A Multi-Scale Deformable Convolutional Neural Network with Adaptive Adjustment for Robust Packaging Image Recognition

Wen Sun
College of Art and Design, Guangdong University of Science and Technology, Dongguan City, Guangdong Province, 532083, China
E-mail: 13728181407@163.com

Keywords: deep convolutional neural network (DCNN), packaging images, adaptive adjustment module (AAM), feature extraction, automatic recognition

Received: June 3, 2025

This paper presents an innovative image recognition model for package inspection, designed to fulfill the demands of real-time and precise categorization in difficult industrial environments. Traditional techniques reliant on human feature extraction frequently underperform when confronted with lighting variability, background interference, and deformation of package items. To mitigate these limitations, the proposed model integrates a multi-scale convolutional architecture that captures both local and global characteristics through the use of parallel convolutional filters of varying sizes. An adaptive adjustment method is incorporated into the network to dynamically alter the placement of convolutional operations according to image content, hence improving flexibility and feature representation. A thorough data augmentation strategy incorporating geometric transformation, brightness modification, and semantic-level blending is implemented to boost the model's robustness and generalization capacity. Experiments performed on a bespoke industrial packaging dataset comprising 10,000 labeled images reveal that the proposed model attains a classification accuracy of 96.8 percent, a recall of 95.3 percent, and an F1-score of 93.8 percent, with an inference time of 11.2 milliseconds and a parameter count of 21.3 million. In comparison to current deep learning architectures like Residual Networks, the model demonstrates considerable enhancements in accuracy and speed. These results support its appropriateness for practical packaging inspection systems.

Povzetek: Razvit je večmerni DCNN z deformabilnimi konvolucijami in prilagoditvenim modulom (AAM) ter hibridno augmentacijo. Je robusten, realnočasovni model za prepoznavo embalažnih slik. Na podatkovni zbirki PackNet-10K doseže odlične rezultate, prekaša ResNet/Inception pri hitrosti in točnosti.

1 Introduction

In the current context of intelligent manufacturing and Industry 4.0, the packaging industry is facing an increasing demand for intelligence. With the advancement of technology, traditional packaging image recognition methods, especially those relying on manual feature extraction techniques such as SIFT and HOG, have gradually shown problems of low efficiency and poor robustness, and cannot meet the requirements of modern industry for high speed, high precision, and high robustness [1]. Traditional manual feature extraction often struggles to handle the complex backgrounds, lighting changes, and deformations of images in various practical application scenarios [2].

The advantage of its ability to automatically extract features not only improves the automation level of feature learning but also dramatically enhances the accuracy and efficiency of image recognition [3]. However, when applied to the recognition of packaging images, existing deep learning models still have some shortcomings, particularly in addressing common issues such as complex backgrounds, lighting variations, and object deformations in packaging images. The performance of these models has not yet reached an ideal level. Therefore, building a high-precision and highly robust packaging image recognition model that can adapt to complex scenarios is of great practical significance for promoting the intelligent development of the packaging industry and improving production efficiency. Currently, research in the field of packaging image recognition has made significant progress both domestically and internationally. Traditional feature extraction methods, such as those based on color histograms and texture analysis, have, to some extent, solved the problem of feature extraction. However, these methods have weak generalization ability and are difficult to adapt to the ever-changing packaging image scenes. With the introduction of Convolutional Neural Networks (CNNs), shallow CNN models such as LeNet-5 have achieved good results in simple image classification tasks [4]. However, shallow models such as LeNet-5 are unable to effectively capture multi-scale image features, which limits their application in packaging image recognition.

Deep convolutional neural networks, including models like ResNet and Inception, demonstrate improved feature expression capabilities through residual connections and multi-branch structures, which produce superior results in image classification and object detection tasks. The most advanced deep learning models currently face challenges involving automated feature extraction for packaging images when processing complex background environments, along with diverse lighting conditions and packaging shape variations. Research on packaging image analysis requires immediate enhancements to understand complex packaging environments better [5].

