illustrated in Figure 2.

To address the challenges of attenuated features and complex backgrounds in ASD tasks, recent academic efforts have explored three pivotal dimensions: feature fidelity, edge reinforcement, and loss function optimization [28,29]. First, conventional networks predominantly employ strided convolutions for downsampling, which inevitably leads to the excessive compression of spatial information. To mitigate the loss of tiny object features, the recently proposed SPD-Conv mechanism utilizes space-to-depth transformations for lossless feature reassembly, establishing a new theoretical paradigm for preserving fine-grained spatial topological information [30]. Furthermore, since aircraft skin defects—such as scratches and cracks—typically manifest as linear edge signals, researchers have attempted to integrate attention mechanisms to bolster edge representation [21,31]. However, achieving precise discriminative decoupling of valid defect boundaries under intense background noise (e.g., brushed metal textures) remains a formidable challenge in current multi-scale perception research [32]. Finally, traditional IoU loss functions, such as CIoU, are largely

based on static geometric constraints. To handle the uneven distribution of sample quality in industrial datasets, the newly developed Wise-IoU (WIoU) offers a dynamic weighting scheme to mitigate the interference from low-quality samples [33].

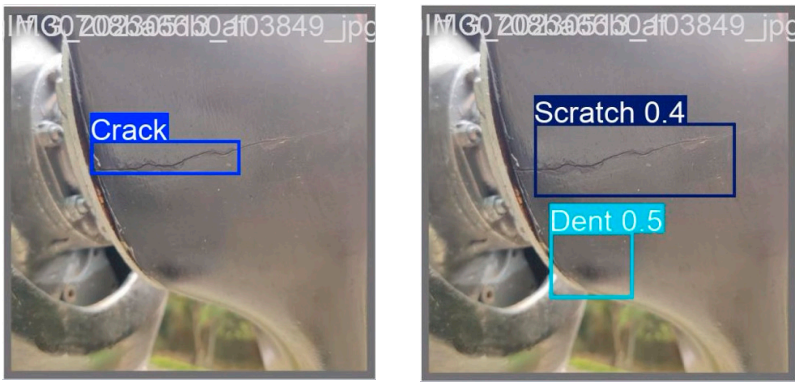


Figure 2. Detection results of the YOLOv8 model: the **left** image shows the ground truth labels, while the **right** image presents the predictions. The high inter-class similarity in local textures between different types of defects, compounded by adverse illumination effects, leads to model misclassifications.

In summary, although existing studies have improved ASD detection, several challenges remain, particularly in preserving fine-grained features, suppressing background interference, and improving localization stability. Some approaches focus mainly on feature enhancement, while others emphasize geometric constraints or sample-quality modeling. In this study, DEE-Net is developed as a YOLO11-based framework that integrates feature preservation, selective feature enhancement, and regression optimization. By combining and adapting existing techniques with the proposed ASE module, the framework aims to improve defect representation under complex industrial backgrounds. To facilitate a clear understanding of the overall design, a high-level schematic of DEE-Net is presented in Figure 3.

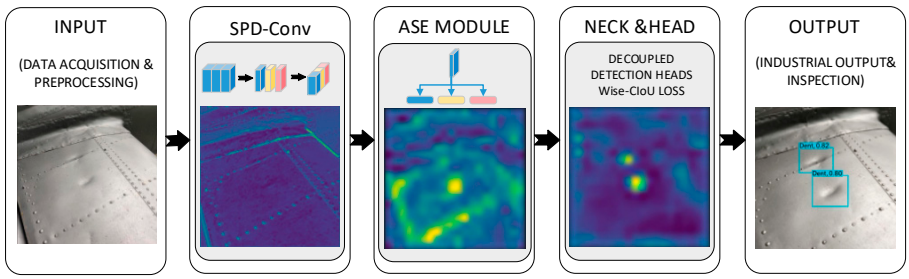


Figure 3. The overall framework of the proposed DEE-Net. It illustrates the data trajectory from raw image acquisition to final industrial inspection, highlighting the space-to-depth reconfiguration in SPD-Conv, the selective edge enhancement within the ASE module, and the dynamic regression process guided by the Wise-CIoU loss function.

3. Methods

To address the challenges of aircraft skin defect detection, including the loss of microscopic target features, background texture interference, and localization instability, this study presents DEE-Net, a deep learning framework based on an optimized YOLO11 architecture. The design of the network focuses on feature preservation and selective enhancement for fine-grained defect detection.

3.1. Architecture

The overall architecture of DEE-Net is illustrated in Figure 4. The framework integrates several components across the backbone, neck, and prediction head. In the backbone, SPD-

Conv is incorporated to perform space-to-depth transformations, which helps preserve fine-grained spatial information during downsampling. The resulting features are further processed in the feature fusion network by the ASE module, which employs parallel multi-scale paths and an edge enhancement component to improve the representation of defect-related features while reducing the influence of background textures. Finally, the prediction head integrates the Wise-CIoU loss function to adaptively adjust regression gradients for samples of different qualities, contributing to more stable localization under complex inspection conditions.

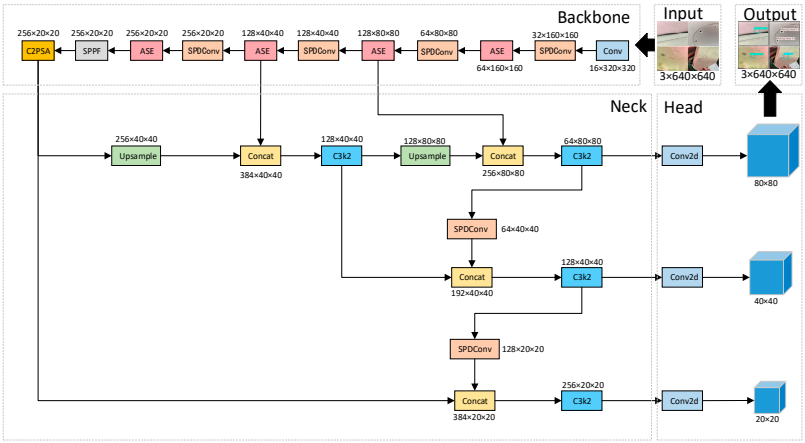


Figure 4. The detailed architectural diagram of DEE-Net. The framework consists of a backbone optimized with SPD-Conv for lossless feature reassembly, a feature fusion neck integrated with the ASE module, and a prediction head guided by the Wise-CIoU loss function for robust bounding box regression.

3.2. Feature Reassembly Mechanism

In the detection of microscopic defects smaller than 1 mm, conventional strided convolutions or pooling operations lead to the excessive compression or even the total loss of fine-grained spatial information. This research posits that subsequent feature fusion mechanisms are insufficient to rectify the irreversible information dissipation occurring within the initial shallow layers of the network. To address this, DEE-Net incorporates the SPD-Conv module. Specifically, this component is designed to address the irreversible loss of features for extremely small defects (under 1 mm) typically caused by continuous pooling operations, thereby achieving lossless downsampling and preserving critical sub-pixel information.

