Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Cigarette-Group End-Face Defect Detection Based on an Improved DEIM

**Xuhui Ye**
Hubei University of Technology
**Lilu Huang**
Hubei University of Technology
**Zhong Li**
Hubei University of Technology
**Miao Zheng**
Hubei University of Technology
**Yan Chen**
dhxychenyan@wdu.edu.cn

Wuhan Donghu University

**Article**

**Keywords:** cigarette-group end face, defect detection, DEIM, lightweight network, edge deployment

**Posted Date:** November 24th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11256168/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Cigarette-Group End-Face Defect Detection Based on an Improved DEIM

Xuhui Ye [1], Lilu Huang [1], Zhong Li [1], Miao Zheng [1], and Yan Chen [2,*]

1. School of Mechanical Engineering, Hubei University of Technology, Wuhan, Hubei 430068, China.

2*. School of Mechanical and Electrical Engineering, Wuhan Donghu University, Wuhan, Hubei 430212, China.

*Corresponding author(s). E-mail(s): dhxychenyan@wdu.edu.cn;

Contributing authors: yxh89@hbut.edu.cn ; 102510152@hbut.edu.cn; 102610148@hbut.edu.cn； 102611993@hbut.edu.cn; dhxychenyan@wdu.edu.cn

**Abstract**

To address the diverse forms of cigarette-group end-face defects, the weak visual features of small defects, and the limited computing resources of edge devices, this study proposes a lightweight detection method based on an improved DEIM-S model. First, MobileNetV4 replaces the original backbone, using its Universal Inverted Bottleneck structure to reduce the computational cost of feature extraction. Second, a C2F-IDWB module combines the C2F structure with InceptionDWBlock to strengthen multiscale defect feature extraction and fusion. Finally, an Efficient Multi-Scale Attention (EMA) mechanism is introduced at the encoder's P5 feature output to enhance responses in defective regions and suppress irrelevant background information. Experiments were conducted on a self-developed da-taset containing five defect categories: flattened ends, reversed cigarettes, missing tobacco, broken cigarette tips, and broken cigarette bodies. The improved model achieved AP50 and AP50:95 values of 99.1% and 81.07%, respectively, exceeding the baseline by 3.0 and 3.54 percentage points. Its computational cost fell by 28.3% to 17.82 G FLOPs, and its model size decreased by 21.9% to 30.6 MB. Under the comparative experimental conditions used in this study, it achieved a detection speed of 189.56 FPS. A PTH → ONNX → RKNN model conversion procedure and a deployment scheme for the RK3588 edge platform are also presented. By improving detection accuracy while reducing computational cost, the proposed method provides a technical reference for real-time detection of cigarette-group end-face defects in GDX2 packaging machines.

**Keywords:** cigarette-group end face; defect detection; DEIM; lightweight network; edge deployment

## 0 Introduction

China is one of the world's largest tobacco-consuming countries [1]. In 2025, its ciga-rette output exceeded 2.47 trillion cigarettes [2], and the tobacco industry generated RMB 1.657 trillion in combined industrial and commercial taxes and profits [3], making a sub-stantial contribution to national revenue. The appearance quality of cigarettes is an im-portant indicator of production standards and product consistency. During cigarette packaging, cigarettes that have undergone rod making and conveying are formed into groups inside the packaging machine and arranged in a prescribed pattern before entering subsequent packaging stages. Continuous conveying, stacking, compression, and jogging can subject cigarettes to uneven forces, excessive local pressure, and changes in their ori-entation as they fall. These conditions may cause blockages, missing cigarettes, tilting, and deformation, which can in turn lead to appearance defects such as flattened ends, missing material, breakage, and reversed cigarettes. If these defects are not identified and removed promptly, they may compromise product consistency and cause packaging equipment to stop for intervention, reducing production efficiency. Existing inspection methods rely largely on manual sorting or conventional image processing algorithms [4]. Their low effi-

ciency, limited real-time performance, and high miss rates make it difficult to meet the re-quirements of cigarette packaging lines for accurate, real-time, and stable inspection.

In recent years, deep learning has been widely applied to industrial defect detection because of its ability to learn features automatically [5]. Two-stage object detectors based on region proposals [6–8] locate and identify objects through proposal generation, classi-fication, and regression, but generally incur substantial computational costs. In contrast, single-stage methods represented by YOLO (You Only Look Once) [9] and SSD (Single Shot MultiBox Detector) [10] directly predict object locations and classes, offering an ad-vantage in detection speed. Transformer-based detectors use attention mechanisms for feature interaction, providing another approach to defect detection in complex scenes. Ap-plying these methods to cigarette packaging equipment requires model designs that ac-count for both the characteristics of cigarette-group end-face defects and the available computing resources.

Previous studies on defect detection in cigarettes and related products have improved feature extraction, attention mechanisms, and model compression. Yuan et al. [11] incor-porated ASPP, SENet, and an improved bounding-box loss into YOLOv4 and used K-means++ to optimize anchor boxes, enhancing the extraction of defect features. Liu et al. [12] proposed C-CenterNet, which combines CBAM attention, deformable convolution, and the ACON activation function to improve detection accuracy, although it also in-creases the parameter count and computation time. For cigarette-pack defect detection, Dong et al. [13] proposed AMC-YOLO and achieved good performance in identifying mul-tiple defect categories. Their work provides a reference for visual inspection of cigarette products; however, differences in appearance and imaging conditions between cigarette packs and cigarettes in the feed channel call for model design and validation tailored to the specific inspection target.

To enhance defect features while reducing model complexity, Wu Qinghua et al. [14] proposed TCN-Net, which integrates a triplet attention mechanism with feature-focused diffusion to strengthen salient defect features and improve discrimination between similar defects. To address missed detections of small cigarette defects, Zhang Pengtao et al. [15] proposed CSDD-YOLO. They added a P2 detection head to improve the detection of fine defects, compressed the model using LAMP pruning, and validated its deployment in a C++ environment. These studies show that stronger defect representation and lower model complexity can both support practical cigarette inspection. Nevertheless, feature en-hancement modules generally add computational cost, while model compression may weaken the representation of fine details. Detection accuracy and deployment efficiency therefore require further balancing.

To address these challenges, this paper proposes a lightweight method for detecting cigarette-group end-face defects based on an improved DEIM-S model. First, MobileNetV4 replaces the original backbone, and its universal inverted bottleneck structure reduces the computational cost of feature extraction. Second, the C2F structure is combined with In-ceptionDWBlock to form a C2F-IDWB module that strengthens the extraction and fusion of multiscale defect features. Finally, an EMA attention mechanism is introduced at the en-coder’s P5 feature output to enhance responses to defective regions and suppress irrele-vant background information. The proposed method is evaluated on a self-developed da-taset containing five categories of cigarette appearance defects through ablation studies, comparative experiments, and visualization. It is also deployed on an RK3588 edge plat-form, providing a reference for detecting cigarette-group end-face defects in cigarette feed channels.

# 1 Cigarette-Group Forming and Inspection System

## 1.1 System Design

Cigarette-group forming is the final critical stage in packaging machines such as the GDX2 before the cigarettes are enclosed in a pack. After rod making and conveying, ciga-rettes are arranged in a fixed

pattern. Within the cigarette feed channel, they pass sequen-tially through first-stage jogging, a buffering partition, and second-stage jogging before entering the pack-loading channel in an orderly manner and being pushed into the form-ing pocket. As shown in Figure 1, the cigarette-group end-face defect inspection system mainly consists of a visual acquisition unit, a first-stage jogging mechanism, an S-shaped buffering partition, and a second-stage jogging mechanism.

