Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería

Tech Science Press

https://doi.org/10.32604/rimni.2026.076107

ARTICLE

SAT-YOLO: A Multi-Module Enhanced Network for High-Precision Small Target Detection in Industrial Paper Cups

Shuaian Zhang[1,2], **Ziliang Peng**[3], **Jian Huang**[2,*], **Ruiyin Chen**[3] **and Daojin Dai**[3]

[1]School of Mechanical Engineering, Zhejiang Sci-Tech University, Hangzhou, China
[2]Zhejiang Provincial Innovation Center of Advanced Textile Technology, Shaoxing, China
[3]Zhejiang New Debao Machinery Co., Ltd., Wenzhou, China

*Corresponding Author: Jian Huang. Email: 2024220503220@mails.zstu.edu.cn

Received: 14 November 2025; Accepted: 18 March 2026; Published: 06 November 2026

ABSTRACT: During industrial paper-cup manufacturing, surface defects caused by oil leakage can degrade appearance and may pose health risks, making accurate and deployable inspection essential. However, small defects are easily overwhelmed by complex backgrounds, exhibit weak semantic cues, and suffer from unstable localization. To address these issues, we propose SAT-YOLO, featuring a SKBAttention module for defect-focused spatial—channel enhancement to suppress background interference, an AAConv module to reduce fine-grained information loss during downsampling, and a TSCODE decoupled head to alleviate classification—regression conflicts and stabilize localization. On a real paper-cup defect dataset, SAT-YOLO improves overall mAP50 by 4.7% and by 9.2% on the most challenging small-spot defects, with consistent gains in Precision, Recall, and mAP50–95. These improvements reduce missed subtle defects and strengthen quality control on high-speed production lines, at the cost of a moderate complexity increase and slightly lower inference speed.

KEYWORDS: Paper cups; small target detection; SKBAttention; AAConv

1 Introduction

As a widespread throwaway consumer product, paper cups have become a crucial vessel for the food and beverage sector due to their convenience, sanitation, and environmental benefits [1]. The quality of paper cups is intrinsically linked to consumer health and user experience, particularly given the growing concern regarding food safety. Surface defects in paper cups, such as spots, stains, and wrinkles, can adversely affect product aesthetics and brand image, potentially raising safety issues [2,3]. Therefore, achieving high precision in defect detection of paper cups is critical to ensure a safer and superior product experience for consumers while maintaining product quality, strengthening production process automation, and bolstering enterprise competitiveness.

The conventional manual inspection techniques are challenging to match with the pace of high-speed production lines, leading to several industrial goods remaining undetected promptly, significantly hindering production efficiency. Moreover, the manual inspection is influenced by the energy and focus of the employees [4]. It is susceptible to leakage detection and misdetection, challenging the assurance of product quality and demanding a substantial workforce, hence increasing operational expenses for organizations. In addition, while the conventional machine vision inspection method mitigates some limitations of manual inspection, it heavily depends on the amalgamation of various disciplines, including optics, mechanics,

Copyright © 2027 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

electronics, software, and algorithms, which causes elevated costs for system development and upkeep [5]. It may represent a considerable economic strain for small and medium-sized organizations.

In recent years, deep learning technology has advanced considerably in visual inspection and has increasingly become the predominant method for industrial defect detection. Contemporary approaches in deep learning-based inspection categorize detection algorithms into two distinct paradigms: single-stage detection algorithms (e.g., YOLO series) [6,7] and two-stage detection algorithms (e.g., Faster R-CNN) [8].

Single-stage detection algorithms possess a straightforward structure and rapid detection speed, enabling direct prediction of target category and position. The YOLO series algorithms have emerged as an essential component in industrial defect detection, owing to their lightweight architecture, adaptability for optimization, and real-time detection capabilities [9,10]. For example, Guo et al. [11] proposed YOLO-IRS by incorporating global-attention modeling and enhanced feature representation, achieving higher small-target accuracy with competitive speed. Yuan et al. [12] presented YOLO-SSW through higher-resolution feature fusion, improved fine-detail extraction, and training reweighting, which better preserves small-target details and improves robustness under noisy samples.

Among recent YOLOv8 to YOLOv12 variants, the dominant trend is to improve the overall accuracy and speed balance by strengthening feature aggregation in the backbone and neck, refining detection head design, and optimizing training and inference strategies. However, paper cup surface defect inspection has distinctive properties: defect cues are extremely small and often low in saliency, easily overwhelmed by complex background textures, and prone to being suppressed or lost during downsampling; meanwhile, coupled detection heads may further weaken classification and localization consistency for these subtle defects. In contrast, SAT-YOLO explicitly targets these paper-cup specific challenges by (i) enhancing faint, small defect cues while filtering out distracting texture responses via SKBAttention, (ii) retaining discriminative fine details during downsampling with AAConv to reduce feature degradation, and (iii) improving classification and localization consistency for subtle defects through the task-aware TSCODE head, which incorporates SCE and DPE to mitigate head coupling interference, thereby making the novelty clear at an early stage.

In summary, this study makes the following contributions: (1) It proposes SAT-YOLO for small-defect detection in industrial paper-cup inspection, where SKBAttention enhances weak defect cues under complex textures, AAConv preserves fine-grained details during downsampling, and the TSCODE decoupled head alleviates the conflict between classification and localization. (2) It establishes a real industrial paper-cup defect dataset with explicit defect categories and small-target criteria, improving the reproducibility of spot, stain, and wrinkle detection. (3) Through ablation, comparative, and cross-domain experiments, it verifies that SAT-YOLO improves overall mAP50 by 4.7% and spot-defect mAP50 by 9.2% over YOLOv5s, while results on RSOD and HIT-UAV further confirm the generalizability of the proposed architecture and its key modules.

The remainder of this paper is organized as follows. Section 2 reviews related studies on attention mechanisms and downsampling methods for small-object detection. Section 3 presents the improved YOLOv5-based framework, including the baseline model, the SKBAttention module, the AAConv module, the TSCODE decoupled head, and the overall SAT-YOLO architecture. Section 4 describes the paper-cup defect dataset, experimental settings, evaluation metrics, ablation studies, comparative experiments, and qualitative analysis. Section 5 reports the cross-domain generalization experiments on RSOD and HIT-UAV. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

2.1 Attention Mechanisms

In small-object detection, attention mechanisms can effectively improve feature representation by enhancing target-related responses and suppressing background interference. For example, Zhang et al. [13] introduced the Spatial Context Aware Module (SCAM) in FFCA-YOLO, which improves the discriminability of small objects by emphasizing salient spatial regions and informative channels. However, its feature enhancement remains insufficiently adaptive to large scale variations, which may limit precise localization of small targets. Liang et al. [14] further proposed the Separated and Enhancement Attention Module (SEAM), which decouples spatial and channel attention and improves localization stability under cluttered backgrounds and partial occlusion by combining cross-scale spatial interaction with cross-level channel reweighting. However, this method still mainly relies on feature reweighting and lacks explicit adaptive receptive-field modeling.

Although these methods have achieved certain progress in feature enhancement, they still exhibit limitations in adapting to scale variations and modeling receptive fields adaptively. Motivated by this gap, we propose the Selective Kernel and Bilateral Attention (SKBA) module, which explicitly models multi-receptive-field responses through multi-kernel branches and a maximum-response aggregation mechanism, thereby improving the representation of weak defects at different scales. To provide a clearer comparison of recent and relevant studies, Table 1 summarizes representative attention-based methods in terms of their core strategies, strengths, limitations, and application areas.

Table 1: Meta-analysis of representative attention-based studies for small-object detection.

Reference	Method	Strengths	Limitations	Application Area
Zhang et al. [13]	SCAM	Better small-object discrimination	Weak scale adaptability	Remote sensing
Liang et al. [14]	SEAM	Robust localization under clutter/occlusion	Lacks receptive-field adaptation	Aerial detection
This work	SKBA	Stronger weak-defect enhancement	Higher complexity and cost	Paper-cup defect detection

2.2 Downsampling Convolution

In modern object detectors, downsampling is a key operation for extracting hierarchical features, enlarging the receptive field, and constructing multi-scale representations, thereby exerting an important influence on detection performance. For example, Sunkara and Luo [15] proposed SPD-Conv, which replaces conventional strided convolution with a space-to-depth rearrangement followed by non-strided convolution, thereby preserving local details related to small objects while reducing resolution. However, its improvement mainly lies in structural preservation and remains limited in selectively suppressing texture-dominant background interference. Qiao et al. [16] introduced Switchable Atrous Convolution (SAC), which adaptively combines convolutions with different dilation rates to enlarge the receptive field while maintaining spatial resolution, thereby enhancing contextual modeling for small objects. However, SAC mainly strengthens context aggregation rather than explicitly enhancing weak defect signals, which may restrict its effectiveness in industrial scenarios with low-contrast defects and repetitive textures.