This mechanism achieves lossless downsampling through Space-to-Depth (SPD) transformation. Given an input tensor $X \in R^{S\times S\times C_1}$ (where S is the spatial resolution and C_1 is the number of channels), the SPD operation performs periodic sampling on the spatial dimensions through a set stride (set to 2 in this paper) to generate 4 sub-feature maps $f_{x,y}$. Its mapping relationship can be expressed as:

$$f_{x,y} = X[x : S : 2, y : S : 2, :] \quad x, y \in \{0, 1\} \tag{1}$$

Here, $x : S : 2$ indicates sampling from index x to S with a stride of 2. Through the aforementioned slicing operation, the network extracts four sub-feature maps $f_{0,0}$, $f_{1,0}$, $f_{0,1}$, $f_{1,1}$, each with dimensions of $\frac{S}{2} \times \frac{S}{2} \times C_1$. Subsequently, the module concatenates these four sub-feature maps along the channel dimension:

$$X_{spd} = \text{Concat}(f_{0,0}, f_{1,0}, f_{0,1}, f_{1,1}) \tag{2}$$

This operation downsamples the spatial dimensions by a factor of 2 while expanding the channel dimension by a factor of 4; that is, transforming from $S \times S \times C_1$ to $\frac{S}{2} \times \frac{S}{2} \times 4C_1$.

Finally, to achieve cross-channel feature interaction while preserving micro-scale damage signals, this paper introduces a 1×1 convolutional layer with a stride of 1 (as shown in Figure 5), re-mapping the number of feature channels from $4C_1$ to the target number of channels C_2:

$$X_{out} = Conv_{1\times1}\left(X_{spd}\right) \tag{3}$$

This non-destructive downsampling design ensures that the discriminative signals of micro-scale damage can flow completely into the deep network and detection head, laying a solid data foundation for subsequent accurate regression.

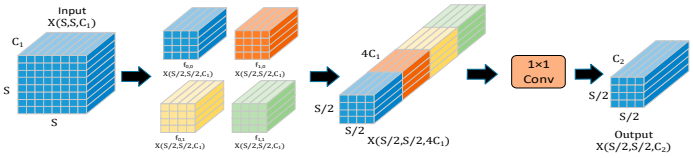


Figure 5. Schematic of the SPD-Conv module. It illustrates the complete pipeline from the input tensor through spatial sampling and reassembly, channel concatenation, to the final convolutional output. This process demonstrates the transformative logic of converting spatial details into the channel dimension without information loss.

3.3. Adaptive Selective Edge-Enhancement Module

The ASE module is the core sensing unit of DEE-Net, and its detailed structure is shown in Figure 6. This module aims to construct a feature extraction system with multi-scale representation capabilities. By simulating the human visual perception logic that first observes contours and then distinguishes details, it achieves high-sensitivity capture of microscopic damage on the skin surface.

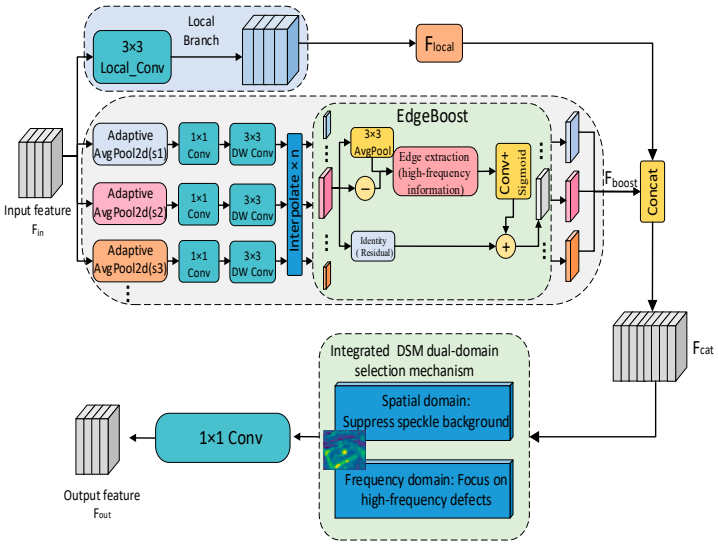


Figure 6. Architecture of the proposed ASE. After the input features are bifurcated, the local convolution branch maintains spatial details, while the multi-scale branch group ex-tracts features across multiple receptive fields; the EdgeBoost unit extracts high- frequency edge weights through a subtraction operation; the concatenated features pass through the DSM to suppress background noise, ensuring the model focuses on critical damage regions.

3.3.1. Multi-Scale Local Information Anchoring and Edge Sensitivity Reinforcement

To address the receptive field mismatch problem caused by the variable shapes of skin defects, the ASE module adopts a dual-path parallel processing strategy. Given the input

feature F_{in}, the module first utilizes n parallel adaptive average pooling layers to construct multi-scale perceptual paths:

$$F_{S_i} = DWConv_{3 \times 3}(Conv_{1 \times 1}(AdaptiveAvgPool(F_{in}, S_i))) \quad (4)$$

where $S_i \in \{S_n\}$ represents different spatial scales (only 3 scales are shown for simplicity in the diagram), aiming to actively capture local semantic signals under different physical sizes. Subsequently, to compensate for the spatial information loss caused by the pooling operation, each scale feature is first upsampled to the original resolution via bilinear interpolation, denoted as X:

$$X = Upsample(F_{Si}) \quad (5)$$

Then, the module features a built-in specialized edge enhancer (EdgeBoost, EB). For the input feature X, a 3×3 average pooling layer is first utilized for smoothing to obtain low-frequency background information, followed by the extraction of high-frequency information F_{high} containing edge details through residual subtraction:

$$F_{high} = X - AvgPool2d(X) \quad (6)$$

Finally, the high-frequency features are fed into a convolutional layer with a Sigmoid activation function to generate an edge attention weight map, which is then superimposed back onto the original features to complete the adaptive activation and enhancement of edge features:

$$F_{boost} = X + \sigma\left(Conv\left(F_{high}\right)\right) \quad (7)$$

where σ represents the Sigmoid function. This differential operation effectively extracts high-frequency signals such as fine scratches on the skin surface.

3.3.2. DSM

To strip away artifact noise such as spectral reflections on the skin surface, the ASE module introduces DSM after multi-scale feature concatenation. Although DSM was originally designed for highly challenging image restoration tasks to filter out physical degradation noise, its intrinsic logic of signal-noise separation is highly consistent with aviation inspection scenarios.

In the DEE-Net architecture, this mechanism plays the role of a feature purifier: its spatial-domain branch is dedicated to suppressing uneven background interference, while the frequency-domain branch locks in key high-frequency information such as cracks and scratches through waveform filtering. This paradigm shift from pixel reconstruction to feature selection significantly enhances the system's discriminative robustness under complex lighting environments.