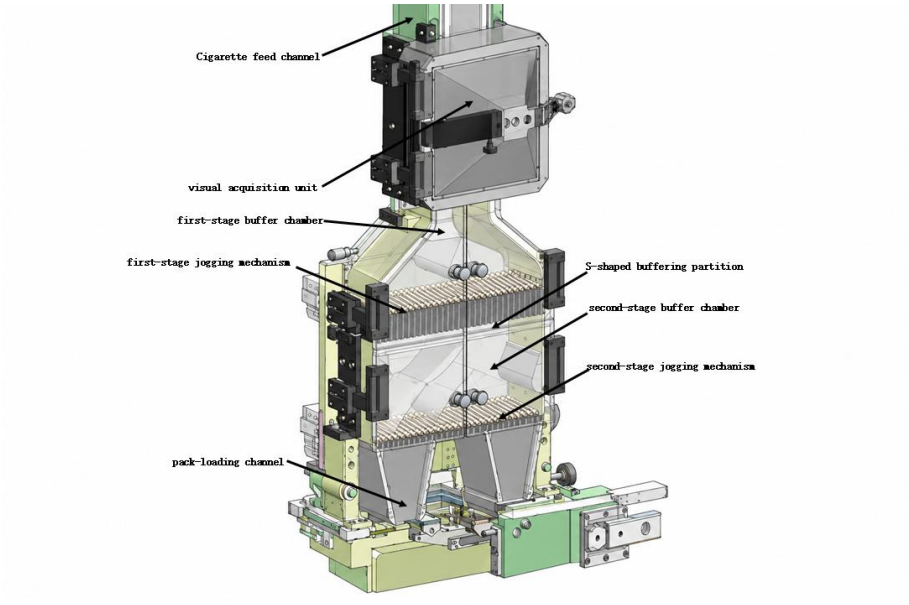
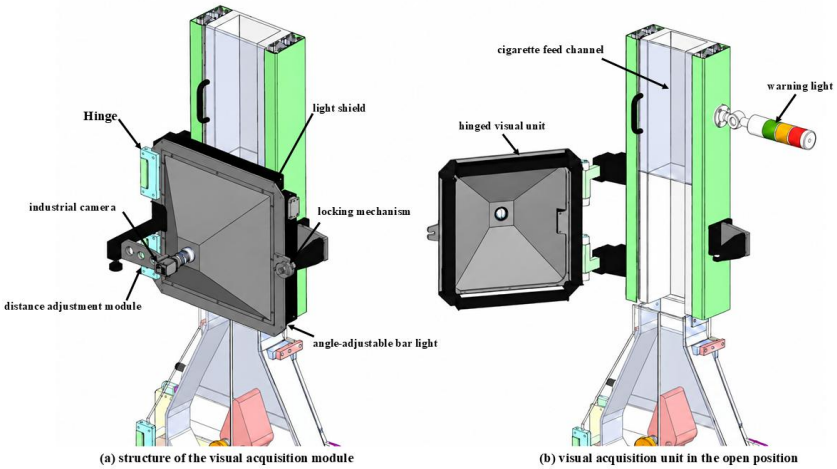



**Figure 1.** Cigarette-group end-face defect detection system.

The visual acquisition unit captures high-resolution images of cigarettes in the ciga-rette feed channel and transmits them to the image processing module. As shown in Figure 2(a), an industrial camera is used with angle-adjustable bar lights. A distance adjustment module sets the imaging position, while a light shield reduces interference from ambient light. In Figure 2(b), the visual unit can be opened on a hinge to expose the cigarette feed channel, allowing operators to inspect the cigarettes and maintain the equipment. The image acquisition module consists of a 1.6-megapixel color industrial camera and bar lights. The lights are positioned above, below, and on both sides of the inspection area, and their angles can be adjusted according to the imaging results. This arrangement pro-vides uniform illumination of the imaged surface while reducing ambient-light interfer-ence, making abnormal cigarettes easier to identify. The S-shaped buffering partition dis-tributes the pressure generated during cigarette conveying, reducing the load on cigarettes in the lower layers and limiting compression-induced deformation.




(a) structure of the visual acquisition module (b) visual acquisition unit in the open position

**Figure 2.** Structural diagram of the image acquisition module.

## 1.2 System Workflow

The system workflow is shown in Figure 3. When cigarettes pass through the visual acquisition area above the cigarette feed channel, the bar lights illuminate them, and the industrial camera captures images for preprocessing and anomaly identification by the image processing module. If an abnormal cigarette is detected, a warning light is activated. The operator then opens the hinged door of the first-stage buffer area, removes the abnor-mal cigarette with tweezers, and closes the door. Normal cigarettes pass through first-stage jogging, the S-shaped buffering partition channel, and the second-stage buffer chamber. They then enter the pack-loading channel through the existing second-stage jog-ging mechanism before being pushed into the forming pocket.

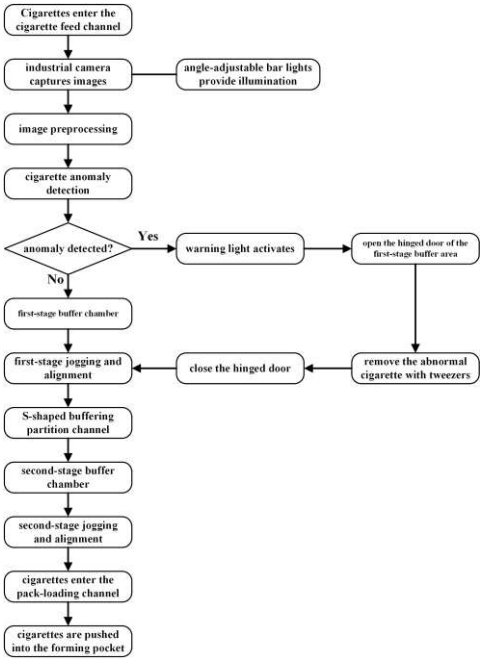



**Figure 3.** Workflow of the proposed system.

## 2 Model Construction

DEIM has a modular architecture that supports rapid optimization and customiza-tion [16]. For example, feature extraction modules in its backbone and encoder can be re-placed or combined with other modules to meet the requirements of different tasks. Its flexible architecture also facilitates the integration of attention mechanisms to improve ob-ject detection accuracy. As a lightweight version of the DEIM series, DEIM-S has relatively few parameters while maintaining high detection accuracy. It can therefore operate effi-ciently with limited computing resources and is suitable for defect detection scenarios that require both accuracy and real-time performance. This study improves DEIM-S in three respects. The resulting architecture is shown in Figure 4.

1) MobileNetV4 is introduced as a lightweight backbone. Its Universal Inverted Bot-tleneck (UIB) structure extracts features while reducing the number of model parameters and computational complexity without substantially compromising feature representa-tion.

2) A C2F-IDWB module is designed to improve feature extraction in the encoder. By combining the C2F structure with InceptionDWBlock (IDWB), the module strengthens multiscale feature extraction and fusion while reducing parameter count and computation, thereby maintaining detection accuracy.

3) An Efficient Multi-Scale Attention (EMA) mechanism is introduced to enhance the deep features output by the encoder. By modeling dependencies between spatial positions and channel features, it

emphasizes informative features in defective regions and sup-presses redundant background information, further improving detection accuracy.

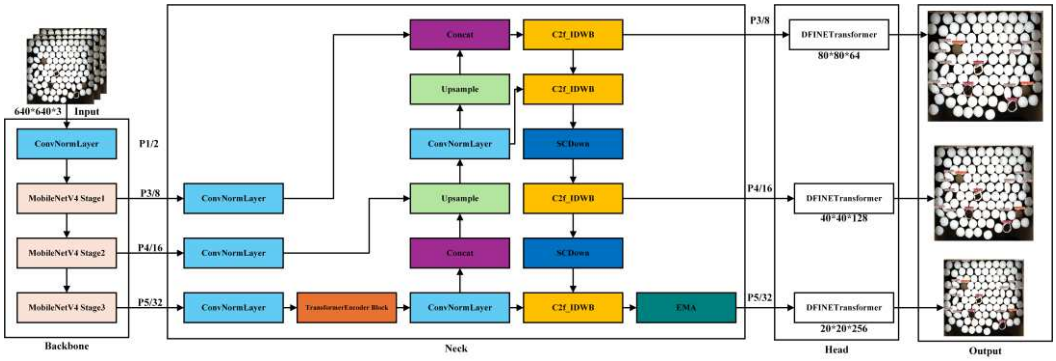



**Figure 4.** Architecture of the improved DEIM-S network.

## 2.1 MobileNetV4

In the task of cigarette-group end-face defect detection, the computational complexity of the original DEIM-S backbone makes deployment difficult in resource-constrained settings. Therefore, this study adopts the lightweight and efficient MobileNetV4 as the backbone for feature extraction. Its architecture is shown in Figure 5. MobileNetV4 uses the Universal Inverted Bottleneck (UIB) block, which provides a unified representation of different convolutional blocks by flexibly incorporating two optional depthwise convolutions into a conventional inverted bottleneck. UIB can adjust how spatial and channel information interact according to feature extraction requirements, maintaining strong feature representation while reducing the number of parameters and computational complexity.The first depthwise convolution is placed before channel expansion. It extracts initial spatial information from the input features and improves the network's sensitivity to local surface textures and subtle cigarette defects. The second is placed between channel expansion and projection. It further facilitates spatial feature interaction and strengthens high-level semantic information. By combining these two depthwise convolutions in different ways, UIB can form different feature extraction structures and efficiently model spatial and channel information in a low-dimensional space. In addition, MobileNetV4 uses hardware-aware neural architecture search to balance accuracy and computational efficiency across computing platforms [17].