Researchers developed a precise and sturdy model for recognizing packaging images through the implementation of deep convolutional neural networks (DCNN). The traditional DCNN architecture received an enhancement through the development of multi-scale convolutional layers and feature fusion modules, which work to extract local and global features better while introducing an adaptive convolution kernel adjustment mechanism. The model achieves adaptive parameter optimization through convolution kernel modification, which allows it to flexibly respond to complex backgrounds and changing lighting conditions.

2 Related work

Recently, packaging image recognition technology has become a key part of intelligent packaging production lines. However, traditional image feature extraction methods have shown significant limitations when dealing with complex backgrounds and large-scale data. Traditional manual feature extraction methods, such as Scale Invariant Feature Transform (SIFT) and Local Binary Patterns (LBP), have, to some extent, solved the problem of image feature extraction, but they are not a foolproof solution [6], [7].

Medus et al. [3] developed a real-time food packaging inspection system using hyperspectral imaging combined with convolutional neural networks to detect contaminated heat-sealed trays. The system classified up to eleven types of contaminants (e.g., plastic, rubber) with over 94% accuracy and a processing time of 70–105 milliseconds, enabling inspection speeds of up to 14 trays per second. A custom dataset and flexible CNN configurations allowed optimization for either accuracy or fault rejection, demonstrating the practical effectiveness of deep learning in high-speed industrial inspection. Zhang et al. [8] introduced an improved fully convolutional network for picture segmentation in packaging design, drawing inspiration from natural language processing methodologies. The model integrated superpixels, multi-branch networks, and attention mechanisms. It attained an accuracy of 96.84% and a segmentation error rate of 1.42%, indicating enhanced efficiency and precision relative to conventional approaches. Siddiqua et al. [9] created a Faster R-CNN model integrated with InceptionV2 to identify dengue mosquitoes utilizing photos from diverse surroundings. The model surpassed R-FCN and SSD, attaining a detection accuracy of 95.19%. Evaluation metrics encompassed precision, recall, false positives, and false negatives.

Compared to traditional methods, the emergence of deep learning, intense convolutional neural networks (CNNs), has brought revolutionary progress to the field of image recognition. AlexNet, as the first network to achieve breakthrough results in large-scale image classification competitions, has for the first time validated the effectiveness of deep CNN in image recognition. AlexNet utilizes a deeper network structure and ReLU activation function, significantly improving the accuracy of image classification [10]. However, as the depth of the network increases, the problems of gradient vanishing and exploding gradually become apparent, limiting the possibility of further deepening the network. To address this issue, ResNet successfully overcame the vanishing gradient problem by introducing a residual learning mechanism, enabling the network to reach deeper layers while maintaining stable training, thereby improving image recognition performance. At the same time, with the continuous innovation of network structures, attention mechanisms are gradually being applied in the field of image recognition. For example, the SE module (Squeeze and Excitation) enhances the ability to distinguish the importance of features by weighting the channels, thereby improving the performance of the model in complex scenes [10].

Although deep learning technology has achieved significant results in the field of image recognition, it still faces several challenges in packaging image recognition. The packaging industry faces a considerable issue because packaging image datasets are extremely limited [11]. The packaging industry has unique requirements, which make publicly available packaging datasets very limited in content and insufficient for training deep learning models with strong generalization capabilities [12]. The packaging image recognition process faces significant challenges due to background interference that occurs in packaging production line environments. Production line environments feature uneven lighting and object occlusion alongside object deformation that degrade packaging image quality and compromise feature recognition precision and model performance. Current research requires immediate attention to solve the critical problem of improving packaging image recognition accuracy and robustness when operating in complex and varied interference environments [13]. Wang and Song [14] introduced a convolutional neural network-based technique for classifying network traffic and detecting anomalies, overcoming the shortcomings of conventional methods in managing intricate traffic patterns. Their model attained high accuracy on the CIC-IDS2017 (98.5%) and ISCX VPN-NOVPN (99.2%) datasets, while also markedly enhancing recall and F1 score. Through the examination of several network architectures, they diminished the false alarm rate to 1.5%, showcasing significant robustness and adaptability in practical settings. Lu [15] utilized Convolutional Neural Networks to assist in the assessment of promotional graphic designs showcasing traditional Fengxiang clay sculptures. The outputs of CNN

corresponded with manual evaluations, validating its efficacy. The research demonstrated that CNNs improved the examination and aesthetic quality of culturally inspired designs.