Specifically, the network first performs a joint concatenation of the features F_{boost_i} enhanced by EdgeBoost from each multi-scale branch and the features F_{local} extracted by the local perception branch in the channel dimension to form the global fused feature F_{cat}:

$$F_{cat} = Concat(F_{local}, F_{boost_1}, \cdots, F_{boost_n}) \quad (8)$$

Subsequently, DSM acts as a feature purifier, simultaneously performing global evaluation and filtering of F_{Cat} in the spatial and frequency dimensions. Finally, the purified features pass through a 1×1 convolution for channel dimensionality reduction and feature reorganization, outputting the final result F_{out} of the ASE module:

$$F_{out} = Conv_{final}(DSM(F_{cat})) \quad (9)$$

Tailored to mitigate specular reflections and complex background noise on aircraft skin surfaces, this mechanism filters out visual artifacts and enhances authentic defect edges through a dual-domain selection strategy in both spatial and frequency domains. This dual-domain weighting mechanism can adaptively filter out key features highly relevant to the target task, significantly optimizing the expression accuracy of edge features while suppressing speckle noise.

3.4. Dynamic Regression Loss via Wise-CIoU

Regarding boundary box localization, this study achieves a deep coupling of geometric constraints and sample quality perception by integrating the Wise-CIoU loss. To handle the boundary ambiguity and low-quality samples inherent in manual aircraft defect labeling, a dynamic focusing coefficient is utilized to prevent the model from being biased by poor-quality training data, ensuring more stable convergence. This approach retains the comprehensive penalty terms of CIoU (L_{CIoU}), which account for overlap area, center distance, and aspect ratio, to ensure geometric continuity in localization. Its core innovation lies in the introduction of a non-monotonic focusing coefficient, γ, which performs dynamic re-weighting of the loss as

$$L_{WCIoU} = \gamma \cdot L_{CIoU}, \gamma = \frac{\beta}{\delta^{\alpha-\delta}} \quad (10)$$

Within this framework, $\beta = \frac{L_{IoU}}{\overline{L}_{IoU}}$ is defined as the outlierness used to measure the prediction quality of the samples. When a sample exhibits low outlierness (representing a high-quality sample), γ assigns a higher gradient weight to facilitate refined convergence of the model; conversely, when facing outliers with excessive outlierness, γ undergoes a non-monotonic decrease, thereby automatically attenuating the interference of abnormal samples on the gradients. This dynamic allocation mechanism guides the model toward convergence at higher IoU thresholds without increasing computational costs.

4. Experimental Results Analysis

4.1. Experimental Setup

4.1.1. Dataset

This study utilizes the publicly available aircraft skin defect dataset (ASDD). The dataset contains a total of 3007 high-resolution images sourced from the website (https://universe.roboflow.com/project-5lf3h/dataset2-69an0, accessed on 18 December 2025), covering five typical types of surface damage in aircraft inspection: Crack, Dent, Missing-head, Paint-off, and Scratch. To ensure that the experimental results are comparable with existing public benchmarks, this experiment does not adopt the conventional 8:1:1 split but strictly follows the original distribution protocol of the dataset, as shown in Table 1.

Table 1. Distribution of the ASDD.

Dataset Split	Quantity	Description
Train	2856 images	Used for deep feature learning of the model.
Val	102 images	Used for hyperparameter tuning and training monitoring.
Test	49 images	Used to evaluate the final performance of the model on unseen samples.

To address the issues of imbalanced sample distribution and complex geometric variations, this study employs a dual-strategy framework combining offline and online data augmentation. Specifically, offline augmentation is first applied to the original training data to enhance sample diversity, followed by online augmentation techniques during training.

It is important to note that all data augmentation operations are applied exclusively to the training data. The validation and testing sets, as well as the cross-validation partitioning, are strictly based on the original (non-augmented) images to prevent data leakage and ensure a fair evaluation.

To ensure a fair and reliable evaluation, special attention was given to the data partitioning process and potential similarity between training and testing samples. All data splits were performed at the level of original images, and no augmented samples were used for validation or testing. In particular, augmented variants derived from the same original image were strictly confined to the training set, ensuring that no visually similar samples appeared across different subsets. This design effectively prevents potential data leakage. For robustness evaluation, a 5-fold cross-validation protocol was adopted on the entire set of original (non-augmented) images. The dataset was randomly shuffled and partitioned into five subsets of approximately equal size. In each fold, four subsets were used for training, and one subset was used for evaluation. To preserve the data distribution, a stratified sampling strategy was applied.

To facilitate reproducibility, all experiments were conducted using fixed training settings, including a consistent number of training epochs, learning rate, and batch size across all folds. The best-performing model weights were selected based on validation performance. In addition, a fixed random seed was used during data partitioning to ensure consistent results across repeated runs.

4.1.2. Implementation Details

The experiments are implemented based on the PyTorch 2.5.1 deep learning framework using Python 3.7. All models are trained and evaluated on a workstation running Windows 11. The detailed hardware and software configurations are summarized in Table 2.

Table 2. Experimental Environment Configuration.

Category	Item	Configuration
Hardware	CPU	Intel Core i7-14700KF @ 3.60 GHz
	GPU	NVIDIA GeForce RTX 4090 (24 GB)
	RAM	64 GB
Software	OS	Windows 11
	Language	Python 3.7
	Framework	PyTorch 2.5.1
	Acceleration	CUDA 12.1

To ensure optimal convergence and facilitate a fair comparison, all models are trained from scratch without loading any pre-trained weights. This strategy is adopted to rigorously validate the feature extraction capability of the proposed method specifically for the aircraft skin defect detection task. The input images are resized to 640 $\times$ 640 pixels. The stochastic gradient descent (SGD) optimizer is employed for weight updates. The detailed training hyperparameters are summarized in Table 3.

Table 3. Training Hyperparameter Settings.

Parameter	Value	Description
Input Size	640 $\times$ 640	Resolution of input images
Epochs	300	Total number of training iterations
Batch Size	32	Number of samples per gradient update
Optimizer	SGD	Stochastic Gradient Descent

Table 3. *Cont.*

Parameter	Value	Description
Patience	50	Epochs to wait before early stopping
Momentum	0.937	Momentum factor for SGD
Weight Decay	0.0005	L2 regularization coefficient
Initial LR ($lr\ 0$)	0.01	Initial learning rate
Final LR Factor ($lr\ f$)	0.01	Factor to determine the final learning rate
Pre-trained	False	Train from scratch

4.1.3. Evaluation Metrics

To comprehensively evaluate the performance of DEE-Net in aircraft skin defect detection, we selected evaluation metrics across three dimensions: detection accuracy, localization precision, and model efficiency. Regarding detection accuracy, we employ Precision (P), Recall (R), and their harmonic mean, the F1-Score, which are defined as

$$\mathrm{P} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FP})} \tag{11}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{12}$$

Localization precision is primarily assessed using the mean average precision at an IoU threshold of 0.5 (mAP@0.5) and mAP@0.5:0.95, the latter of which averages mAP values across IoU thresholds from 0.5 to 0.95 with a step size of 0.05. This metric imposes stricter requirements on defect boundary localization and provides a more realistic reflection of the model's spatial positioning capabilities in complex industrial environments. Furthermore, model efficiency is measured by the number of the computational cost represented by GFLOPs, where the latter signifies the floating-point operations during a single forward pass and is critical for real-time detection on UAV-mounted edge terminals.