Despite these advances, existing methods still face difficulty in simultaneously preserving fine-grained defect details, suppressing background texture interference, and enhancing weak target cues. As a result, subtle defect information may still rapidly decay during feature compression or be overwhelmed by complex background textures in industrial defect detection. Motivated by this gap, we propose the Adaptive Attention Convolution (AAConv) module, which integrates local spatial preservation, attention-guided reweighting, and residual fusion to enhance weak small-defect cues more effectively during downsampling. To provide a clearer comparison of recent and relevant studies, Table 2 summarizes representative downsampling-related methods in terms of their core strategies, strengths, limitations, and application areas.

Table 2: Meta-analysis of representative downsampling-related studies for small-object detection.

Reference	Method	Strengths	Limitations	Application Area
Sunkara and Luo [15]	SPD-Conv	Better detail preservation	Weak background suppression	Small-object detection
Qiao et al. [16]	SEAM	Better contextual modeling	Weak subtle-defect enhancement	General detection
This work	SKBA	Stronger weak-defect preservation	Higher complexity and cost	Paper-cup defect detection

3 Improved YOLOv5 Model

3.1 YOLOv5

Although newer YOLO variants (e.g., YOLO11/YOLO12/YOLOv13) have been proposed recently, we adopt YOLOv5s as the baseline primarily because it provides a strong and widely validated foundation for small-object detection. First, YOLOv5s offers a favorable accuracy—efficiency trade-off with a lightweight design, making it practical for detecting tiny defects while remaining deployable on resource-constrained industrial hardware. Second, its backbone—neck–head pipeline is highly modular and has proven effective for preserving and aggregating multi-scale features (especially high-resolution cues that are critical for small targets), which allows our proposed modules to be integrated with minimal changes and enables a clearer, more controlled analysis of their contributions to small-defect detection. Third, many newer versions involve substantial architectural and training-pipeline changes; using YOLOv5s as a controlled baseline helps isolate the impact of our small-target–oriented designs and improves the interpretability of module-level improvements. The detailed architecture of YOLOv5s is shown in Fig. 1.

3.2 The Selective Kernel and Bilateral Attention Module

Due to their diminutive feature dimensions, small target defects are readily overshadowed or lost by background information during feature extraction. To address the drawback, the Selective Kernel and Bilateral Attention (SKBAttention) module is designed into the backbone network of YOLOv5. The SKBAttention module is an enhancement to SKNet (Channel Attention Module), which is an attention mechanism that exclusively addresses channel variations without accounting for spatial positional variation. Nevertheless, small targets typically exhibit greater dependence on spatial structure; thus, channel-level enhancement may not achieve the intended outcomes for small target detection, particularly when the target's features are spatially sparse. The SKBAttention module implements a spatial attention mechanism derived from the SKNet module, enhancing the model's sensitivity to the spatial distribution of targets by dynamically

adjusting the weights of the spatial dimensions in the feature maps, thereby significantly increasing the detection precision of defects in small targets.

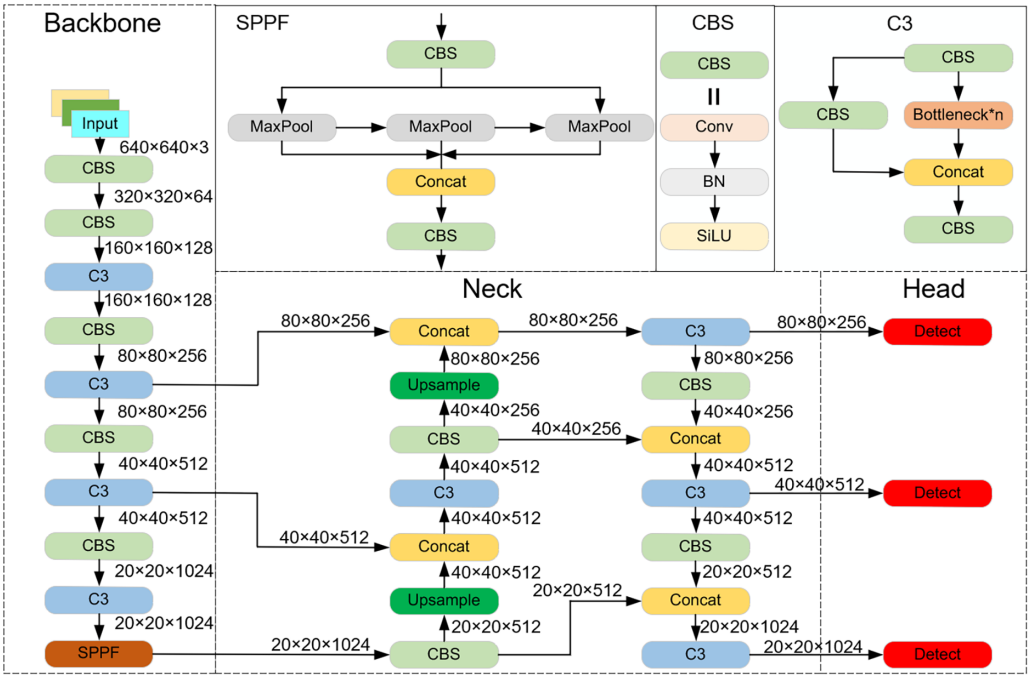


Figure 1: The structure of the YOLOv5s.

Fig. 2 illustrates the structure of the SKBAttention module, wherein the input feature map X undergoes initial processing through several convolutional branches. Each branch possesses a convolution kernel of different sizes (e.g., 3×3, 5×5), resulting in an output that contains multi-scale features. The convolution operation for 3×3 is shown in Eq. (1), while the convolution operation for 5×5 is shown in Eq. (2):

$$\widetilde{\mathcal{F}} : X \rightarrow \widetilde{U} \tag{1}$$

$$\widehat{\mathcal{F}} : X \rightarrow \widehat{U} \tag{2}$$

where, $\widetilde{U}$ and $\widehat{U}$ respectively denote the feature maps obtained by applying 3×3 and 5×5 convolutions to the input feature map X.

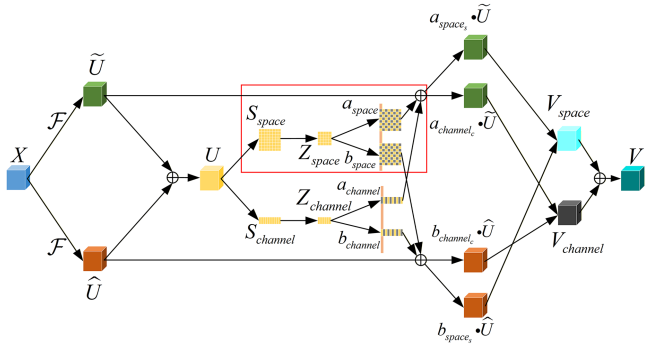


Figure 2: The structure of the SKBAttention module.

The output features from all convolutional branches are summed up element-wise to produce the fused feature map U. The precise formula is shown in Eq. (3):

$$U = \widetilde{U} + \widehat{U} \tag{3}$$

Subsequently, spatial attention computation and channel attention computation are executed for U, respectively.

Spatial attention is derived by initially compressing the channel dimension of the fused feature map U to produce a spatial-level representation S_{space}. In short, the sth element of S_{space} is computed by compressing the channel dimension C. The precise formula is shown in Eq. (4):

$$S_{space_s} = \mathcal{F}_{gp}(U_s) = \frac{1}{C}\sum_{k=1}^{C} U_s(k) \tag{4}$$

where, $\mathcal{F}_{gp}$ denotes the operation for compressing the channel dimension, U_s denotes the $s\text{-}th$ element of U, and S_{space_s} denotes the $s\text{-}th$ element of S_{space}.