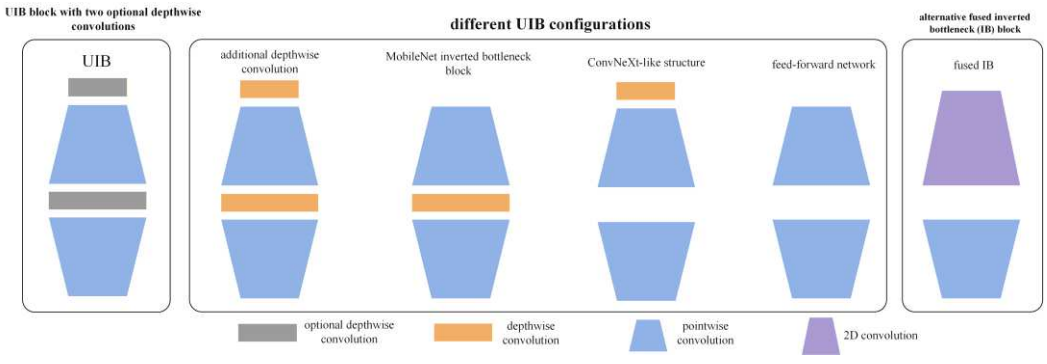



**Figure 5.** Architecture of the MobileNetV4 network.

In the MobileNetV4 architecture, UIB serves as a core component. It efficiently extracts and fuses input features through channel expansion, depthwise convolution, and channel projection. Let $X$ denote the input feature map. The feature transformation performed by UIB can be expressed as:

$$Y=PW_{proj}\left(DW\left(PW_{\exp}\left(X\right)\right)\right) \tag{1}$$

where $PW_{\text{exp}}$ denotes the $1 \times 1$ pointwise convolution used for channel expansion, $DW$ denotes depthwise convolution, $PW_{\text{proj}}$ denotes channel projection, and $Y$ denotes the output feature map of the UIB block. First, pointwise convolution expands the feature channels to increase the feature representation dimension. Depthwise convolution then extracts local spatial information. Finally, pointwise convolution projects the features to the target dimension, enhancing the network's representational capacity while controlling computational cost.

Unlike standard convolution, depthwise convolution operates independently on each input channel. For a convolution with a kernel size of $K \times K$, $C_{\text{in}}$ input channels, and $C_{\text{out}}$ output channels, the parameter counts of standard and depthwise convolution can be expressed as:

$$P_{Conv} = K^2 C_{in} C_{out}, P_{DW} = K^2 C_{in} \tag{2}$$

where $P_{\text{Conv}}$ and $P_{\text{DW}}$ denote the parameter counts of standard and depthwise convolution, respectively. As shown in Eq. (2), depthwise convolution does not perform full spatial convolution across every pair of input and output channels, substantially reducing the model's parameter count and computational complexity.

## 2.2 C2F-IDWB Module

DEIM-S uses the C2f module for feature extraction during multiscale feature fusion. The module further extracts and fuses features from different levels, combining shallow detail with deep semantic information to provide richer representations for the subsequent detection head. During deep feature fusion, however, extensive use of standard convolutions still incurs computational and memory costs. This increases model complexity and creates a bottleneck in balancing detection accuracy with deployment efficiency on resource-constrained devices.To address this issue, this study introduces IDWB to optimize the C2f module. IDWB serves as the core unit in the feature transformation branch of the C2f two-branch structure, while the information flow in the other branch remains unchanged. The resulting module is termed C2F-IDWB. IDWB uses depthwise convolution to model spatial information efficiently and an Inception-style multibranch design to strengthen interactions among local features at different scales. It thus obtains richer feature representations while reducing parameter count and computational complexity. C2F-IDWB is applied to the fusion of P3, P4, and P5 multiscale features in the DEIM-S encoder, further enhancing features at different levels after fusion and improving the model's representational capacity. The architecture of the proposed C2F-IDWB module is shown in Figure 6.

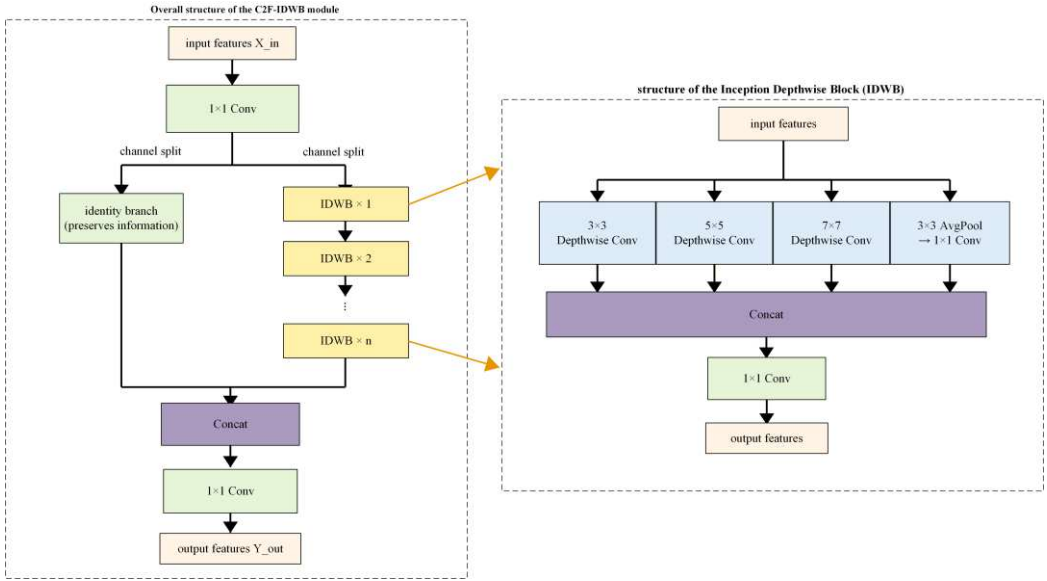



**Figure 6.** Architecture of the C2F-IDWB module.

## 2.3 EMA Multiscale Attention Mechanism

DEIM-S effectively aggregates information from different feature levels through multiscale feature fusion. In cigarette-group end-face defect detection, however, deeper features may contain substantial background and redundant information as the network depth increases, interfering with features that characterize defective regions. This problem is particularly pronounced for small defects with subtle texture differences, whose responses can be weakened by complex background information. To address it, this study introduces an Efficient Multi-Scale Attention (EMA) mechanism at the deep-feature output of the encoder to further refine the P5 features fused by C2F-IDWB. Its structure is shown in Figure 7. By grouping channels and enabling feature interaction at different scales, EMA models dependencies between spatial positions and channel information without substantially increasing computational complexity. It enhances responses in defective regions and suppresses irrelevant background information, directing the network's attention toward key defect areas on cigarette surfaces and improving the representation of deep semantic features.

In the EMA mechanism, the input features are first divided into several groups along the channel dimension, allowing feature interaction within lower-dimensional subspaces. Let the input feature map be $X$. The grouping process can be expressed as:

$$X=\left\{X_1, X_2, \cdots, X_{\mathrm{g}}\right\},\ X_i \in R^{(C/g \times H \times W)} \tag{3}$$

where $C$、$H$、$W$ denote the number of input channels, height, and width, respectively, and $g$ denotes the number of feature groups. The grouping factor is set to $g = 8$. Channel grouping reduces the feature dimension involved in attention computation while preserving local information within each subspace.

EMA then processes the grouped features using branches at different scales. One branch aggregates features along the horizontal and vertical directions to establish long-range spatial dependencies. The other uses a $3 \times 3$ convolution to extract local spatial information and obtains attention weights through cross-spatial information interaction. The resulting attention weights are multiplied elementwise with the original input features. The output can be expressed as:

$$Y = X \otimes W_{EMA} \tag{4}$$

where $W_{EMA}$ denotes the attention weights generated by EMA, $\otimes$ denotes elementwise multiplication, and $Y$ denotes the attention-enhanced output features. In this way, EMA adaptively adjusts feature responses according to the importance of each spatial position, enhancing information related to defects while suppressing irrelevant responses from cigarette backgrounds and redundant textures.

In this study, EMA is placed at the P5 feature output of the DEIM-S encoder to further enhance the 256-channel deep features processed by C2F-IDWB. Applying attention only at this point recalibrates high-level semantic features before they enter the decoder, enhancing responses in key defective regions while limiting additional computation. The EMA-enhanced P5 features are fed into DFINETransformer together with the P3 and P4 features for further multiscale interaction, improving the model's representation of cigarette-group end-face defects.

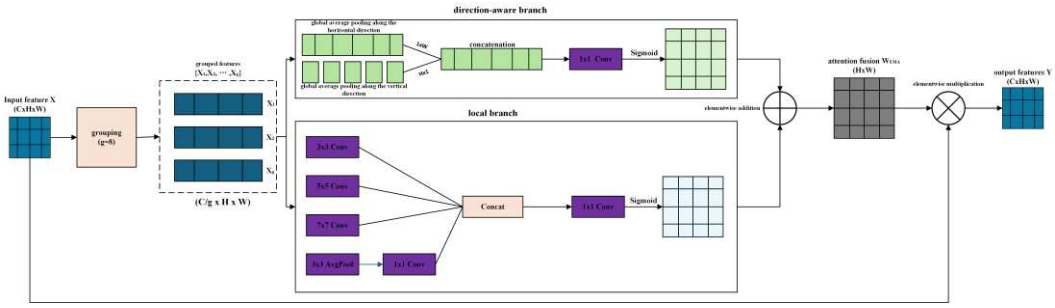


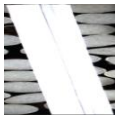

Figure 7. Architecture of the EMA mechanism.