4.2. Cross-Validation Protocol and Results

To further evaluate the robustness of the proposed method and reduce the influence of a single fixed data partition, we conducted a 5-fold cross-validation experiment on all original non-augmented images. The dataset was randomly divided into five subsets of approximately equal size. In each fold, four subsets were used for training, and the remaining subset was used for evaluation. Data augmentation was applied only to the training subset in each fold, while the evaluation subset remained unchanged. This setting prevents augmented variants of the same original image from appearing in both training and evaluation subsets, thereby reducing the risk of data leakage.

The final cross-validation results are reported as the mean and standard deviation across the five folds. The overall results are shown in Table 4, while the category-level cross-validation results for Scratch are presented in Table 5.

Table 4. Overall 5-Fold Cross-Validation Results.

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall
YOLO11	0.688 ± 0.036	0.393 ± 0.032	0.719 ± 0.123	0.663 ± 0.064
DEE-Net	0.727 ± 0.052	0.396 ± 0.023	0.741 ± 0.102	0.665 ± 0.045

Table 5. 5-Fold Cross-Validation Results for the Scratch Category.

Class	mAP@0.5	mAP@0.5:0.95	Precision	Recall
Scratch	0.697 ± 0.215	0.405 ± 0.117	0.709 ± 0.219	0.594 ± 0.211

4.3. Ablation Study

Ablation experiments are conducted to systematically verify the performance contribution of each core improvement module within DEE-Net. This study utilizes YOLO11 as the baseline model and progressively incorporates the SPD-Conv, ASE, and Wise-CIoU loss function to evaluate their respective impacts. The comprehensive comparative results of the ablation study for each improvement module are presented in Table 6.

Table 6. Comparison of Ablation Study Results for Different Improvement Modules.

No.	Baseline	SPD-Conv	ASE(DSM)	Wise-CIoU	mAP@0.5	mAP@0.5:0.95	Precision	Recall	F1-Score
1	✓				0.8032	0.4937	0.7115	0.8491	0.7645
2	✓	✓			0.8137	0.4870	0.7833	0.7843	0.7767
3	✓	✓	✓		0.8325	0.4749	0.8012	0.7730	0.7702
4	✓	✓	✓	✓	0.8747	0.5180	0.8254	0.8382	0.8212

Note: The symbol ✓ denotes that the corresponding module is enabled, while a blank entry indicates that the module is not used.

Quantitative Analysis of Ablation Study

The baseline model (YOLO11) exhibits a relatively high recall rate of 84.91% in the skin detection task, demonstrating its architectural potential for candidate object discovery. However, its precision is limited to 71.15%. A fine-grained analysis of the prediction results reveals that the baseline model suffers from significant perceptual deficiencies when processing high-frequency micro-features, such as cracks. Moreover, it is susceptible to artifact interference generated by the highly reflective background of the aircraft skin, which leads to a higher rate of false positives (FPs) and significant localization deviations.

By incorporating the SPD-Conv structure to replace traditional strided convolutions, the model's precision significantly improved from 71.15% to 78.33%, while mAP@0.5 reached 81.37%. This mechanism utilizes a space-to-depth conversion strategy to reorganize sub-pixel structural features into the channel dimension, effectively alleviating the information loss inherent in conventional downsampling operations. Experimental evidence demonstrates that SPD-Conv bolsters the model's perceptual sensitivity to defect edge signals, thereby suppressing the generation of invalid predictions at the source of feature extraction.

Following the integration of the ASE module, the mAP@0.5 steadily climbed to 83.25%. The core performance gain is manifested in the enhanced capture of geometrically complex categories, such as "Dent," and defects with blurred edges. This improvement is primarily attributed to the high-frequency signal compensation provided by the internal EB component of the ASE module, as well as the filtering efficacy of the dual-domain selection mechanism (DSM). Functioning as a selective information bottleneck, DSM adaptively decouples target features from background noise, allowing the model to accurately lock onto discriminative semantics highly relevant to the detection task even within high-interference environments.

With the final introduction of the Wise-CIoU loss function, DEE-Net achieves its optimal comprehensive performance. The mAP@0.5:0.95 reaches a peak of 51.80%, representing an increase of 4.31 percentage points compared to the version incorporating only the ASE module. Notably, for the highly challenging "Crack" category, the mAP@0.5 eventually reaches 69.78%, significantly outperforming the 60.87% achieved by the baseline model. This validates the superiority of the non-monotonic focusing mechanism when handling

uneven sample quality distributions, successfully guiding the model to converge toward sub-pixel level optimal bounding boxes.

4.4. Comparison of Detection Performance

To comprehensively evaluate the performance of the proposed DEE-Net algorithm in complex defect detection tasks, we conducted a quantitative comparison with current mainstream object detection algorithms under identical experimental settings. To ensure the fairness and objectivity of the comparison, all models were trained and optimized using a unified hyperparameter configuration within the same hardware environment. The comparative models include the classic two-stage detector Faster R-CNN (ResNet50) and a series of state-of-the-art one-stage detectors: YOLOv8s, YOLOv10s, YOLO11s, and YOLO12s.

Table 7 summarizes the comprehensive performance of each model across key metrics, including Precision, Recall, mAP@0.5, mAP@0.5:0.95, GFLOPs, and FPS. Furthermore, Tables 8–13 provide a detailed breakdown of the specific detection results for each individual defect category.

Table 7. Comparison experiments.

Method	Backbone	mAP@0.5	mAP@0.5:0.95	GFLOPs	FPS	Precision	Recall
Faster R-CNN	ResNet50	0.6547	0.3083	948.18	50.7	0.4493	0.6393
YOLOv8s	-	0.7470	0.4947	6.8	97.15	0.6859	0.7085
YOLOv10s	-	0.7269	0.4916	6.5	**127.30**	0.6370	0.6814
YOLO11s (Baseline)	-	0.8032	0.4937	6.3	98.18	0.7115	**0.8491**
YOLO12s	-	0.7485	0.4347	**5.8**	84.57	0.7542	0.7040
DEE-Net (Ours)	-	**0.8747**	**0.5180**	11.5	85.23	**0.8254**	0.8382

Note: Bold values indicate the best performance for each evaluation metric.