Subsequently, the fully connected layer compresses S_{space} into a low-dimensional representation Z_{space}. The precise formula is shown in Eq. (5):

$$Z_{\text{space}} = \mathcal{F}_{fc}\left(S_{\text{space}}\right) = \delta\left(B\left(W_{S_{\text{space}}}\right)\right) \tag{5}$$

where $Z_{space} \in R^{\text{J}\times\text{L}\times 1}$, $\text{J} \times \text{L} < \text{H} \times \text{W}$, δ represents the ReLU function and B represents the BatchNorm function, W is a weight matrix.

The spatial weights for each branch are generated by inputting Z_{space} again into the fully connected layer corresponding to each branch. The weights are modified to conform to the shape of $[bs, 1, h, w]$ and stacked to $[k, bs, 1, h, w]$, indicating the spatial attention of the k branches. The weights are normalized using Softmax to obtain the spatial attention weights a_{space} and b_{space}, guaranteeing that the weights constitute a probability distribution over the branches. a_{space_s} and b_{space_s} are defined in Eq. (6):

$$a_{space_s} = \frac{e^{A_s z}}{e^{A_s z} + e^{B_s z}}, \quad b_{space_s} = \frac{e^{B_s z}}{e^{A_s z} + e^{B_s z}} \tag{6}$$

where a_{space} and b_{space} denote the soft attention vectors of $\widetilde{U}$ and $\widehat{U}$, respectively, a_{space_s} is the sth element of a_{space}, and b_{space_s} is the sth element of b_{space}.

Subsequently, the feature tensors, following convolution in the initial phase, are weighted and summed using spatial attention weights to get the spatial-enhanced feature V_{space}. The precise formula is shown in Eq. (7):

$$V_{space} = a_{space_s} \cdot \widetilde{U} + b_{space_s} \cdot \widehat{U} \tag{7}$$

where $\widetilde{U}$ and $\widehat{U} \in R^{\text{H}\times\text{W}\times\text{R}}$, a_{space} and $b_{space} \in R^{\text{H}\times\text{W}\times 1}$.

Channel attention is derived by initially compressing the spatial dimension of the fused feature map U to produce a channel-level representation $S_{channel}$. In short, the cth element of $S_{channel}$ is computed by compressing the spatial dimension $\text{H} \times \text{W}$. The precise formula is shown in Eq. (8):

$$S_{\text{channel}_c} = \mathcal{F}_{gp}\left(U_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} U_c(i, j) \tag{8}$$

where, $\mathcal{F}_{gp}$ denotes the operation for compressing the spatial dimension, U_c denotes the $c\text{-}th$ element of U, and $S_{channel_c}$ denotes the $c\text{-}th$ element of $S_{channel}$.

Subsequently, the fully connected layer compresses $S_{channel}$ into a low-dimensional representation $Z_{channel}$. The precise formula is shown in Eq. (9):

$$Z_{\text{channel}} = \mathcal{F}_{fc}\left(S_{\text{channel}}\right) = \delta\left(B\left(W_{S_{\text{channel}}}\right)\right) \tag{9}$$

where $Z_{\text{channel}} \in R^{d\times 1}$, $d < C$.

The channel weights for each branch are generated by inputting $Z_{channel}$ again into the fully connected layer corresponding to each branch. The weights are modified to conform to the shape of $[bs, c, 1, 1]$ and stacked to $[k, bs, c, 1, 1]$ to represent the channel attention of the k branches. The weights are normalized using Softmax to obtain the channel attention weights $a_{channel}$ and $b_{channel}$, guaranteeing that the weights constitute a probability distribution over the branches. $a_{channel_c}$ and $b_{channel_c}$ are defined in Eq. (10):

$$a_{channel_c} = \frac{e^{A_c z}}{e^{A_c z} + e^{B_c z}}, \quad b_{channel_c} = \frac{e^{B_c z}}{e^{A_c z} + e^{B_c z}} \tag{10}$$

where $a_{channel}$ and $b_{channel}$ denote the soft attention vectors of $\widetilde{U}$ and $\widehat{U}$, respectively, $a_{channel_c}$ is the cth element of $a_{channel}$, and $b_{channel_c}$ is the cth element of $b_{channel}$.

Subsequently, the feature tensors $\widetilde{U}$ and $\widehat{U}$, following convolution in the initial phase, are weighted and summed using channel attention weights to get the channel-enhanced feature $V_{channel}$. The precise formula is shown in Eq. (11):

$$V_{channel} = a_{channel_c} \cdot \widetilde{U} + b_{channel_c} \cdot \hat{U} \tag{11}$$

The spatial enhancement feature V_{space} and the channel enhancement feature $V_{channel}$ are summed to produce the final output feature V. The precise formula is shown in Eq. (12):

$$V = V_{space} + V_{channel} \tag{12}$$

Through strategic optimization of spatial-channel information integration, the network's discriminative power in feature representation is substantially strengthened. This methodology successfully alleviates the challenge of small target features suffering from background-induced information suppression, consequently elevating the recognition precision for defects characterized by small target features.

3.3 The Adaptive Attention Convolution Module

In backbone network feature extraction procedures, standard convolutional downsampling operations progressively compress feature map dimensions through stride amplification mechanisms. This substantial reduction in feature map size may cause small target defects, which count for only a small portion of the original image, to merge into a more extensive background area. Consequently, this can result in a severe loss of fine-grainedness of the small target defects, challenging the model's ability to detect these defects in subsequent layers efficiently. To mitigate the limitations of conventional convolutional downsampling methods in detecting small target defects, the Adaptive Attention Convolution (AAConv) module downsampling method is designed in place of the conventional convolutional downsampling method of the backbone network to improve the detection accuracy of small target defects. The AAConv module enhances the ability of the feature expression of small targets while minimizing background interference by fine-processing the small target features, hence improving the precision of small target detection.

The structure of the AAConv module is illustrated in Fig. 3. First, the input feature map X undergoes Maxpool to diminish its size. After the Maxpool operation, the feature map is split into two segments based on the channels: X_1 (1/4 channel segment of X) and X_2 (3/4 channel segment of X). This split strategy routes a small portion of channels through Slice and SKBAttention to preserve and enhance high-frequency details of small targets, while directing the majority of channels through ADown to provide stable downsampling and contextual suppression. Processing of the main branch is divided by X_1. Initially, X_1 is processed by $Slice$ to enhance the local information extraction ability of the feature of the small target. Subsequently, the SKBAttention module is employed to augment the global information extraction ability of the feature of the small target. Finally, the $Sigmoid$ function is used to produce the weights of X_1. Processing of the residual branch is divided by X_1. X_1 undergoes initial processing via $Slice$, subsequently through the SKBAttention module, and ultimately, the processed residual branch is weighted and summed with the weights generated by the $Sigmoid$ function of the main branch. Processing segment X_2: X_2 is executed by the Adown module. The processed X_1 and X_2 are connected along the channel dimension to produce the final output. The module's multi-scale feature extraction, residual connection, and attention mechanism enhance the AAConv module's flexibility and efficiency in addressing small targets inside intricate environments.

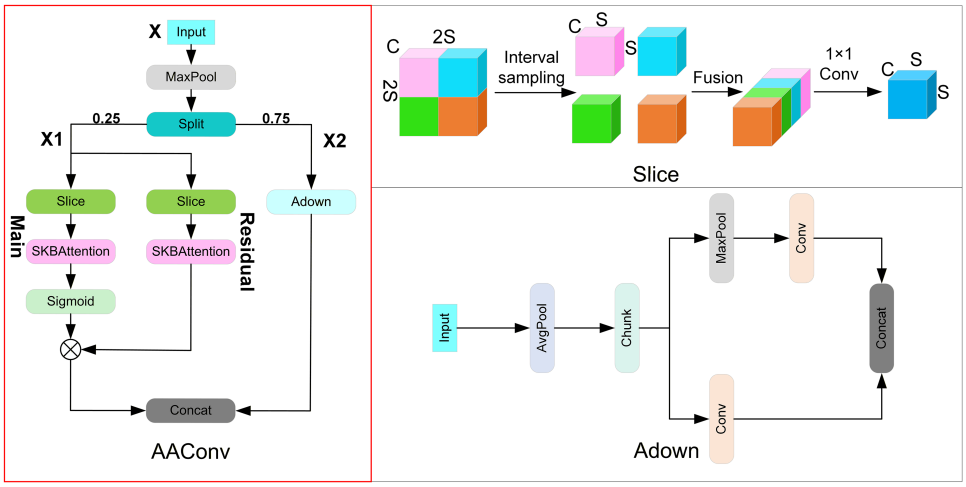


Figure 3: The structure of the AAConv module.