# 3 Experimental Analysis

## 3.1 Dataset Collection and Preparation

Severe defects occur infrequently and unpredictably in actual production. When defects appear continuously or on a large scale, they usually indicate abnormal equipment operation, requiring the production line to stop promptly for inspection. It is therefore difficult to collect a sufficiently large and class-balanced set of defect samples directly from an operating production line over an extended period. To improve the model's ability to learn different defect forms and combinations, this study supplemented real defect samples with images of typical defects and various sample combinations captured using a simulated acquisition setup. The device shown in Figure 8 was used to acquire images of cigarette defects. The cigarettes were Huanghelou (Tianxia Minglou), produced by Hubei China Tobacco Industrial Co., Ltd., as shown in Figure 9. All cigarette samples used for image acquisition were purchased through commercial channels and were not supplied by the manufacturer. A total of 1,269 images containing cigarette appearance defects were collected. The dataset covers five defect categories: flattened end, reversed cigarette, missing tobacco, broken cigarette tip, and broken cigarette body, as listed in Table 1. The images were annotated in YOLO format using LabelImg and divided into training, validation, and test sets at a ratio of 8:1:1. All experimental results were evaluated on the test set. Images were resized to 640×640 pixels to match the network input.

**Table 1.** Defect categories and counts in the dataset.

| Defect category | Description | Example | Training set | Validation set | Test set |
|---|---|---|---|---|---|
| Flattened end | Face A has an oval shape |  | 9439 | 1170 | 1172 |
| Reversed cigarette | Faces A and B are interchanged |  | 3642 | 416 | 449 |
| Missing tobacco | Tobacco material is missing from face B |  | 3548 | 433 | 428 |
| Broken cigarette tip | A broken tip lies on face A |  | 185 | 29 | 33 |
| Broken cigarette body | A broken section of the cigarette body lies on face A |  | 50 | 5 | 4 |

Note: Breakage defects are extreme cases and occur very rarely in actual production; consequently, only a small number of such samples were collected.

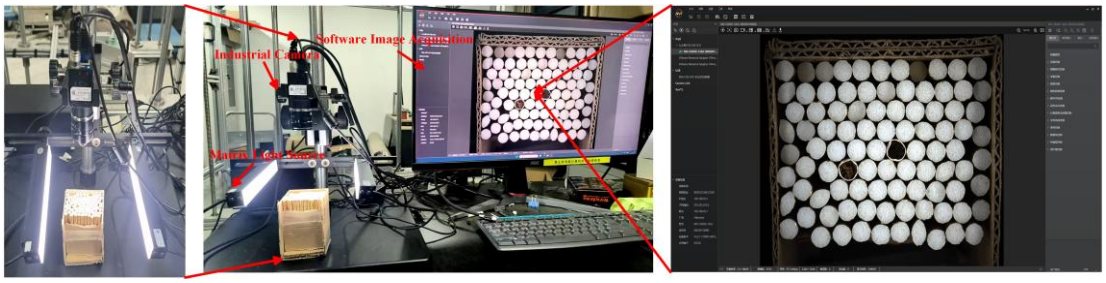



**Figure 8.** Image acquisition setup.




**Figure 9.** Schematic diagram of cigarette components.

## 3.2 Experimental Environment and Parameter Settings

To evaluate the effectiveness of the proposed model, a training platform was built using Windows 11 as the operating system and PyTorch as the deep learning framework. All experiments were conducted on this platform. Its configuration is listed in Table 2.

**Table 2.** Configuration of the model training platform.

| Component | Specification |
|---|---|
| Operating system | Windows11 |
| CPU | Intel(R) Core(TM) i5-14600KF CPU 3.50GHz |
| GPU | NVIDIA GeForce GTX 5060 Ti 16G |
| CUDA | 13.0 |
| RAM | 32GB |
| Deep learning framework | Pytorch2.9.0 |
| Programming language | Python 3.9.0 |

The same training parameters and hyperparameters were used throughout all experiments, as shown in Table 3.

**Table 3.** Training parameters and hyperparameters.

| Category | Parameter | Value |
|---|---|---|
| Training parameter | Image size | 640×640 |
| | Batch size | 16 |
| | Epochs | 300 |
| | Number of workers | 8 |
| Hyperparameter | Optimizer | AdamW |
| | Initial learning rate | 0.0004 |
| | Backbone learning rate | 0.0002 |
| | β1 | 0.9 |

| | | |
|---|---|---|
| | β2 | 0.999 |
| | Weight decay | 0.0001 |

## 3.3 Evaluation Metrics

Precision ($P$), recall ($R$), average precision (AP), model size, parameter count (Params), floating-point operations (FLOPs), and frames per second (FPS) were used as evaluation metrics. Detection accuracy was assessed using COCO evaluation criteria, primarily AP50, AP75, and AP50:95. AP50 and AP75 denote average precision at intersection-over-union (IoU) thresholds of 0.50 and 0.75, respectively. AP50:95 denotes the mean of AP values calculated at IoU thresholds from 0.50 to 0.95 in increments of 0.05. The metrics were calculated as follows.

$$P=\frac{TP}{TP+FP}\times 100\% \tag{5}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{6}$$

In the equations, $TP$ denotes the number of correctly detected targets (true positives), $FP$ denotes the number of incorrect positive predictions (false positives), and $FN$ denotes the number of existing targets that the model failed to detect correctly (false negatives).

$$P_{\text{interp}}(r)=\max_{\tilde{r}\geq r} P\left(\tilde{r}\right) \tag{7}$$

$$AP_c(\tau)=\frac{1}{101}\sum_{r\in\{0,0.01,\ldots,1\}} P^{c,\tau}_{\text{interp}}(r) \tag{8}$$

$$AP_{50}=\frac{1}{M}\sum_{c=1}^{M} AP_c(0.50) \tag{9}$$

$$AP_{75}=\frac{1}{M}\sum_{1}^{M} AP_c(0.75) \tag{10}$$

$$AP_{50:95}=\frac{1}{10M}\sum_{c=1}^{M}\sum_{k=0}^{9} AP_c(0.50+0.05k) \tag{11}$$

In the equation, $r$ denotes a recall sampling point, and $\tilde{r}$ denotes a recall value greater than or equal to the current sampling point. $M$ denotes the total number of defect categories, $c$ denotes the category index, and $\tau$ denotes the IoU threshold.

## 3.4 Ablation Study

Ablation experiments were conducted to assess the effectiveness of each modification. The results are presented in Table 4. Replacing the original DEIM-S backbone with lightweight MobileNetV4 reduced the parameter count, computational cost, and model size by 12.9%, 21.5%, and 18.6%, respectively. At the same time, AP50 increased by 1.4 percentage points and AP50:95 by 2.32 percentage points. These results indicate that MobileNetV4 improves detection accuracy while reducing model complexity.To further reduce the parameter count and strengthen feature extraction, the original feature

extraction module was replaced with C2F-IDWB. The parameter count decreased from 2.78 M to 319.1 K, while computational cost and model size decreased by 8.9% and 4.1%, respectively. AP50:95 increased by a further 0.34 percentage points, indicating that detection performance was maintained as the model was compressed.Finally, the EMA attention mechanism was introduced to improve detection accuracy. Compared with the baseline, the resulting model reduced the parameter count, computational cost, and model size by 90.0%, 28.3%, and 21.9%, respectively, while increasing AP50, AP75, and AP50:95 by 3.0, 2.4, and 3.54 percentage points. Compared with Model 2, adding EMA increased AP50, AP75, and AP50:95 by 2.1, 3.3, and 0.88 percentage points, respectively. The parameter count and model size remained unchanged, and computational cost increased by only 0.06 G. These results support the effectiveness of EMA in enhancing defect features and improving detection accuracy.

**Table 4.** Ablation study results.

| Model | MobileNetV4 | C2F-IDWB | EMA | AP50:95/% | AP75/% | AP50/% | Params | Model Size/MB | FLOPs/G |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | | | | 77.53 | 94.6 | 96.1 | 3.19 M | 39.2 | 24.84 |
| Model 1 | √ | | | 79.85 | 95.7 | 97.5 | 2.78 M | 31.9 | 19.49 |
| Model 2 | √ | √ | | 80.19 | 93.7 | 97.0 | 319.1 K | 30.6 | 17.76 |
| Proposed model | √ | √ | √ | 81.07 | 97.0 | 99.1 | 319.1 K | 30.6 | 17.82 |

## 3.5 Comparative Experiments

To comprehensively evaluate the proposed model for cigarette defect detection, this section compares lightweight backbones, feature extraction modules, attention mechanisms, and detection algorithms. All models in these experiments were trained using the same environment and parameter settings.