Table 8. Detailed detection results of Faster R-CNN across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.2439	0.4348	0.3125	0.3045	0.1199
Dent	0.6957	0.7619	0.7273	0.8043	0.3267
Missing-head	0.3784	1.000	0.5490	0.8592	0.4252
Paint-off	0.4286	0.5000	0.4615	0.4722	0.0834
Scratch	0.5000	0.5000	0.5000	0.8333	0.5865
All (average)	0.4493	0.6393	0.5101	0.6547	0.3083

Table 9. Detailed detection results of YOLOv8s across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.7674	0.6087	0.6789	0.6560	0.3679
Dent	0.8895	0.7673	0.8239	0.8954	0.4822
Missing-head	0.9145	1.0000	0.9554	0.9950	0.8128
Paint-off	0.4680	0.6667	0.5500	0.5827	0.3369
Scratch	0.3901	0.5000	0.4383	0.6061	0.4739
All (average)	0.6859	0.7085	0.6893	0.7470	0.4947

Table 10. Detailed detection results of YOLOv10s across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.7070	0.4783	0.5706	0.5453	0.3240
Dent	0.7286	0.7619	0.7449	0.8197	0.4259
Missing-head	0.8426	1.0000	0.9145	0.9950	0.8106
Paint-off	0.6559	0.6667	0.6612	0.6369	0.4029
Scratch	0.2512	0.5000	0.3344	0.6379	0.4946
All (average)	0.6370	0.6814	0.6451	0.7269	0.4916

Table 11. Detailed detection results of YOLO11s across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.7494	0.6501	0.6962	0.6087	0.3557
Dent	0.7971	0.7619	0.7791	0.8390	0.4045
Missing-head	0.7587	1.0000	0.8628	0.9950	0.8100
Paint-off	0.6343	0.8333	0.7203	0.7450	0.3819
Scratch	0.6181	1.0000	0.7640	0.8283	0.5161
All (average)	0.7115	0.8491	0.7645	0.8032	0.4937

Table 12. Detailed detection results of YOLO12s across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.8089	0.5528	0.6568	0.6803	0.3051
Dent	0.9312	0.6457	0.7626	0.8339	0.3681
Missing-head	0.7909	1.0000	0.8832	0.9857	0.7764
Paint-off	0.6212	0.8216	0.7075	0.6854	0.3514
Scratch	0.6188	0.5000	0.5531	0.5575	0.3726
All (average)	0.7542	0.7040	0.7126	0.7485	0.4347

Table 13. Detailed detection results of DEE-Net across different categories.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.8809	0.6436	0.7438	0.6978	0.3585
Dent	0.8529	0.8290	0.8408	0.9250	0.4462
Missing-head	0.7931	1.0000	0.8846	0.9950	0.8233
Paint-off	0.6002	0.8333	0.6978	0.7609	0.3806
Scratch	1.0000	0.8849	0.9389	0.9950	0.5815
All (average)	0.8254	0.8382	0.8212	0.8747	0.5180

From the perspective of overall detection performance, DEE-Net achieves competitive performance across most evaluation metrics. Specifically, DEE-Net obtains an mAP@0.5 of 87.47% and an mAP@0.5:0.95 of 51.80%, corresponding to improvements of 7.15% and 2.43%, respectively, compared to the YOLO11s baseline. These results suggest that the proposed modifications contribute to improved feature extraction and multi-scale representation. While maintaining a relatively high recall (83.82%), DEE-Net also improves precision to 82.54%, compared to 71.15% for YOLO11s. This indicates that the model is better able to reduce false positives under complex background conditions. Such characteristics are beneficial for industrial inspection scenarios where both detection accuracy and reliability are important.

Compared with the two-stage detector Faster R-CNN, DEE-Net achieves higher detection performance while maintaining significantly lower computational cost. Although Faster R-CNN employs a ResNet50 backbone, it achieves an mAP@0.5 of 65.47% with a computational cost of 948.18 GFLOPs, which is substantially higher than that of DEE-Net. These results indicate that the proposed method provides a more favorable trade-off between accuracy and efficiency. This significant disparity in computational cost stems from fundamental architectural differences: the GFLOPs of Faster R-CNN are substantially higher than those of the YOLO series, primarily because the large number of proposals generated by the region proposal network (RPN) in its two-stage architecture must be processed individually by the RoI Head. Notably, the reported GFLOPs correspond to the full inference cost including all proposal processing steps. In contrast, as a one-stage detector, the computational cost of the YOLO series is primarily concentrated in the convolutional operations of the backbone and neck and is typically reported as MACs. By comparison, DEE-Net reduces computational cost by approximately 98.8% (to only 11.5 GFLOPs), while

improving inference speed by 68% (from 50.7 FPS to 85.23 FPS) and achieving a 22% increase in accuracy, fully demonstrating the dual superiority of the proposed method in both efficiency and accuracy.

A category-level analysis shows that detection performance varies among different defect types. Although the "Scratch" category achieves very high performance under the original fixed split, this result should be interpreted cautiously due to the limited number of test samples and the relatively distinguishable visual patterns of some scratch instances. The 5-fold cross-validation results reported in Section 4.2 provide a more conservative estimate, suggesting that the near-100% performance under the original split may be influenced by the specific data partition. Therefore, the conclusions of this study are mainly based on the aggregated performance across all folds and categories. The performance on underrepresented categories will be further investigated in future work using larger and more balanced datasets.

In terms of computational efficiency, the introduction of additional feature enhancement modules increases the computational cost of DEE-Net to 11.5 GFLOPs, compared to approximately 6 GFLOPs for the YOLO11 baseline. This results in a reduction in inference speed from 98.18 FPS to 85.23 FPS. Nevertheless, the processing speed remains above the commonly required threshold for real-time industrial inspection. This trade-off between computational cost and detection performance reflects the improved representation capability of the model, particularly for detecting fine-grained defects. Overall, the results suggest that DEE-Net provides a balanced compromise between accuracy and efficiency for practical inspection scenarios.

Overall, the proposed method demonstrates consistent performance improvements over baseline models while maintaining real-time capability, indicating its potential applicability in aircraft surface inspection tasks.

4.5. Robustness Evaluation Under Environmental Perturbations

To further evaluate the generalization ability of DEE-Net under degraded imaging conditions, we conducted robustness experiments using an expanded challenge test set. In practical UAV-assisted aircraft inspection, image quality may be affected by factors such as insufficient illumination and camera motion. Therefore, the challenge test set (N = 147) was constructed to simulate two common degradation factors: low-light conditions and motion blur.

The overall performance comparison is summarized in Table 14, while the detailed detection results for each specific defect category are presented in Tables 15 and 16.

Table 14. Robustness Evaluation under Simulated Environmental Perturbations (mAP@0.5).

Model	Original Test Set	Stress-Test Set (Dark/Blur)	Absolute Drop	Relative Degradation
Baseline	0.8032	0.5895	21.37%	26.6%
Ours	0.8747	0.7736	**10.11%**	**11.5%**

Note: Bold values indicate the best performance for each evaluation metric.