3.4 The Task-Specific Context Decoupling Module

Object detection is typically categorized into two primary tasks: classification and regression. The classification task focuses on category recognition and semantic discrimination, which generally relies more on higher-level features, whereas the regression task focuses on precise localization and bounding-box refinement, which benefits more from lower-level details and boundary cues. In YOLOv5, classification and regression share the same feature input and are optimized jointly, which may introduce task interference: the gradients and feature preferences of the two tasks can be inconsistent, and this inconsistency becomes more pronounced for small objects whose semantic cues are weak and whose boundaries are easily blurred after downsampling. As a result, shared-head learning may lead to unstable classification confidence and imprecise localization for small targets.

To mitigate this issue, we adopt the TSCODE decoupling head [17], which explicitly separates classification and regression into task-specific branches to reduce mutual interference. TSCODE consists of Semantic Context Encoding (SCE) and Detail Preserving Encoding (DPE), where SCE emphasizes discriminative context cues for classification, and DPE focuses on preserving fine-grained details for accurate localization. By

decoupling the two tasks and providing task-aligned feature encoding, TSCODE helps maintain more stable semantic discrimination and more precise boundary regression, thereby improving detection robustness and accuracy for small targets.

3.5 SAT-YOLO

The proposed SAT-YOLO is designed to overcome the limitations of YOLOv5 in small-target detection, and its overall architecture is illustrated in Fig. 4. Specifically, the model improves the detection pipeline in a progressive manner, covering feature enhancement, detail preservation, and task prediction. First, the SKBAttention module is introduced to strengthen discriminative spatial-channel responses, thereby enhancing weak defect features under complex backgrounds. Based on this, the AAConv module is incorporated into the downsampling process to reduce the loss of fine-grained information and preserve subtle target cues during feature compression. Furthermore, the original coupled head of YOLOv5 is replaced with the TSCODE decoupled head, where the SCE and DPE modules are adopted to alleviate the conflict between classification and localization. In this way, the features enhanced by SKBAttention and preserved by AAConv can be more effectively exploited in the TSCODE decoupled head, thereby enabling more accurate defect recognition and more precise localization. Through the coordinated interaction of these three modules, SAT-YOLO achieves more reliable small-target detection performance.

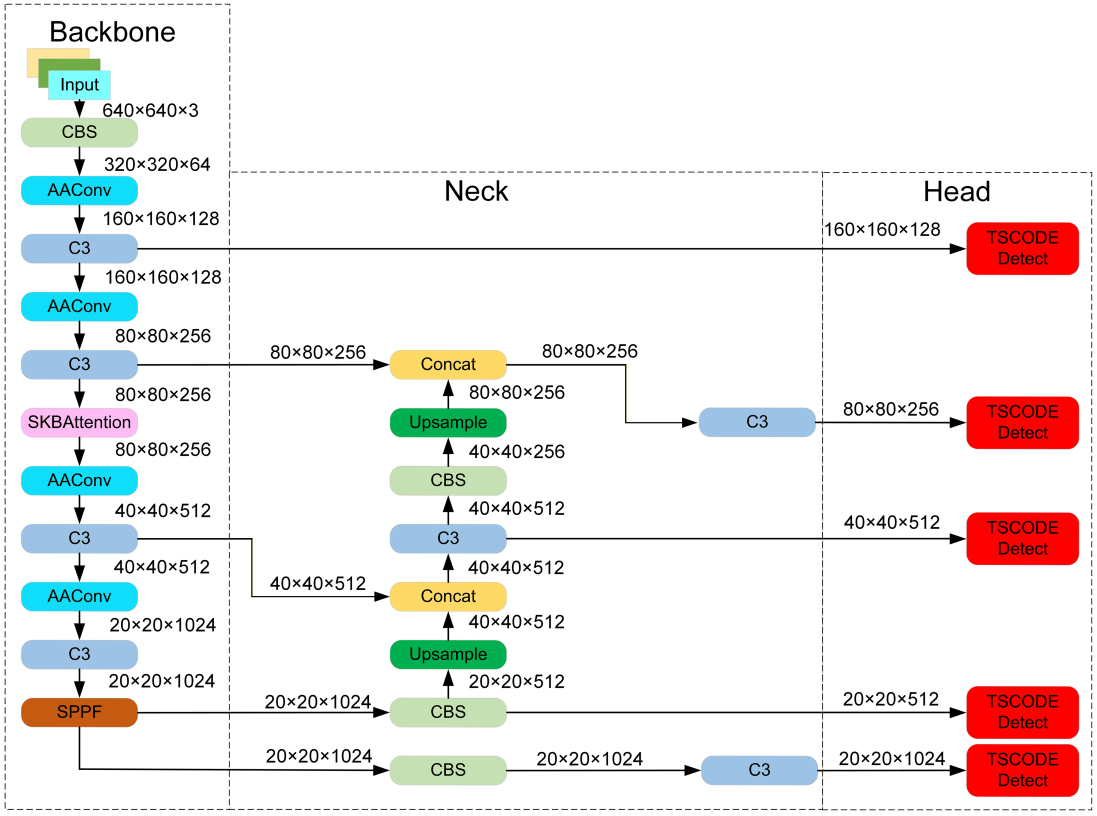


Figure 4: The structure of the SAT-YOLO model.

4 Experimental Results and Analyses

4.1 Dataset Preparation

4.1.1 Paper-Cup Datasets

Due to the absence of a publicly accessible dataset for defect detection in paper cups, this paper constructs a dataset of 375 original images of paper cup defects, utilizing actual images obtained from industrial settings. Defects are divided into three common types based on their defective features: spot, stain, and wrinkle, with schematic diagrams of each type of defect shown in Fig. 5. To improve annotation consistency and reproducibility, we further define these defect categories using both appearance and size criteria. Specifically, spots are small, dark (blackish) dot-like defects with annotated area ranging from 4^2 to 96^2; stains are dark (blackish) smudge-like defects with a larger area range from 4^2 to 418^2; and wrinkles are white (bright) fold-like defects with annotated area ranging from 24^2 to 186^2.

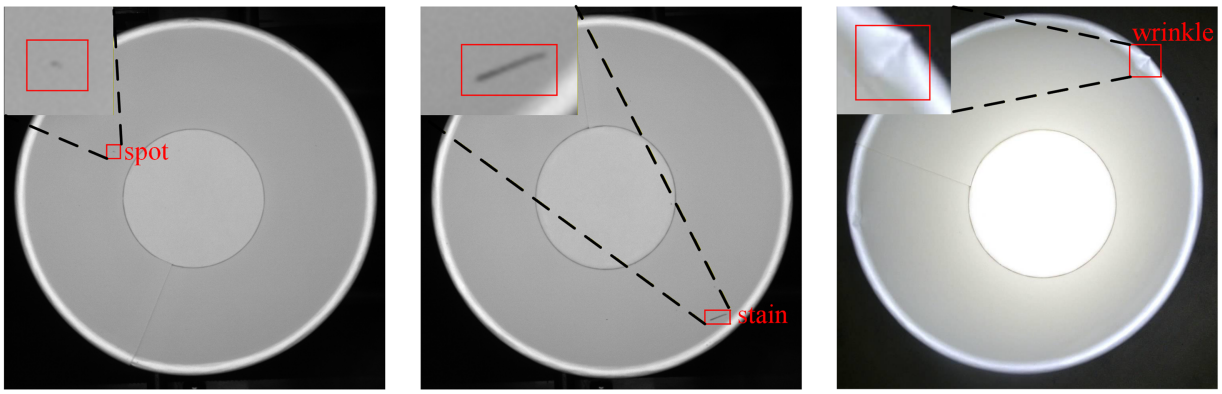


Figure 5: Schematic diagram of the defects of various types of paper cups.