### 3.5.1 Comparison of Lightweight Backbones

The proposed model is made lightweight primarily by replacing the backbone. To evaluate the effectiveness of this approach in reducing model size and parameter count, MobileNetV4 was compared with other lightweight backbones, including EfficientNetV2 [18], GhostNet [19], and MobileNetV3 [20], using the baseline model as the starting point. The results are shown in Table 5. Except for MobileNetV4, the candidate backbones generally produced larger models. Using them would leave limited room under the resource budget for subsequent changes to the neck and head, such as introducing attention mechanisms or modifying the loss function to improve accuracy. MobileNetV4 achieved stronger compression than the other candidate backbones, leaving more capacity for subsequent accuracy improvements. It therefore enabled gains in AP while meeting the strict model-size constraint, consistent with the overall objective of deployment in resource-constrained settings.

**Table 5.** Comparison of lightweight backbones.

| DEIM-S | Params/M | Model Size/MB | FLOPs/G | AP50-95/% | AP75/% | AP50/% |
|---|---|---|---|---|---|---|
| +EfficientNetV2 | 2.86 | 175.9 | 65.07 | 80.03 | 96.5 | 97.8 |
| +GhostNet | 2.81 | 44.4 | 21.07 | 78.15 | 91.3 | 97.2 |
| +MobileNetV3 | 2.81 | 34.0 | 20.65 | 79.22 | 97.0 | 98.6 |
| +MobileNetV4 | 2.78 | 31.9 | 19.49 | 79.85 | 95.7 | 97.5 |

### 3.5.2 Comparison of Feature Extraction Modules

To evaluate the effects of different feature extraction modules on model performance, StarBlock [21], LocalBlock [22], PSConv [23], GlobalPool [24], RepHMS [25], and IDWB were compared. The

results are presented in Table 6. With IDWB, the model achieved an AP50:95 of 80.19%, the highest among the modules evaluated and 0.50 percentage points above the next-best module, PSConv. Its AP50 reached 97.0%, exceeding those of StarBlock, LocalBlock, PSConv, GlobalPool, and RepHMS by 1.2, 1.8, 0.9, 0.8, and 4.2 percentage points, respectively. Although its AP75 of 93.7% was slightly below GlobalPool's 93.9%, IDWB achieved the best results for both AP50:95 and AP50. The IDWB model had 319.10 K parameters, a model size of 30.6 MB, and 17.76 G FLOPs, giving it the best overall performance.

**Table 6.** Comparison of feature extraction modules.

| Feature extraction module | Params/K | Model Size/MB | FLOPs/G | AP50:95/% | AP75/% | AP50/% |
|---|---|---|---|---|---|---|
| +StarBlock | 319.10 | 29.9 | 16.92 | 77.04 | 88.5 | 95.8 |
| +LocalBlock | 319.10 | 29.5 | 16.45 | 77.26 | 88.6 | 95.2 |
| +PSConv | 319.10 | 29.3 | 17.42 | 79.69 | 93.6 | 96.1 |
| +GlobalPool | 319.10 | 28.7 | 15.29 | 78.02 | 93.9 | 96.2 |
| +RepHMS | 345.73 | 28.0 | 15.05 | 75.64 | 88.8 | 92.8 |
| +IDWB | 319.10 | 30.6 | 17.76 | 80.19 | 93.7 | 97.0 |

### 3.5.3 Comparison of Attention Mechanisms

To evaluate the effects of different attention mechanisms on detection performance, CA, ECA, CBAM, MOH, SE, and EMA were compared. The results are presented in Table 7. The EMA model achieved the best overall detection performance, with AP50:95, AP75, and AP50 values of 81.07%, 97.0%, and 99.1%, respectively, all higher than those of the other methods. Its AP50:95 was 2.72 percentage points higher than that of CA, the next-best method on this metric; its AP75 was 3.5 percentage points higher than that of MOH; and its AP50 was 0.4 percentage points higher than that of CA. The EMA model had 319.1 K parameters, a model size of 30.6 MB, and 17.82 G FLOPs, adding only a small computational overhead compared with other lightweight attention mechanisms. Considering both detection accuracy and model complexity, EMA effectively enhances key defect features at a relatively low computational cost.

**Table 7.** Comparison of attention mechanisms.

| Attention mechanism | Params/K | Model Size/MB | FLOPs/G | AP50:95/% | AP75/% | AP50/% |
|---|---|---|---|---|---|---|
| +CA | 319.1 | 30.6 | 17.76 | 78.35 | 89.2 | 98.7 |
| +ECA | 319.1 | 30.6 | 17.76 | 75.96 | 91.7 | 93.7 |
| +CBAM | 319.1 | 30.6 | 17.76 | 76.80 | 92.3 | 94.8 |
| +MOH | 319.1 | 31.6 | 17.97 | 76.79 | 93.5 | 95.3 |
| +SE | 319.1 | 30.6 | 17.76 | 76.41 | 91.7 | 93.2 |
| +EMA | 319.1 | 30.6 | 17.82 | 81.07 | 97.0 | 99.1 |

### 3.5.4 Comparison with Other Detection Algorithms

To evaluate the advantages of the proposed model, it was compared with models from the YOLO, RT-DETR, and DEIM families, as well as conventional object detection algorithms. The results are presented in Table 8. The proposed model performed well in terms of both model efficiency and detection accuracy.

**Table 8.** Comparison of different detection algorithms.

| Model | Params/M | FLOPs/G | Model Size/MB | AP50/% | AP50:95/% | FPS |
|---|---|---|---|---|---|---|
| retinanet_r50_fpn | 36.41 | 154.0 | 277.0 | 98.7 | 80.4 | 10.00 |
| cascade-rcnn_r50_fpn | 69.21 | 186.0 | 522.0 | 98.9 | 79.0 | 7.00 |
| ddq-detr-4scale_r50 | 48.32 | 213.0 | 445.0 | 99.3 | 80.8 | 0.60 |
| faster-rcnn_r50_fpn_giou | 41.37 | 158.0 | 414.0 | 99.0 | 75.4 | 8.90 |
| rtdetr-r50 | 42.79 | 134.8 | 171.8 | 99.3 | 83.0 | 52.70 |
| rtdetr-r50-Faster | 37.22 | 120.0 | 149.6 | 99.2 | 82.4 | 55.95 |
| rtdetr-EfficientViT | 10.91 | 28.7 | 44.6 | 99.2 | 80.5 | 151.36 |
| deim_hgnetv2_s | 10.18 | 24.8 | 39.2 | 96.1 | 77.5 | 146.47 |
| deim_rtdetrv2_r18 | 19.88 | 59.9 | 76.2 | 95.8 | 77.4 | 102.69 |
| YOLOv3 | 98.48 | 262.6 | 197.4 | 98.9 | 77.8 | 51.90 |
| YOLOv5 | 7.83 | 19.0 | 16.0 | 98.3 | 76.3 | 265.60 |
| YOLOv6 | 15.99 | 43.0 | 32.2 | 98.5 | 76.3 | 173.48 |
| YOLOv8 | 9.84 | 23.6 | 20.0 | 99.0 | 77.8 | 224.40 |
| YOLOv9 | 6.32 | 22.7 | 13.3 | 99.1 | 78.9 | 190.13 |
| YOLOv10 | 8.07 | 24.8 | 16.6 | 98.4 | 74.8 | 252.92 |
| YOLO11-fasternet | 7.53 | 15.7 | 15.4 | 99.1 | 75.8 | 275.11 |
| YOLO11-efficientViT | 7.29 | 14.4 | 15.5 | 99.1 | 75.5 | 161.33 |
| YOLO11-mobilenetv4 | 9.54 | 27.9 | 19.5 | 99.2 | 76.8 | 194.19 |
| YOLOv12 | 9.26 | 21.5 | 19.0 | 98.5 | 79.0 | 192.31 |
| YOLOv26 | 9.95 | 22.5 | 20.3 | 99.0 | 80.0 | 294.00 |
| Proposed model | 0.319 | 17.8 | 30.6 | 99.1 | 81.1 | 189.56 |

The results in Table 8 reveal substantial differences among the models in accuracy, efficiency, and deployment cost. Although the two-stage detection models (RetinaNet, Faster R-CNN, and Cascade R-CNN) all achieved mAP50 values close to 99%, they had relatively large parameter counts and model sizes. Their inference speeds were only 7.0–10.0 FPS, making it difficult to meet the throughput and latency requirements of real-time industrial inspection; their deployment costs were also considerably higher. Transformer-based detectors offered advantages in localization quality, but generally incurred greater computational costs and varied substantially in efficiency. Among them, rtdetr-r50 achieved the highest mAP50:95 of 83.0%, indicating stronger bounding-box regression at high IoU thresholds. Although ddq-detr-4scale_r50 also achieved high accuracy, its speed of just 0.6 FPS was insufficient for real-time inspection. The EfficientViT-based lightweight rtdetr reduced its parameter count to 10.91 M and FLOPs to 28.7 G, but its mAP50:95 fell to 80.5%, indicating some loss of localization quality after model compression.