Table 15. Detailed detection performance of the baseline model under the combined challenge of low-light and motion-blur perturbations.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.6435	0.4203	0.5085	0.4605	0.2064
Dent	0.6719	0.5873	0.6305	0.6093	0.3021
Missing-head	0.8510	0.9048	0.8770	0.9277	0.7140
Paint-off	0.4394	0.4444	0.4419	0.4595	0.2198
Scratch	0.4259	0.6667	0.5197	0.4907	0.2759
All (average)	0.6081	0.6047	0.5955	0.5895	0.3436

Table 16. Detailed detection performance of the DEE-Net under the combined challenge of low-light and motion-blur perturbations.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.6352	0.6057	0.6201	0.5790	0.2944
Dent	0.8336	0.8744	0.8535	**0. 9076**	0.3989
Missing-head	0.7734	0.8938	0.8292	0.9184	0.7449
Paint-off	0.5647	0.7932	0.6597	**0. 6485**	0.2804
Scratch	0.8218	0.7755	0.7980	**0. 8147**	**0. 4807**
All (average)	0.7257	0.7885	0.7521	0.7736	0.4398

The results show that both the baseline YOLO11s and DEE-Net experience performance degradation under environmental perturbations. Specifically, the mAP@0.5 of YOLO11s decreases from 0.8032 to 0.5895, corresponding to an absolute drop of 21.37 percentage points. In comparison, DEE-Net decreases from 0.8747 to 0.7736, with an absolute drop of 10.11 percentage points. This smaller degradation suggests that DEE-Net maintains better detection performance under low-light and motion-blur conditions.

It should be noted that the challenge test set is generated through simulated environmental perturbations and therefore does not fully replace validation on independent real-world datasets. Nevertheless, these experiments provide additional evidence regarding the environmental robustness of the proposed method and complement the 5-fold cross-validation results reported in Section 4.2.

Furthermore, we conducted supplementary evaluations of Faster R-CNN on the challenge test set, as shown in Table 17. Compared with both Faster R-CNN and YOLO11s, DEE-Net achieves higher mAP@0.5 under the same degraded imaging conditions while maintaining real-time inference capability. These results suggest that the proposed method provides a more favorable balance between robustness and efficiency in simulated low-quality inspection scenarios.

Table 17. Detailed detection performance of the Faster R-CNN under the combined challenge of low-light and motion-blur perturbations.

Class Name	Precision	Recall	F1-Score	mAP@0.5	mAP@0.5:0.95
Crack	0.2768	0.4493	0.3425	0.3066	0.1230
Dent	0.7231	0.7460	0.7344	0.8046	0.3146
Missing-head	0.3679	0.9286	0.5270	0.7912	0.3814
Paint-off	0.4211	0.4444	0.4324	0.2997	0.1059
Scratch	0.7500	0.5000	0.6000	0.6772	0.4635
All (average)	0.5078	0.6137	0.5273	0.5758	0.2777

4.6. Visualization Analysis

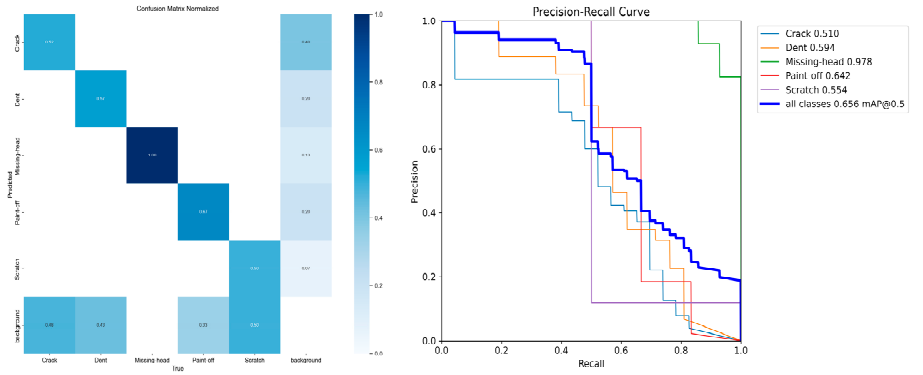
4.6.1. Visual Comparison of Detection Results

To provide a visual validation of the effectiveness of DEE-Net, this study conducted inference tests on aircraft skin surface images captured under complex backgrounds. DEE-Net demonstrates a superior continuous detection capability compared to the baseline, particularly for shallow surface scratches. It effectively avoids the fragmentation phenomenon where a single long scratch is erroneously identified as multiple disjointed targets.

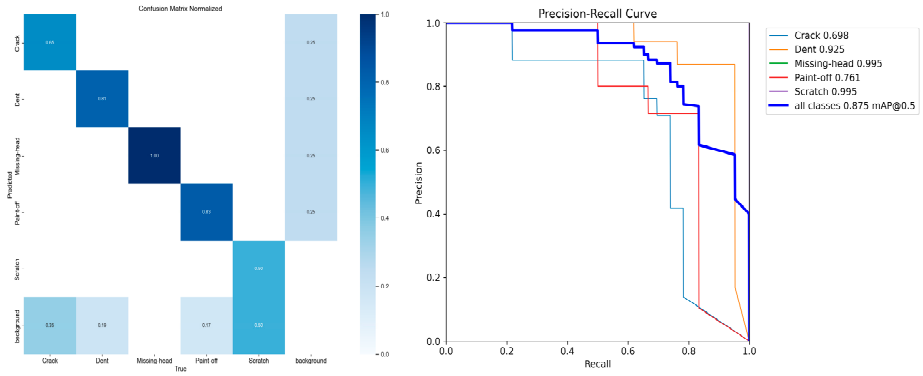
As shown in the visualization comparison in Figure 7, DEE-Net exhibits a significant technological generational gap in terms of robustness when handling extremely small targets. In the detection results of the baseline model, there is a prominent risk of missing detections for subtle scratches and early-stage micro-cracks. This is primarily due to the loss

of pixel-level details during the traditional convolutional downsampling process, which leads to the "information annihilation" of critical defect features in deeper network layers.

In stark contrast, leveraging the introduced SPD-Conv lossless downsampling structure, DEE-Net can effectively preserve the spatial topological information of the original image. Even when faced with damage at extremely small scales that is difficult to discern with the naked eye, the model still achieves precise feature triggering and capture.



(**a**)



(**b**)

Figure 7. (**a**) Normalized confusion matrix and PR curve of YOLO11s (Baseline); (**b**) Normalized confusion matrix and PR curve of DEE-Net (Ours).

4.6.2. Heatmap-Based Interpretability Analysis

To further validate the feature extraction advantages of DEE-Net from the perspective of visual interpretability, this study employs heatmap analysis to compare the feature attention regions of the proposed model with those of the baseline algorithm, as illustrated in Figure 8. The results indicate that the baseline model (YOLO11s) exhibits a diffuse feature response distribution when processing subtle skin defects. Due to its limited capability in capturing sub-pixel features, the baseline is highly susceptible to interference from metallic textures and complex lighting conditions. This leads to significant recognition biases and lower confidence levels, particularly for defects such as "Dent".