In addition, to facilitate a clear evaluation of small-target performance, we explicitly define "small targets" in our dataset as defect instances whose bounding-box area ($w \times h$, in pixels) falls within 0 to 32^2. The paper cup defects were labeled using the labelImg tool, subsequently expanding the dataset to 2250 images by data augmentation methods, including translation, noise injection, cutout, rotating, scaling, and zoom. To guarantee adequate training data for the model and to enable better model evaluation, this paper randomly divides the dataset into train, valid, and test sets in an 8:1:1 ratio.

Fig. 6 illustrates the statistical amounts of distinct defect types in the dataset following data augmentation, while Fig. 7 illustrates the aspect ratio (width/height) distribution of distinct defect types. Figs. 6 and 7 illustrate that spots are the primary defect in paper cups. The smaller aspect ratio of the spots within the image complicates industrial detection; therefore, detecting small target defects (spots) in paper cups is highly significant.

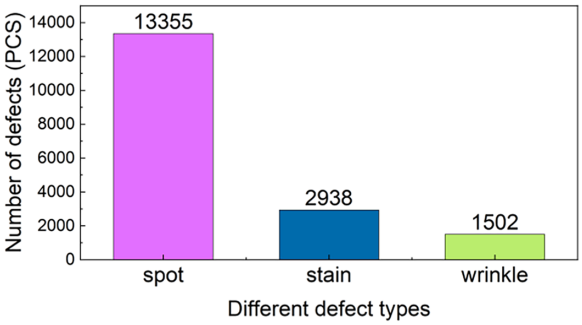
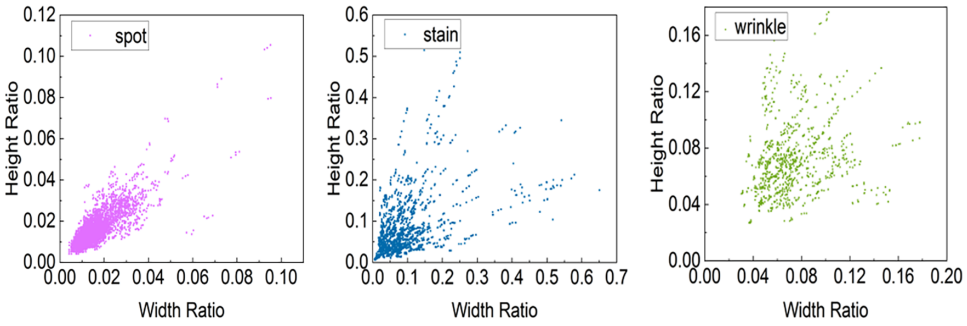


Figure 6: Number of defects in various types of paper cups.



(a) spot **(b)** stain **(c)** wrinkle

Figure 7: Distribution of aspect ratio for various types of paper cup defects.

4.1.2 Other Datasets

This paper employs the publicly available RSOD dataset to evaluate the generalization capacity of the SAT-YOLO model in small target detection tasks [18]. The open-access dataset contains 976 annotated aerial images featuring four distinct object categories: aircraft, oil tank, overpass, and playground. Experimental validation focuses specifically on these predefined classes to assess model performance across different remote sensing scenarios. This paper utilizes the HIT-UAV dataset to evaluate the generalization capabilities of the SKBAttention and AAConv modules in infrared UAV detection [19]. The open-access repository contains 2898 thermal infrared aerial images annotated with five predefined object classes: Person, Car, Bicycle, OtherVehicle, and DontCare. Experimental validation specifically focuses on cross-category detection performance across these distinct high-altitude UAV targets.

4.1.3 Domain Difference Discussion

Our industrial paper-cup defect images are close-range RGB shots dominated by repetitive textures and subtle low-contrast defects, whereas RSOD consists of large-scale aerial RGB scenes with significant scale variation and complex backgrounds, and HIT-UAV provides thermal infrared UAV imagery with weaker textures and higher noise. Despite these domain gaps, all three datasets share the core challenges of small-object detection: small target-to-image ratio, fine-detail loss during downsampling, and severe background interference. Therefore, the proposed SKBAttention and AAConv modules, which explicitly enhance spatial—channel discriminability and preserve fine-grained information during downsampling, together with the TSCODE decoupled head that alleviates classification—localization conflict, can provide

domain-agnostic improvements. The consistent gains on RSOD and HIT-UAV verify the architecture-level generalization of SAT-YOLO across industrial and aerial imagery.

4.2 Experimental Setting and Training Parameters

All experiments' environment configurations and training parameters are consistent to avoid external interference and experimental errors. The environment configurations are shown in Table 3, and the training parameters are shown in Table 4.

Table 3: Training environment.

Parameter	Configuration
Operating system	Windows 11
GPU	RTX4090 24 G
Programming language	Python 3.10
Framework	PyTorch 2.3.0
GPU accelerated	CUDA 12.1

Table 4: Training parameters.

Training Parameter	Value
Image Size	640 × 640
Epochs	300
Batch size	16
Initial learning rate	0.01
Workers	8
Momentum	0.937
Optimizer	SGD

4.3 Indicators of Evaluation

In this experiment, Precision (P), Recall (R), Mean Average Precision at an IoU confidence threshold of 0.5 (mAP50), and Mean Average Precision at an IoU confidence threshold ranging from 0.5 to 0.95 (mAP50-95) are used as evaluation metrics for model detection accuracy. The specific equations are shown in Eqs. (13)~(16):

Precision (P) is calculated as:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

Recall (R) is calculated as:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

Mean Average Precision (mAP) is calculated as:

$$AP = \int_0^1 P(R)dR \tag{15}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{16}$$

In Eqs. (13)~(16): True Positives (*TP*) represents the number of correctly classified positive instances, while False Positives (*FP*) denotes negative samples that are incorrectly assigned to the positive class. Conversely, False Negatives (*FN*) characterizes positive-class examples that are erroneously misclassified as negative instances. N is the total number of categories; and AP_i is the ith category's average precision.

4.4 Ablation Experiments

4.4.1 Ablation Experiments with the SAT-YOLO Model

This paper assesses the detection performance of each improved module in the SAT-YOLO model through ablation experiments, with results shown in Table 5.

Table 5: Results of ablation experiments with the SAT-YOLO model.

YOLOv5s	SKB	AAConv	TSCODE	P/%	R/%	mAP 50/%	mAP 50%–95/%	mAP50 (spots)/%	Params/M	FPS
✓				87.4 ± 0.1	75.1 ± 0.1	76.9 ± 0.2	47.6 ± 0.2	40.9 ± 0.1	7.02	112.6
✓	✓			87.5 ± 0.2	76.6 ± 0.3	78.6 ± 0.1	51.2 ± 0.1	43.4 ± 0.2	8.44	86.3
✓		✓		89.3 ± 0.2	77.0 ± 0.2	79.4 ± 0.1	52.1 ± 0.2	45.4 ±0.1	8.67	104.5
✓			✓	89.8 ± 0.2	78.5 ± 0.1	80.5 ± 0.2	53.2 ± 0.3	47.5 ± 0.3	16.17	67.5
✓	✓	✓		89.8 ± 0.3	76.9 ± 0.1	79.6 ± 0.1	52.8 ± 0.2	48.5 ± 0.1	9.34	96.2
✓	✓		✓	89.2 ± 0.1	79.0 ± 0.1	80.9 ± 0.1	54.3 ± 0.3	49.5 ± 0.1	17.78	65.8
✓		✓	✓	89.7 ± 0.2	77.9 ± 0.2	81.1 ± 0.2	55.4 ± 0.1	49.3 ± 0.2	18.23	63.2
✓	✓	✓	✓	89.9 ± 0.1	79.7 ± 0.1	81.6 ± 0.1	55.6 ± 0.3	50.1 ± 0.2	19.38	53.8

Table 5 indicates that the introduction of the SKBAttention module, the AAConv module, and the TSCODE decoupling head module enhances the mAP50 of the SAT-YOLO model on the paper cup defect dataset by 1.7%, 2.5%, and 3.6%, respectively, in comparison to the original YOLOv5s model. The mAP50 for small target defect (spots) detection was enhanced by 2.5%, 4.5%, and 6.6% for the three modules, respectively, indicating the effectiveness of each improved module. Notably, the TSCODE head yields the largest gain for the most challenging small-target category (spots), improving mAP50(spots) by 6.6%, compared with 2.5% and 4.5% brought by SKBAttention and AAConv, respectively. This observation supports that decoupling classification and localization is especially beneficial for paper-cup defects, where weak semantics and ambiguous boundaries make the task conflict more severe; thus, TSCODE provides a more suitable optimization pathway than a coupled head for small defect detection.