By comparison, the YOLO models generally maintained lower complexity and higher throughput, with most achieving substantially higher FPS and thus better meeting the needs of real-time industrial inspection. Although YOLOv3 achieved an mAP50 of 98.9%, its parameter count and FLOPs reached 98.48 M and 262.6 G, respectively, while its speed was only 51.90 FPS, making it markedly less efficient than later versions. Performance varied more clearly under the stricter mAP50:95 metric. For example, despite its high FPS, YOLOv10 achieved an mAP50:95 of only 74.8%, showing that high throughput

does not necessarily translate into better localization at higher IoU thresholds. YOLOv26, recently introduced by Ultralytics [26], achieved the highest FPS and combined high speed with good localization quality, but its parameter count remained substantially higher than that of the lightweight model proposed here.

Overall, the proposed model achieved a favorable balance between deployment efficiency and detection performance. It had only 0.319 M parameters, 17.82 G FLOPs, and a model size of 30.6 MB. Its parameter count was the lowest among all compared models, while its FLOPs and model size also remained low. Its mAP50 reached 99.1%, nearly matching the best result in Table 8 and demonstrating competitive detection accuracy. The model achieved 189.56 FPS. Although this was lower than that of high-throughput configurations such as YOLOv5 and YOLOv26, it was substantially faster than the two-stage and Transformer-based models and met the basic real-time inference requirements of industrial applications [27]. At stricter IoU thresholds, its mAP50:95 reached 81.07%, exceeding all YOLO models in the table but remaining below rtdetr-r50 and rtdetr-r50-Faster. This indicates that its localization precision could be improved further. Taken together, the proposed model achieved near-saturated detection accuracy and substantial real-time performance with an exceptionally small parameter count, making it well suited to deployment on resource-constrained edge devices.

## 3.6 Visualization of Detection Results

To compare the detection results of the proposed model and the original model, several images were randomly selected from the test set. The results are shown in Figure 10, where missed targets are outlined in red.

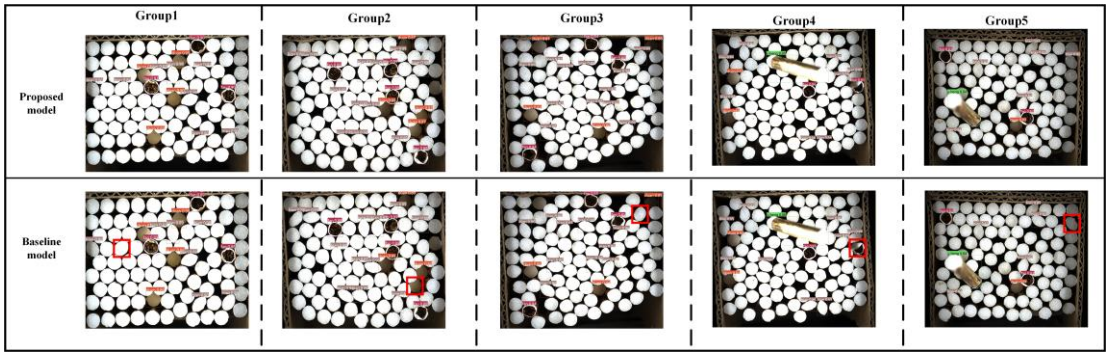



**Figure 10.** Comparison of detection results.

As shown in Figure 10, the proposed model achieved higher detection accuracy than the original DEIM-S model. Specifically, it detected targets missed by DEIM-S, thereby reducing missed detections.

## 3.7 Heatmap Visualization

To examine which features the proposed model attends to when detecting cigarette-group end-face defects, its feature maps were visualized as heatmaps. The model's best checkpoint (best.pth) was used to generate heatmaps with GradCAMPlusPlus [28]. Three groups of images (Group 1–Group 3) were selected from the test set to assess feature extraction by visualizing the regions on which the model focused. The results are shown in Figure 11. Red indicates high attention, while blue indicates low attention. The proposed model concentrated its attention on defective targets, whereas the baseline model's attention was more dispersed.

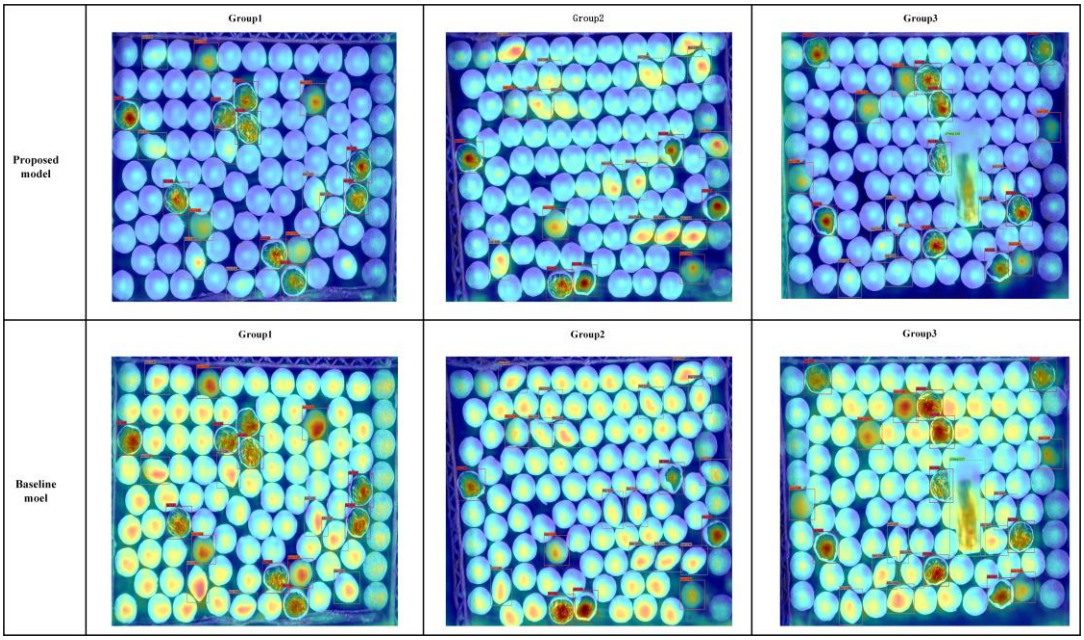



**Figure 11.** Comparison of feature map visualizations using heatmaps.

## 4 Experimental Validation of Edge Deployment

To evaluate the proposed model on an edge computing device, it was deployed on a LubanCat-5 development board (specifications are provided in Table 9). The LubanCat-5 is a low-power, high-performance single-board computer developed by Wildfire Technology based on the Rockchip RK3588. Its built-in neural processing unit (NPU) provides 6 TOPS of computing power [29]. Its power consumption is approximately 1% of that of a GPU with comparable computing power. The board supports mainstream deep learning frameworks, including TensorFlow, PyTorch, Caffe, and ONNX, enabling local, offline AI processing in complex industrial environments.

**Table 9.** Hardware specifications.

| Component | Specification |
|---|---|
| CPU | Quad-core ARM Cortex-A55, Quad-coreARM Cortex-A76,Neon and FPU |
| GPU | ARMMali-G610 MP4, OpenGLES 1.1/2.0/3.2, Vulkan 1.2, openCL 2.2 |
| AI performance | 6TOPS |
| NPU | Support INT4/INT8/INT16/FP16 |
| Memory/storage | LPDDR4X-4 GB/eMMC-32 GB |
| Dimensions | 190mm x125mm |
| Operating system | Ubuntu-20.04.6 |

The proposed model is a Transformer-based model stored in PTH format. Because the RK3588 board's NPU accelerates inference only for models in RKNN format, the model must be converted to RKNN for NPU inference. Direct conversion is difficult because the RK platform does not fully support all Transformer operators. ONNX models, however, offer good compatibility with a range of hardware platforms and inference engines. Therefore, an indirect conversion procedure was adopted: PTH → ONNX → RKNN. Mod-el size changes during conversion because the formats differ in how model data are en-coded, compressed, and stored. The model sizes before and after the proposed improve-ments are shown in Table 10.

**Table 10.** Model sizes.

| Model | Transformer model/MB | ONNX model/MB | RKNN model/MB |
|---|---|---|---|
| Proposed model | 30.6 | 60.3 | 38.1 |
| Baseline model | 39.2 | 77.3 | 43.9 |

The detection performance of the RK3588 edge computing platform was systematically evaluated using cigarette images from the test set. The time from image input to detection output was recorded, and frame rate was used as the primary performance metric. The video detection process in Figure 12 illustrates the model's operation on the RK3588 platform. Detection boxes and confidence scores are displayed in real time, together with detection status and frame rate. Figure 13 shows the detection results obtained by the proposed model on the RK3588 platform.