In contrast, the activated regions in the DEE-Net heatmaps are highly consistent with the ground-truth physical contours of the defects. The model demonstrates a superior ability to penetrate background noise and precisely lock onto the core representations of "Paint-off" and "Dent." This remarkable visual focusing capability provides strong empirical evidence for the high-frequency edge signal compensation provided by the ASE module, as well as the filtering efficacy of the DSM mechanism in suppressing irrelevant information across both spatial and frequency domains. Consequently, the model is able to accurately decouple discriminative features from complex industrial backgrounds, fun-

damentally underpinning its exceptional performance in detecting extremely fine defects such as “Crack” and “Scratch”.

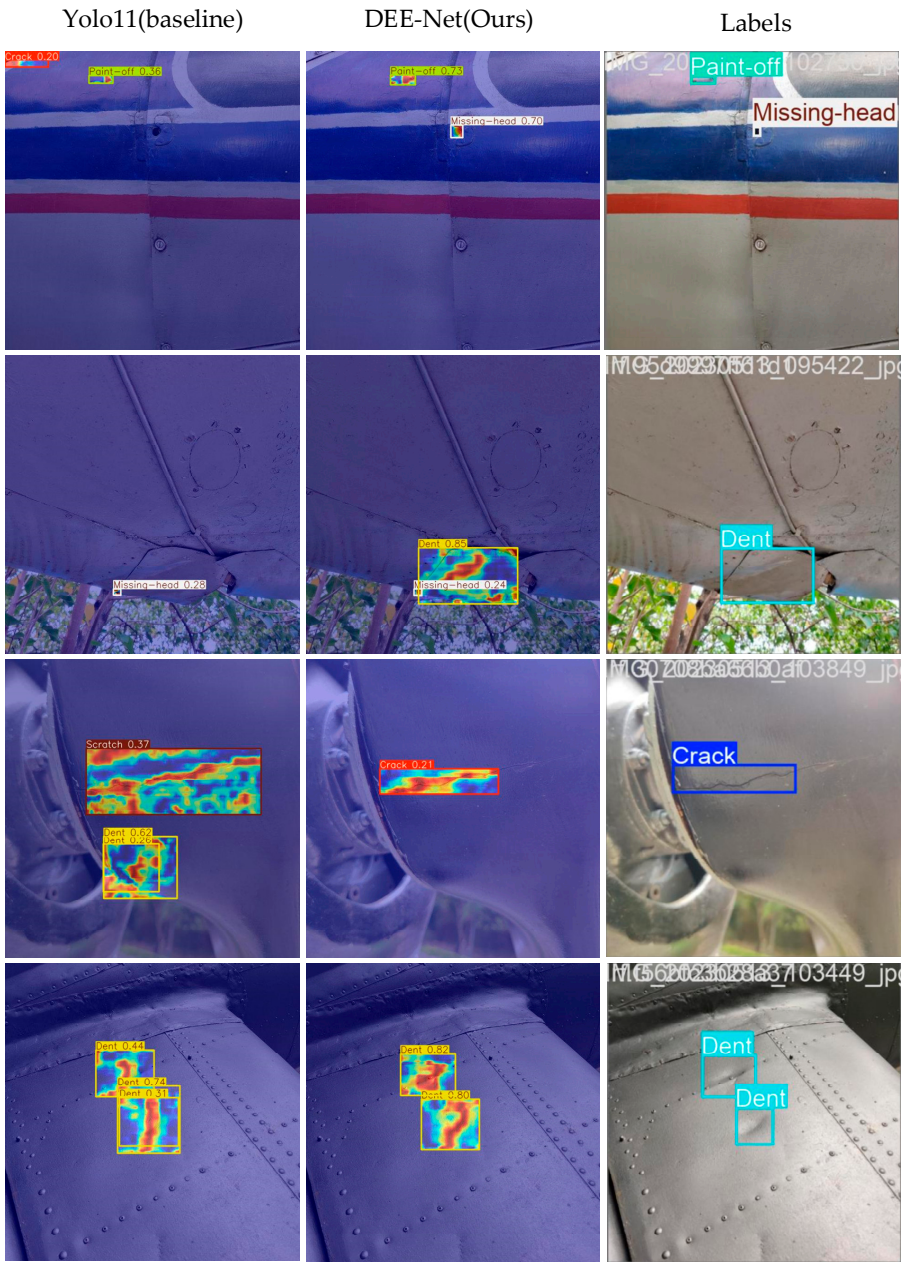


Figure 8. Heatmap visualization comparison between YOLO11 (Baseline) and DEE-Net (Ours). The columns from left to right represent the YOLO11 (Baseline) detection results, the DEE-Net (Ours) results, and the ground truth labels, respectively. Warmer colors represent higher activation intensity, while cooler colors represent lower activation intensity.

5. Conclusions

Addressing the urgent requirements for structural integrity monitoring and flight safety assurance of aircraft in complex operating environments, this paper proposes DEE-Net, a high-efficiency algorithm for aircraft skin defect detection. The core contributions of this methodology are manifested through several integrated dimensions: first, the introduction of the SPD-Conv lossless feature reorganization mechanism utilizes space-to-depth conversion to reduce feature scales while preserving critical sub-pixel information, effectively addressing the information annihilation of micro-defects during the down-

sampling process at the source. Building upon this, the designed ASE module leverages high-frequency differential logic to compensate for the attenuation of subtle signals—such as cracks and scratches—inherent in traditional convolutions. Furthermore, the integration of the DSM enables the model to simultaneously focus on discriminative features in both the spatial and frequency domains, significantly suppressing interference from metallic reflections and complex background noise. To further refine the training process, the Wise-CIoU loss function, incorporating a non-monotonic focusing mechanism, was introduced to dynamically adjust gradient weights for samples of varying quality, thereby achieving a deep coupling of geometric constraints and sample quality awareness while markedly improving bounding box regression accuracy and convergence stability.

Comparative experiments show that DEE-Net achieves an mAP@0.5 of 87.47% on the ASDD dataset, representing an improvement over the YOLO11s baseline. The results suggest that the proposed method provides enhanced capability for detecting fine-grained defects. Heatmap analysis based on Grad-CAM further indicates that the model tends to focus on defect-related regions, supporting its effectiveness in capturing relevant features. In terms of efficiency, DEE-Net achieves an inference speed of 85.23 FPS while maintaining competitive detection performance. This suggests that the proposed method has the potential to meet real-time requirements in UAV-assisted inspection scenarios, providing a reasonable balance between accuracy and computational cost.

Despite the excellent performance of DEE-Net at this stage, its robustness under complex and variable climatic conditions—such as rain, fog, and intense glare—remains to be further validated. Consequently, subsequent research will focus on several critical avenues to enhance the practical utility of the model. Specifically, we aim to investigate cross-scene generalization capabilities by employing transfer learning and domain adaptation techniques to ensure detection consistency across different aircraft models and varying illumination environments. Furthermore, research into knowledge distillation and extreme lightweighting will be conducted to explore advanced model compression techniques, with the goal of achieving lower-latency deployment on power-constrained embedded inspection devices without compromising precision. In addition, the research focus will shift from qualitative recognition toward the deep quantitative assessment of defects, such as the precise measurement of defect length and area, to provide more valuable structural damage evaluation data for aircraft maintenance decision-making.