Moreover, the combined SKBAttention–AAConv module yields a clear improvement, increasing the mAP50 for spot by 7.6%, which strongly demonstrates that AAConv reduces information loss during down-sampling and preserves fine-grained representations of small targets, while SKBAttention further enhances target responses and suppresses background interference via multi-scale selective attention; together, they synergistically improve the detectability of small-target defects.

Finally, The SAT-YOLO model, utilizing all of these modules, enhanced the mAP50 by 4.7% for all defect detection in paper cups and by 9.2% for small target defects (spots), thereby significantly enhancing the detection precision of small target defects in paper cups.

To assess the model performance for paper cup detection, this paper visualizes the comparative analysis through multiple dimensions: firstly, Fig. 8 presents a graphical comparison of the mAP50 with epochs curves for the SAT-YOLO model and the original YOLOv5s model; secondly, Fig. 9 shows the comparison of P, R, mAP50, mAP50-95, and mAP50 (spots) evaluation metrics of the two models after training in a bar chart; further, Figs. 10 and 11 plot comparative PR curves for spots and all defects, respectively.

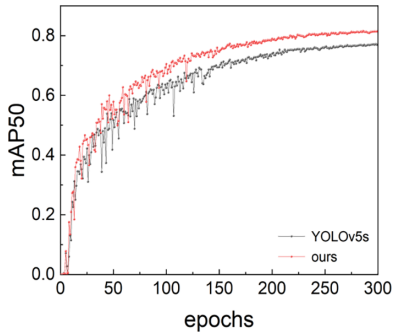


Figure 8: Comparison of mAP50 curves between the SAT-YOLO model and the original YOLOv5s model.

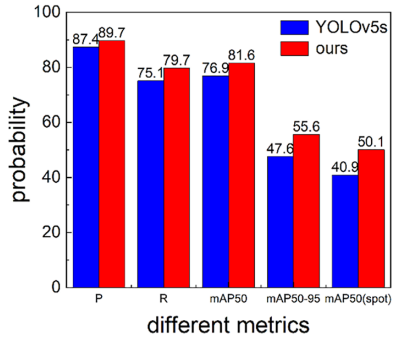


Figure 9: Comparison of evaluation metrics between the SAT-YOLO model and the original YOLOv5s model.

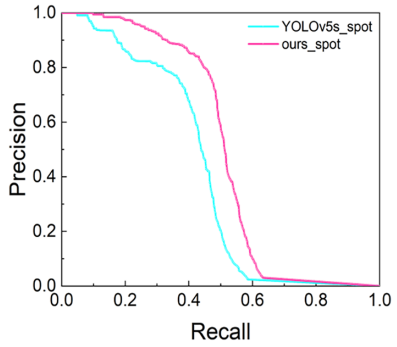


Figure 10: Comparison of P–R curves between the SAT-YOLO model and the original YOLOv5s model for small target defect (spots) detection.

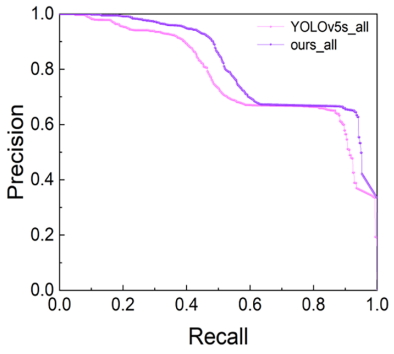


Figure 11: Comparison of P–R curves between the SAT-YOLO model and the original YOLOv5s model for all defect detection.

Fig. 8 illustrates that as training epochs increase, the mAP50 of the SAT-YOLO model consistently surpasses that of the YOLOv5s model, indicating superior training and enhanced final detection precision of the SAT-YOLO model. Fig. 9 illustrates that the P, R, mAP50, mAP50-95, and mAP50 (spots) metrics of the SAT-YOLO model surpass those of the YOLOv5s model, indicating an enhancement in the detection accuracy of the SAT-YOLO model across all cases. As evidenced by the precision-recall (PR) curves in Figs. 10–11, the SAT-YOLO model exhibits a larger area under the curve (AUC) relative to the YOLOv5 model, thereby empirically confirming enhanced detection performance for both small target defect categories (spots) and all defect identification tasks.

4.4.2 Ablation Experiments with the SKBAttention Module

This paper conducts ablation experiments to validate the enhancement impact of the SKBAttention module on the paper cup dataset, with results shown in Table 6.

Table 6: Results of ablation experiments of the SKBAttention module.

Model	P/%	R/%	mAP50/%	mAP50–95/%	mAP50 (spots)/%	Params (M)	GFLOPs
YOLOv5s	87.4	75.1	76.9	47.6	40.9	7.02	15.8
YOLOv5s+SK	87.5	75.9	77.9	50.7	42.1	8.42	33.6
YOLOv5s+SKB	87.8	76.6	78.6	51.2	43.4	8.44	33.8

Table 6 shows that the SKBAttention module effectively improves the detection of small target defects (spots) in paper cups with almost no increase in parameters and computation.

In order to visualize the improvement of the SKBAttention module compared to the SKNet module, this paper visualizes the comparison of the mAP50 between the two modules during the training process of the paper cup dataset so as to clearly present the improvement effect of the SKBAttention module. The variation of mAP50 with epochs of the two modules is shown in Fig. 12.

As shown in Fig. 12, the training effect of the SKBAttention module is significantly better than the SKNet module.

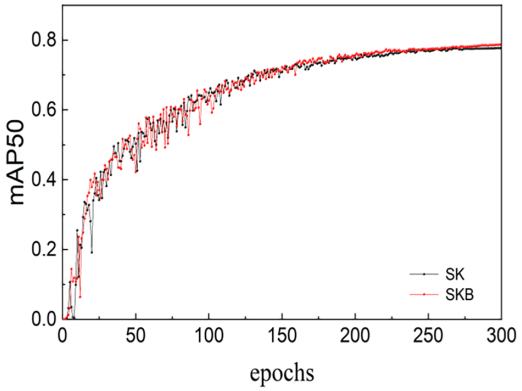


Figure 12: Comparison of mAP50 curves between the SKBAttention and the SKNet modules.

4.5 Comparative Experiments

4.5.1 Comparative Experiments with the SAT-YOLO Model

This paper conducts comparative experiments to validate the effectiveness of the proposed SAT-YOLO. Specifically, SAT-YOLO is compared with representative detectors, including YOLOv5s, YOLOv8s, YOLOv9s, YOLOv10s, YOLOv11s, YOLOv12s, RT-DETR [20], Faster R-CNN, and SSD [21]. For fairness, all models are trained and evaluated on the same dataset split under an identical training protocol, including the same image resolution, batch size, number of epochs, optimizer strategy, data augmentation, and evaluation script. To avoid bias introduced by model-specific tuning, we adopt the official implementations and their recommended default hyperparameters for each baseline, without performing additional per-model hyperparameter search, while ensuring that all methods are trained from the same type of initialization. Detection performance is reported in terms of P, R, mAP50, and mAP50–95. The results are presented in Table 7.

Table 7: Comparison results of different network models.