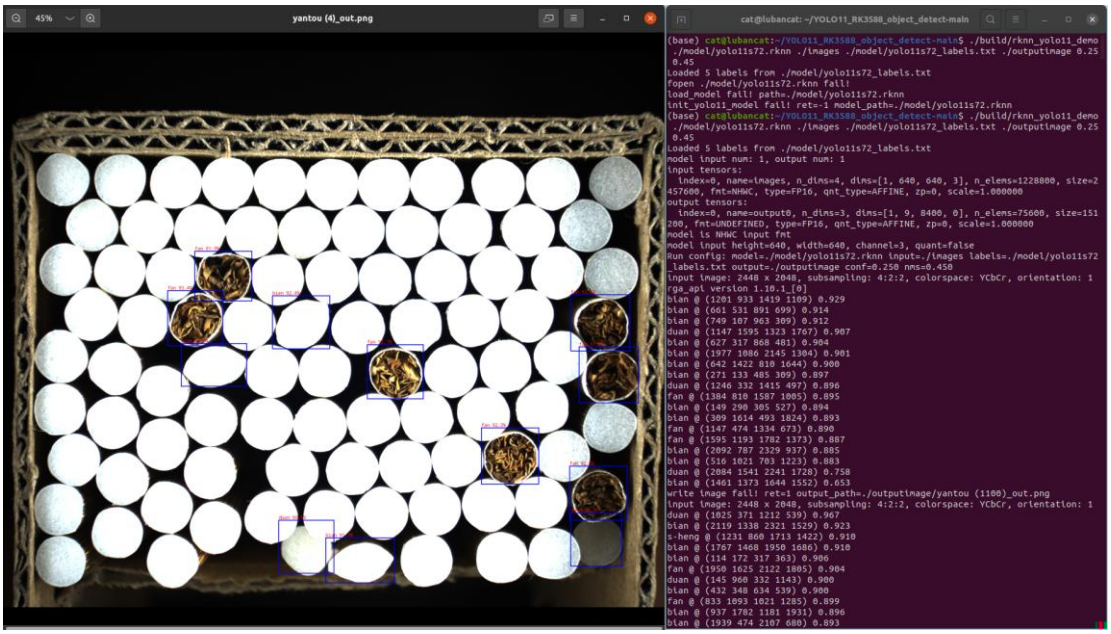



**Figure 12.** Schematic diagram of the detection pipeline on the RK3588 platform.

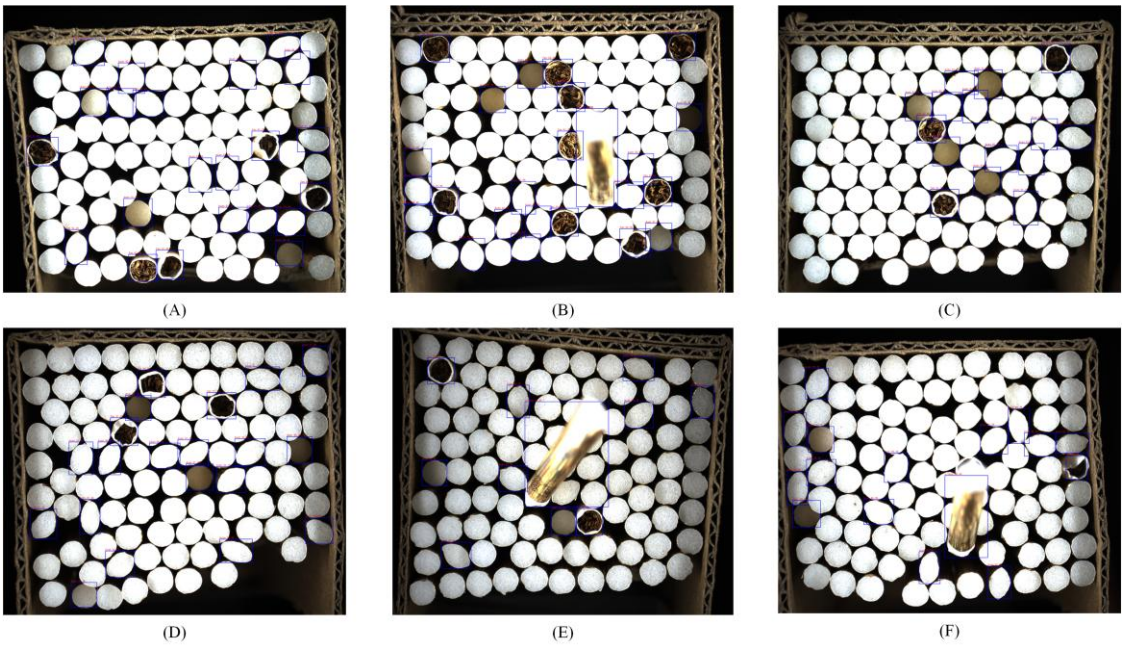



**Figure 13.** Detection results on the RK3588 platform.

To assess the model's performance under continuous production conditions, tests were conducted on an operating GDX2 packaging line. The test sequence comprised normal cigarette groups and samples with flattened ends, reversed cigarettes, missing tobacco, broken cigarette tips, and broken cigarette bodies. Samples of the infrequent breakage defects were supplemented with defective cigarette groups confirmed by manual inspection. The converted RKNN model was run on the RK3588 platform to produce a detection result for each image, with manually reviewed results serving as the reference.Each image was treated as one statistical unit. A normal image was counted as correctly identified if it triggered no defect alarm. A defective image was counted as correctly identified if both the predicted defect category and location matched the annotation at an IoU of at least 0.5. If an image contained multiple defects, all defects had to be matched. Accuracy was calculated as the number of correctly identified images divided by the number of inspected images, multiplied by 100%. The results are shown in Table 11.

**Table 11.** Results and error types in the simulated production-line test.

| Category | Test images | Correctly identified | Missed detections | Misclassifications | False alarms | Accuracy/% |
|---|---|---|---|---|---|---|
| Normal cigarette group | 8,000 | 7,851 | 0 | 0 | 149 | 98.14 |
| Flattened end | 260 | 253 | 5 | 2 | 0 | 97.31 |
| Reversed cigarette | 240 | 232 | 6 | 2 | 0 | 96.67 |
| Missing tobacco | 220 | 215 | 3 | 2 | 0 | 97.73 |
| Broken cigarette tip | 60 | 58 | 1 | 1 | 0 | 96.67 |
| Broken cigarette body | 40 | 39 | 1 | 0 | 0 | 97.50 |
| Total | 8,820 | 8,648 | 16 | 7 | 149 | 98.05 |

As shown in Table 11, the accuracy for normal cigarette groups was 98.14%. The accuracies for flattened ends, reversed cigarettes, and missing tobacco were 97.31%, 96.67%, and 97.73%, respectively; those for broken cigarette tips and broken cigarette bodies were 96.67% and 97.50%. Of the 8,820 images, 8,648 were correctly identified, giving an overall accuracy of 98.05%. Every category exceeded the 95% accuracy requirement for production. Among the 172 incorrectly identified images, there were 16 missed detections, 7 misclassifications, and 149 false alarms on normal cigarette groups. Because samples of breakage defects were limited, their accuracy requires further validation through longer on-site operation and a larger sample set.

## 5 Conclusions

To meet the accuracy and computational efficiency requirements of cigarette-group end-face defect detection in the cigarette feed channel, this study proposed a lightweight detection method based on an improved DEIM-S model. A MobileNetV4 backbone reduced the computational cost of feature extraction; the proposed C2F-IDWB module strengthened the extraction and fusion of multiscale defect features; and an EMA attention mechanism at the encoder's P5 feature output enhanced responses in defective regions. Together, these changes improved cigarette appearance defect detection accuracy while reducing model complexity.

On a self-developed dataset containing flattened ends, reversed cigarettes, missing tobacco, broken cigarette tips, and broken cigarette bodies, the proposed model achieved AP50, AP75, and AP50:95

values of 99.1%, 97.0%, and 81.07%, respectively. These were 3.0, 2.4, and 3.54 percentage points higher than those of the DEIM-S baseline. Computational cost decreased from 24.84 G to 17.82 G FLOPs, a reduction of 28.3%, while model size decreased from 39.2 MB to 30.6 MB, a reduction of 21.9%. Under the comparative experimental conditions used in this study, the detection speed was 189.56 FPS. Ablation and module comparison experiments confirmed the effectiveness of the modifications. Detection results and heatmap visualizations further showed that the improved model focused more strongly on defective regions and reduced missed detections of some targets. In addition, the PTH → ONNX → RKNN conversion procedure and deployment on the RK3588 edge platform provide a practical route for implementing the method on edge devices.