In conclusion, this study presents a deep learning-based approach for aircraft surface defect detection. The experimental results indicate that the proposed DEE-Net achieves improved detection performance while maintaining real-time inference capability. Within the scope of the current dataset and experimental settings, the proposed method shows potential for reducing manual inspection workload and supporting automated inspection processes in aircraft maintenance.

Author Contributions: X.W. conceptualization, methodology, supervision, writing—original draft; M.L. methodology, software, visualization, validation, writing—original draft; Y.L. conceptualization, methodology, supervision; J.Q. methodology, software, visualization, validation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Research Funds of Key Laboratory for Civil Aviation Data Governance and Decision Optimization (No. CAMICCADGDO-2025-(01-01)), the Funds of Sichuan Provincial Engineering Research Center of Smart Operation and Maintenance of Civil Aviation Airports (No. JCZX2024ZZ03), and the Henan Province Key R&D Special Fund (No. 251111242100).

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here ASDD (https://universe.roboflow.com/project-5lf3h/dataset2-69an0, accessed on 18 December 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Plastropoulos, A.; Bardis, K.; Yazigi, G.; Avdelidis, N.P.; Droznika, M. Aircraft skin machine learning-based defect detection and size estimation in visual inspections. *Technologies* **2024**, *12*, 158. [CrossRef]
2. Li, H.; Hou, Z.; Wang, C.; Lu, T. A review of automated aircraft skin defects detection. *Measurement* **2025**, *262*, 120019. [CrossRef]
3. Suvittawat, N.; Kurniawan, C.; Datephanyawat, J.; Tay, J.; Liu, Z.; Soh, D.W.; Ribeiro, N.A. Advances in aircraft skin defect detection using computer vision: A survey and comparison of YOLOv9 and RT-DETR performance. *Aerospace* **2025**, *12*, 356. [CrossRef]
4. Drury, C.G. *Human Factors in Aircraft Inspection*; Defense Technical Information Center: Fort Belvoir, VA, USA, 2000.
5. Xiong, J.; Li, P.; Sun, Y.; Xiang, J.; Xia, H. An Aircraft Skin Defect Detection Method with UAV Based on GB-CPP and INN-YOLO. *Drones* **2025**, *9*, 594. [CrossRef]
6. Wronkowicz-Katunin, A. A brief review on NDT&E methods for structural aircraft components. *Fatigue Aircr. Struct.* **2018**, *2018*, 73–81. [CrossRef]
7. Hung, Y.; Wang, J.; Hovanesian, J. Technique for compensating excessive rigid body motion in nondestructive testing of large structures using shearography. *Opt. Lasers Eng.* **1997**, *26*, 249–258. [CrossRef]
8. Yang, L.; Hung, Y. Digital shearography for nondestructive evaluation and application in automotive and aerospace industries. *J. Hologr. Speckle* **2004**, *1*, 69–79. [CrossRef]
9. Zhao, Q.; Dan, X.; Sun, F.; Wang, Y.; Wu, S.; Yang, L. Digital shearography for NDT: Phase measurement technique and recent developments. *Appl. Sci.* **2018**, *8*, 2662. [CrossRef]
10. Connolly, L.; Garland, J.; O'Gorman, D.; Tobin, E.F. Deep-learning-based defect detection for light aircraft with unmanned aircraft systems. *IEEE Access* **2024**, *12*, 83876–83886. [CrossRef]
11. Yiannakides, D.; Drikakis, D.; Sergiou, C. Quantifying human-centric uncertainty in aircraft maintenance. *Aeronaut. J.* **2025**, *129*, 3402–3445. [CrossRef]
12. Yasuda, Y.D.; Cappabianco, F.A.; Martins, L.E.G.; Gripp, J.A. Automated visual inspection of aircraft exterior using deep learning. In Proceedings of the Conference on Graphics, Patterns and Images (SIBGRAPI), Gramado, Brazil, 18–22 October 2021; pp. 173–176.
13. Feng, M.; Xu, Y.; Dai, W.; Luo, H. A UAV-Based Measurement System for Aircraft Skin Defect Detection Using a State-Space Model Approach. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2546212. [CrossRef]
14. Liao, K.-C.; Lau, J.; Hidayat, M. An innovative aircraft skin damage assessment using you only look once-version9: A real-time material evaluation system for remote inspection. *Aerospace* **2025**, *12*, 31. [CrossRef]
15. Chu, Z.; Weng, G.; Yu, L. Real-time Industrial Surface Defect Detection Based on Lightweight Convolutional Neural Networks. *Artif. Intell. Mach. Learn. Rev.* **2024**, *5*, 36–53.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
17. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: New York, NY, USA, 2016; pp. 779–788.
18. Rengasamy, D.; Morvan, H.P.; Figueredo, G.P. Deep learning approaches to aircraft maintenance, repair and overhaul: A review. In *Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018*; IEEE: New York, NY, USA, 2018; pp. 150–156.
19. Tao, X.; Zhang, D.; Ma, W.; Liu, X.; Xu, D. Automatic metallic surface defect detection and recognition with convolutional neural networks. *Appl. Sci.* **2018**, *8*, 1575. [CrossRef]
20. Liu, Y.; Dong, J.; Li, Y.; Gong, X.; Wang, J. A UAV-based aircraft surface defect inspection system via external constraints and deep learning. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5019315. [CrossRef]
21. Huang, B.; Ding, Y.; Liu, G.; Tian, G.; Wang, S. ASD-YOLO: An aircraft surface defects detection method using deformable convolution and attention mechanism. *Measurement* **2024**, *238*, 115300. [CrossRef]
22. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; IEEE: New York, NY, USA, 2023; pp. 7464–7475.
23. Cha, Y.J.; Choi, W.; Suh, G.; Mahmoudkhani, S.; Büyüköztürk, O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Comput. Aided Civ. Infrastruct. Eng.* **2018**, *33*, 731–747. [CrossRef]
24. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [CrossRef]
25. Cui, Y.; Ren, W.; Cao, X.; Knoll, A. Focal network for image restoration. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023*; IEEE: New York, NY, USA, 2023; pp. 13001–13011.

26. See, J.E. *Visual Inspection: A Review of the Literature*; Sandia National Laboratories: Albuquerque, NM, USA, 2012.
27. Hung, Y.; Chen, Y.S.; Ng, S.; Liu, L.; Huang, Y.; Luk, B.; Ip, R.; Wu, C.; Chung, P. Review and comparison of shearography and active thermography for nondestructive evaluation. *Mater. Sci. Eng. R Rep.* **2009**, *64*, 73–112. [CrossRef]
28. Liang, Y.; Han, Y.; Jiang, F. Deep learning-based small object detection: A survey. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence, Tianjin, China, 18–21 March 2022*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 432–438.
29. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011.
30. Sunkara, R.; Luo, T. No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Grenoble, France, 19–23 September 2019*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 443–459.
31. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
32. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.
33. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. *arXiv* **2023**, arXiv:2301.10051.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.