Model	P	R	mAP50	mAP 50–95	mAP50 (spots)	FPS
YOLOv5s	87.4	75.1	76.9	47.6	40.9	112.6
YOLOv8s	82.6	73.1	74.8	52.0	34.9	98.3
YOLOv9s	77.0	63.1	66.7	38.6	25.3	87.7
YOLOv10s	77.1	68.0	70.2	47.3	29.9	75.5
YOLOv11s	80.7	69.7	72.2	47.3	31.6	103.7
YOLOv12s	77.9	65.2	67.7	38.5	30.0	68.5
RT-DETR	80.7	68.7	70.0	35.6	40.7	46.9
Faster R-CNN	78.3	57.0	62.4	31.1	\	52.1
SSD	76.5	56.1	60.4	29.8	\	54.6
Ours	89.9	79.7	81.6	55.6	50.1	53.8

Table 7 illustrates that the mAP50 for all defects in paper cups detected by the SAT-YOLO model is 81.6%, whereas the mAP50 for small target defects (spots) in paper cups is 50.1%. The mAP50 for all defects in paper cups detected by the SAT-YOLO model surpassed those of all defects detected by YOLOv5s, YOLOv8s, YOLOv9s, YOLOv10s, YOLOv11s, YOLOv12s, RT-DETR, Faster R-CNN, and SSD, with respective differences of 4.7%, 6.8%, 14.9%, 11.4%, 9.4%, 13.9%, 11.6%, 19.2%, and 21.2%. The mAP50 for small target

defects (spots) in paper cups detected by the SAT-YOLO model surpassed those detected by YOLOv5s, YOLOv8s, YOLOv9s, YOLOv10s, YOLOv11s, YOLOv12s, and RT-DETR, exhibiting rises of 9.2%, 15.2%, 24.8%, 20.2%, 18.5%, 20.1%, and 9.4%, respectively. Although SAT-YOLO does not have an advantage in FPS compared with other mainstream models, it still meets the speed requirements for industrial inspection. Comprehensive comparative experimental results show that the SAT-YOLO model possesses considerable advantages in detecting small target defects (spots) in paper cups.

4.5.2 Comparative Experiments with the SKBAttention Module

To verify the detection effect of the SKBAttention module on the paper cup dataset, this paper compares it with different attention mechanisms, and the comparison results are shown in Table 8.

Table 8: Comparison results of different attention mechanisms.

Module	P/%	R/%	mAP 50/%	mAP 50–95/%	mAP50 (spots)/%
SE [22]	88.6	75.9	77.2	48.2	41.5
CA [23]	87.1	76.0	77.6	48.9	41.7
SimAM [24]	85.8	75.1	77.6	48.1	41.4
ECA [25]	86.1	75.0	76.9	48.6	40.8
SA [26]	87.7	74.6	77.2	49.4	40.4
EMA [27]	86.5	75.1	77.7	49.4	42.2
SK [28]	87.5	75.9	77.9	50.7	42.1
SKB	87.8	76.6	78.6	51.2	43.4

Table 8 illustrates that the SKBAttention module designed in this paper markedly surpasses other attention mechanisms for small target detection in paper cups.

4.5.3 Comparative Experiments with the AAConv Module

To verify the detection effectiveness of the AAConv module on the paper cup dataset, this paper compares it with other convolutional modules, and the comparison results are presented in Table 9. As shown in Table 9, the AAConv module proposed in this paper significantly outperforms the other convolutional modules in small-target defect detection on paper cups. Furthermore, to validate the choice of the "split" ratio (1 vs. 3), we conducted additional ratio-comparison experiments, and the results are reported in Table 10. The results indicate that the 1:3 split ratio achieves the best performance for small-target detection.

Table 9: Comparison results of different convolutions.

Module	P/%	R/%	mAP 50/%	mAP 50–95/%	mAP50 (spots)/%
Conv	87.4	75.1	76.9	47.6	40.9
ODConv [29]	88.2	74.6	77.1	47.3	40.7
AKConv [30]	68.5	47.2	48.6	26.1	10.6
SPDConv	85.4	75.3	76.6	47.8	39.5
SAConv	86.6	75.5	77.3	48.1	39.8
RFAConv [31]	87.3	75.1	77.3	48.2	41.9
AAConv	89.3	77.0	79.4	52.1	45.4

Table 10: Comparative experiments of AAConv with different split ratios.

Ratio	P	R	mAP 50	mAP50–95	mAP (spot)
3:1	88.2	75.7	78.1	50.5	44.0
1:2	89.2	77.4	78.9	51.8	44.9
2:1	88.2	75.9	78.2	50.8	44.2
1:1	89.0	76.4	78.7	51.4	44.7
1:3	89.3	77.0	79.4	52.1	45.4

4.6 Visualisation and Analysis

To interpret the quantitative gains more intuitively, we visualise representative detection results of several mainstream detectors on the paper-cup defect dataset (Fig. 13), including YOLOv5s, YOLOv8s, YOLOv11s, RT-DETR and the proposed SAT-YOLO. As illustrated in Fig. 13a–e, the compared baselines exhibit varying degrees of missed detections, particularly for tiny spot defects under low-contrast or texture-dominant backgrounds, and their confidence scores are sometimes relatively conservative. In contrast, SAT-YOLO detects these defects more consistently across the presented cases and generally produces higher-confidence predictions, which qualitatively aligns with the quantitative improvements. This advantage mainly benefits from the synergy of SKBAttention, AAConv, and TSCODE, which collaboratively enhance defect-relevant feature representation, preserve small-scale details during downsampling, and improve the consistency and stability of classification and localisation; consequently, SAT-YOLO delivers more reliable predictions under complex background textures and low-contrast conditions. We also note that SAT-YOLO is not perfect: in Fig. 13c, it still misses a defect, indicating that when the semantic cues of small targets are weak, SAT-YOLO does not adapt well to such scenarios and still has room for further improvement.

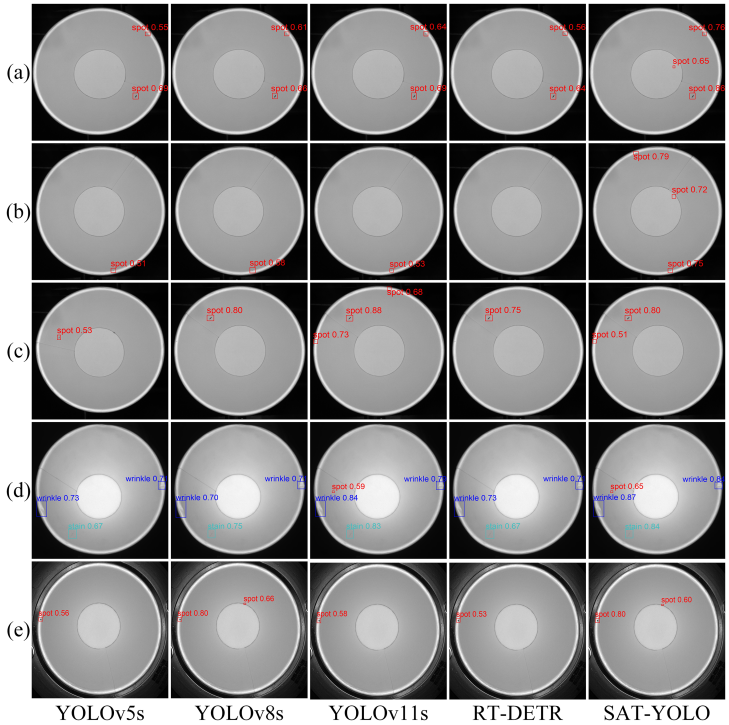


Figure 13: Detection efficacy benchmarking: YOLOv5s model vs. SAT-YOLO model performance.

5 Generalisation Experiments

5.1 Generalisation Experiments with the SAT-YOLO Model

This paper evaluates the cross-domain performance of the SAT-YOLO model through systematic analysis of small target detection recognition accuracy, employing the benchmark RSOD dataset with quantitative outcomes documented in Table 11.

Table 11: Comparison of the SAT-YOLO and YOLOv5s models on the RSOD dataset.

Model	P/%	R/%	mAP50/%	mAP50–95/%
YOLOv5s	94.5	93.3	95.0	63.4
SAT-YOLO	98.3	94.1	96.1	65.9

From Table 11, it can be seen that the SAT-YOLO model is also effective in improving the detection of small targets on public datasets, thus proving that the SAT-YOLO model has a good generalization ability.

5.2 Generalisation Experiments with the SKBAttention Module

This paper conducts comparative experiments on the open-access HIT-UAV dataset to demonstrate SKBAttention's enhanced detection capabilities vs. SKNet in infrared small object scenarios. The quantitative evaluation metrics across both architectures are systematically compared in Table 12.

Table 12: Performance evaluation of the SKBAttention and SKNet modules using the HIT-UAV dataset.