The dataset currently covers a limited range of cigarette varieties and acquisition settings, and contains relatively few samples of broken cigarette tips and bodies. The model's adaptability to other cigarette varieties and complex production conditions therefore requires further validation. Future work will add samples of infrequent defects from operating production lines, expand the dataset across different imaging conditions, and improve localization accuracy for small defects. Operator compatibility and inference optimization on the edge device will also be investigated to assess detection speed and operational stability during continuous production.

**Author Contributions:** Conceptualization, X.Y., L.H.; methodology, X.Y., L.H.; software, X.Y.; validation, X.Y., L.H.; formal analysis, X.Y.; investigation, X.Y., Y.C., Z.L. and M.Z.; resources, X.Y., L.H., Z.L. and Y.C.; data curation, X.Y.; writing—original draft preparation, X.Y.; writing—review and editing, X.Y, L.H; visualization, X.Y.; supervision, Z.L. and M.Z.; project administration, Y.C.; funding acquisition, X.Y. and Y.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the General Program of Hubei Provincial Natural Science Foundation under Grant 2023AFB381, and in part by the Youth Talent Project of the Science and Technology Research Program of the Hubei Provincial Department of Education under Grant Q20231412, and in part by the Key Research and Development Project of the Hubei Province Science and Technology Innovation Plan under Grant 2025BAB045.

**Data Availability Statement:** The model configuration files, code, training parameters, and dataset- splitting description are available from the corresponding author upon reasonable request.

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

## References

1. Leng, W.; Mu, R. Barriers to tobacco control in China: A narrative review. Societies 2020, 10, 101. https://doi.org/10.3390/soc10040101.
2. Chan, K.H.; Xiao, D.; Zhou, M.; Peto, R.; Chen, Z. Tobacco control in China. Lancet Public Health 2023, 8, e1006–e1015. https://doi.org/10.1016/S2468-2667(23)00242-6.
3. Chen, T.; Zhu, J.; Tsuei, S.; Jiang, Y.; Zheng, Y.; Xu, D.; Fu, H. Health and economic effects of increased taxation on tobacco, alcohol, and sugar-sweetened beverages in China: A modelling study. Lancet Public Health 2025, 10, e1025–e1035. https://doi.org/10.1016/S2468-2667(25)00256-7.
4. Ding, Y.; Zhou, H.; Wu, H.; Ma, C.; Yuan, G. SCS-YOLO: A defect detection model for cigarette appearance. Electronics 2024, 13, 3761. https://doi.org/10.3390/electronics13183761.
5. Zhang, C.; Yun, L.; Yang, C.; Chen, Z.; Cheng, F. LRNTRM-YOLO: Research on real-time recognition of non-tobacco-related materials. Agronomy 2025, 15, 489. https://doi.org/10.3390/agronomy15020489.
6. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems; 2015; Volume 28, pp. 91–99.
7. Feng, Y.; Zhao, X.; Tian, R.; Liang, C.; Liu, J.; Fan, X. Research on an intelligent seed-sorting method and sorter based on machine vision and lightweight YOLOv5n. Agronomy 2024, 14, 1953. https://doi.org/10.3390/agronomy14091953.

8. Wu, H.; Zhu, R.; Wang, H.; Wang, X.; Huang, J.; Liu, S. Flaw-YOLOv5s: A lightweight potato surface defect detection algorithm based on multi-scale feature fusion. Agronomy 2025, 15, 875. https://doi.org/10.3390/agronomy15040875.
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016; pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot MultiBox detector. In Computer Vision—ECCV 2016; Springer: Cham, Switzerland, 2016; pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.
11. Yuan, G.; Liu, J.; Liu, H.; Ma, Y.; Wu, H.; Zhou, H. Detection of cigarette appearance defects based on improved YOLOv4. Electron. Res. Arch. 2023, 31, 1344–1364. https://doi.org/10.3934/era.2023069.
12. Liu, H.; Yuan, G.; Yang, L.; Liu, K.; Zhou, H. An appearance defect detection method for cigarettes based on C-CenterNet. Electronics 2022, 11, 2182. https://doi.org/10.3390/electronics11142182.
13. Dong, P.; Wang, Y.; Yu, Q.; Feng, W.; Zong, G. AMC-YOLO: Improved YOLOv8-based defect detection for cigarette packs. IET Image Process. 2024, 18, 4873–4886. https://doi.org/10.1049/ipr2.13272.
14. He, L.; Wei, R.; Ding, Y.; Huang, J.; Wei, X.; Li, R.; Wang, S.; Fu, L. Enhancing dense-scene millet appearance quality inspection based on YOLO11s with overlap-partitioning strategy for procurement. Agronomy 2025, 15, 1284. https://doi.org/10.3390/agronomy15061284.
15. Zhang, P.; Zhang, X.; Huang, J.; Guan, X.; Yu, S.; Liu, Z. YOLOv8n-based algorithm for cigarette defect detection and its deployment. J. Comput. Appl. 2025, Suppl. 2, 305–310.
16. Gao, X.; Yuan, F.; Huo, Y. PCB defect detection algorithm based on improved DEIM. Chin. J. Liq. Cryst. Displays 2026, 41, 700–716. DOI: 10.37188/CJLCD.2026-0045. Available online: https://cjlcd.lightpublishing.cn/zh/article/doi/10.37188/CJLCD.2026-0045/.
17. Qin, D.; Leichner, C.H.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4: Universal models for the mobile ecosystem. In Computer Vision—ECCV 2024; Springer: Cham, Switzerland, 2024; pp. 78–96. https://doi.org/10.1007/978-3-031-73661-2_5.
18. Yin, J.; Lyu, Y.; Suo, K.; Huang, X. PCB defect detection algorithm based on EfficientNetV2. J. Comput.-Aided Des. Comput. Graph. 2025, 37, 1260–1269. https://doi.org/10.3724/SP.J.1089.2023-00551.
19. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; pp. 1577–1586. https://doi.org/10.1109/CVPR42600.2020.00165.
20. Hu, M.; Chen, H.; Xu, W.; Guo, C.; Liu, Q. Surface defect detection on lightweight YOLOv8 steel incorporating MobileNetV3. Sci. Technol. Eng. 2025, 25, 6831–6840. https://doi.org/10.12404/j.issn.1671-1815.2404413.
21. Cao, Z.; Wang, M.; Lu, Y.; Wu, Z.; Wang, J. Research on the lightweight wheel hub defect detection algorithm based on YOLOv8. Mach. Electron. 2025, 43, 31–36.
22. Yue, F.; Peng, Y.; Su, Z.; Zhang, G.; Lian, C.; Yang, B.; Fang, Z. Open-set source cell-phone identification method based on feature interaction and representation enhancement. J. Comput. Appl. 2025, 45, 3813–3819. https://doi.org/10.11772/j.issn.1001-9081.2024121815.
23. Li, D.; Yao, A.; Chen, Q. PSConv: Squeezing feature pyramid into one compact poly-scale convolutional layer. In Computer Vision—ECCV 2020; Springer: Cham, Switzerland, 2020; pp. 615–632. https://doi.org/10.1007/978-3-030-58589-1_37.
24. Meng, R.; Lu, Z.; Cong, Y. Oil injection hole detection algorithm based on lightweight convolution neural network. J. Anhui Univ. Technol. (Nat. Sci.) 2023, 40, 166–172, 190. https://doi.org/10.12415/j.issn.1671-7872.22230.
25. Yang, Z.; Guan, Q.; Yu, Z.; Xu, X.; Long, H.; Lian, S.; Hu, H.; Tang, Y. MHAF-YOLO: Multi-branch heterogeneous auxiliary fusion YOLO for accurate object detection. arXiv 2025, arXiv:2502.04656.

https://doi.org/10.48550/arXiv.2502.04656.

26. Jocher, G.; Qiu, J.; Liu, M.; Lyu, S.; Akyon, F.C.; Kalfaoglu, M.E. Ultralytics YOLO26: Unified real-time end-to-end vision models. arXiv 2026, arXiv:2606.03748. https://doi.org/10.48550/arXiv.2606.03748.
27. Li, H.; Wang, X.; Bu, Y.; David, C.C.; Chen, X. YOLOv8-Orah: An improved model for postharvest Orah mandarin (Citrus reticulata cv. Orah) surface defect detection. Agronomy 2025, 15, 891. https://doi.org/10.3390/agronomy15040891.
28. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision; 2018; pp. 839–847. https://doi.org/10.1109/WACV.2018.00097.
29. Wang, J.; Liao, C.; Li, D.; Zhao, Z.; Chen, J.; Yang, K. An efficient large kernel convolution network designed for neural processing unit. Eng. Appl. Artif. Intell. 2025, 142, 109887. https://doi.org/10.1016/j.engappai.2024.109887.