Model	P/%	R/%	mAP50/%	mAP50–95/%
YOLOv5s	88.3	75.3	81.1	51.7
YOLOv5s+SK	89.4	73.6	80.6	51.1
YOLOv5s+SKB	89.6	78.2	82.6	52.4

From Table 12, it can be seen that the SKBAttention module is also effective in improving the detection of small targets on public datasets, thus proving that the SKBAttention module has a good generalization ability.

5.3 Generalisation Experiments with the AAConv Module

To validate the performance advantages of the AAConv module compared with standard convolution operations for small object recognition, cross-platform evaluations were conducted using two distinct configurations: YOLOv5s implementation on the HIT-UAV public dataset and YOLOv8s deployment on the paper cup dataset. Comparative analyses of detection accuracy are systematically documented in Tables 13 and 14, respectively.

Table 13: Comparison of the AAConv and Conv modules on the HIT-UAV dataset using YOLOv5s.

Module	P/%	R/%	mAP50/%	mAP50–95/%
Conv	88.3	75.3	81.1	51.7
AAConv	88.8	77.8	82.3	51.9

Table 14: Comparison of AAConv and Conv modules on the paper cup dataset using YOLOv8s.

Module	P/%	R/%	mAP50/%	mAP50–95/%	mAP50 (spots)/%
Conv	82.6	73.1	74.8	52.0	34.9
AAConv	83.8	74.4	76.1	53.8	36.7

As can be seen from Tables 13 and 14, the AAConv module is also effective in improving the detection of small targets on the public dataset and YOLOv8s, thus proving that the AAConv module has a good generalization ability.

6 Conclusions

This paper presents SAT-YOLO, an enhanced YOLOv5s-based detector tailored for small-defect detection on industrial paper cups. By integrating SKBAttention, AAConv, and a TSCODE decoupled head, the proposed framework improves spatial—channel feature discrimination, preserves fine-grained cues of tiny defects during feature extraction, and mitigates the interference between classification and localization tasks. Extensive experiments on the paper cup defect dataset demonstrate that SAT-YOLO consistently outperforms the YOLOv5s baseline across standard detection metrics, delivering more reliable predictions and fewer missed detections for small targets. In addition, cross-dataset evaluations on public small-object benchmarks further corroborate that the proposed modules are effective beyond a single industrial dataset, indicating encouraging generalization capability.

Despite these improvements, SAT-YOLO introduces additional model complexity, leading to a higher parameter count and increased computational cost, which may limit real-time deployment on resource-constrained edge devices. Future work will therefore focus on efficiency-oriented optimizations, including structured pruning, knowledge distillation, and quantization-aware training, as well as deployment-level acceleration and realistic FPS reporting under industrial settings. Moreover, we will further enhance robustness to extreme illumination variations—such as over-exposure, strong specular reflections, shadows, and low-light conditions—to improve detection stability and reliability in practical production lines.

Acknowledgement: This work was supported by the Key R&D projects of the Science and Technology Department of Zhejiang Province (2024C01222, 2023C01158, 2024C03118).

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Shuaian Zhang and Daojin Dai; methodology, Shuaian Zhang; software, Shuaian Zhang and Ziliang Peng; validation, Shuaian Zhang, Ziliang Peng, Jian Huang and Ruiyin Chen; formal analysis, Shuaian Zhang; investigation, Shuaian Zhang; resources, Jian Huang and Daojin Dai; data curation, Shuaian Zhang and Ruiyin Chen; writing—original draft preparation, Shuaian Zhang; writing—review and editing, Shuaian Zhang, Ziliang Peng, Jian Huang, Ruiyin Chen and Daojin Dai; visualization, Shuaian Zhang; supervision, Daojin Dai; project administration, Daojin Dai; funding acquisition, Daojin Dai. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author, Jian Huang, upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Guzman-Puyol S. Fluorinated compounds in paper and paperboard based food packaging materials. npj Sci Food. 2024;8(1):82. doi:10.1038/s41538-024-00326-2.
2. Liu C, Samsudin MR, Zou Y. The impact of visual elements of packaging design on purchase intention: brand experience as a mediator in the tea bag product category. Behav Sci. 2025;15(2):181.
3. Aydin A, Güney S. Fault detection system for paper cup machine based on real-time image processing. Eng Appl Artif Intell. 2024;133(3):108070. doi:10.1016/j.engappai.2024.108070.
4. Ibrahim A, Tapamo JR. A survey of vision-based methods for surface defects' detection and classification in steel products. Informatics. 2024;11(2):25. doi:10.3390/informatics11020025.
5. Werheid J, Behnen H, Woltersmann JH, He S, Hamann T, Abdelrazeq A, et al. Machine vision in manufacturing SMEs: a review. Discover Appl Sci. 2025;7(5):371. doi:10.1007/s42452-025-06923-4.
6. Tian Y, Ye Q, Doermann D. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. 2025.
7. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. arXiv:2506.17733. 2025.
8. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press; 2015. p. 91–9.
9. Vijayakumar A, Vairavasundaram S. YOLO-based object detection models: a review and its applications. Multimed Tools Appl. 2024;83:83535–74.
10. Luo S, Xu Y, Zhang C, Jin J, Kong C, Xu Z, et al. LIDD-YOLO: a lightweight industrial defect detection network. Meas Sci Technol. 2025;36(1):016105.
11. Guo L, Wang Y, Guo M, Zhou X. YOLO-IRS: infrared ship detection algorithm based on self-attention mechanism and KAN in complex marine background. Remote Sens. 2024;17(1):20.
12. Yuan T, Jiao Z, Diao N. YOLO-SSW: an improved detection method for printed circuit board surface defects. Micromachines. 2025;13(3):435.
13. Zhang Y, Ye M, Zhu G, Liu Y, Guo P, Yan J. FFCA-YOLO for small object detection in remote sensing images. IEEE Trans Geosci Remote Sens. 2024;62:5611215. doi:10.1109/tgrs.2024.3363057.
14. Liang J, Wang F, Chen J, Huang HY, Dou ZF. Aerial small target detection algorithm based on cross-scale separated attention. PLoS One. 2025;20(11):e0337318. doi:10.1371/journal.pone.0337318.
15. Sunkara R, Luo T. No strided convolutions or pooling: a new CNN building block for low-resolution images and small objects. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin/Heidelberg, Germany: Springer; 2022. p. 450–66.
16. Qiao S, Chen LC, Yuille A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2021. p. 10213–24.
17. Zhuang J, Qin Z, Yu H, Chen X. Task-specific context decoupling for object detection. arXiv:2303.01047. 2023.
18. Long Y, Gong Y, Xiao Z, Liu Q. Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Trans Geosci Remote Sens. 2017;55(5):2486–98. doi:10.1109/tgrs.2016.2645610.
19. Suo J, Wang T, Zhang X, Chen H, Zhou W, Shi W. HIT-UAV: a high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection. Sci Data. 2023;10(1):227.
20. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2024. p. 16965–74.
21. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multibox detector. In: European Conference on Computer Vision (ECCV). Cham, Switzerland: Springer; 2016. p. 21–37.
22. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2018. p. 7132–41.
23. Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2021. p. 13713–22.

24. Yang L, Zhang RY, Li L, Xie X. SimAM: a simple, parameter-free attention module for convolutional neural networks. In: International Conference on Machine Learning (ICML). London, UK: PMLR; 2021. p. 11863–74.
25. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q. ECA-Net: efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2020. p. 11534–42.
26. Zhang QL, Yang YB. SA-Net: shuffle attention for deep convolutional neural networks. In: ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE; 2021. p. 2235–9.
27. Ouyang D, He S, Zhang G, Luo M, Guo H, Zhan J, et al. Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE; 2023. p. 1–5.
28. Li X, Wang W, Hu X, Yang J. Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE; 2019. p. 510–9.
29. Li C, Zhou A, Yao A. Omni-dimensional dynamic convolution. arXiv:2209.07947. 2022.
30. Zhang X, Song Y, Song T, Yang D, Ye Y, Zhou J. LDConv: linear deformable convolution for improving convolutional neural networks. Visual Computing. 2024;149(7–8):105190. doi:10.1016/j.imavis.2024.105190.
31. Zhang X, Liu C, Yang D, Song T, Ye Y, Li K, et al. RFAConv: innovating spatial attention and standard convolutional operation. arXiv:2304.03198. 2023.