Numerous studies have proven that traditional methods for image feature extraction fail to handle large datasets and complicated visual systems effectively. The essential progress of deep learning through convolutional neural networks and attention mechanisms faces multiple challenges in image recognition because of inadequate dataset sizes and background complexities. The research community needs to address these challenges by developing improved model architectures and investigating innovative data enhancement techniques and dynamic learning approaches to enhance image recognition technology to an intelligent level.

2.1 Research gaps and novelties

Notwithstanding recent progress in deep learning for image recognition, packaging inspection tasks in industrial settings are still inadequately investigated due to distinct challenges, such as lighting variability, intricate backgrounds, deformation of packaging shapes, and a scarcity of annotated datasets. Conventional feature extraction techniques like SIFT and LBP, although efficient in controlled environments, do not transfer well across varied real-world situations. Even cutting-edge CNN designs such as ResNet-50 and Inception-V3 demonstrate restricted adaptation to occlusion and geometric distortion, resulting in diminished recognition accuracy in dynamic manufacturing settings. Furthermore, the majority of current models employ static receptive fields and are unable to modify spatial sample positions according to the local visual environment, hence limiting their efficacy in identifying small or irregularly shaped targets typically encountered in packing applications.

This study introduces an innovative packaging image recognition model utilizing a Deep Convolutional Neural Network augmented by two principal components: (1) a Multi-Scale Convolutional Block that integrates local and global features via parallel 3×3 and 5×5 kernels, and (2) an adaptive adjustment module that employs deformable convolution to modify the receptive field through learned offsets dynamically. A hybrid data augmentation technique is developed to enhance resilience by modeling industrial disturbances, including illumination variations, occlusion, and deformation. These enhancements collectively improve the model's capacity to generalize across intricate visual contexts while preserving high recognition accuracy.

Experimental findings on the PackNet-10K dataset indicate that the proposed model substantially surpasses baseline and state-of-the-art techniques in terms of accuracy and inference speed, while preserving a minimal parameter count appropriate for real-time industrial applications.

Table 1 delineates a comparative analysis of traditional methods and contemporary deep learning models, summarizing the performance of various prevalent packaging image recognition techniques about classification accuracy, parameter count, and inference speed, as derived from our experiments utilizing the PackNet-10K dataset. This comparative research emphasizes the shortcomings of state-of-the-art (SOTA) models. It illustrates the necessity for an adaptive design that can accommodate geometric distortions, illumination fluctuations, and intricate packaging scenarios.

Table 1: Summary of baseline methods on the PackNet-10K Dataset

Method	Accuracy (%)	Parameters (M)	Inference Time (ms)	Key Characteristics
SIFT + SVM	68.3	N/A	115.6	Manual feature extraction, sensitive to deformation
LBP + SVM	72.6	N/A	109.3	Texture-based, limited scale robustness
ResNet-50	86.2	25.6	28.5	Deep residual network, poor small object detection
Inception-V3	88.1	23.9	34.2	Multi-branch CNN, slow inference
YOLOv5s	89.4	7.9	16.5	Real-time detection, weaker classification accuracy
Proposed Model	96.8	21.3	11.2	Adaptive MSCB + AAM, fast, high precision

Table 1 illustrates those conventional approaches like SIFT and LBP do not generalize well across differences in illumination and shape, and demonstrate significant processing delay. CNN-based models such as ResNet-50 and Inception-V3, while providing enhanced accuracy, exhibit limited adaptation to non-rigid deformations and demonstrate inefficiency in managing small-scale packaging features. YOLOv5s provides rapid inference; nonetheless, its categorization efficacy is inadequate for detailed packing categories. The proposed model surpasses all baselines by incorporating multi-scale feature fusion and adaptive deformable convolution, resulting in enhanced accuracy and real-time performance.

3 System architecture and algorithm design

3.1 System architecture

The article proposes a method of hierarchical modularization to enhance system performance by optimizing the functionalities of individual modules. The system contains the following basic parts: feature

extraction network, Backbone, adaptive adjustment module (AAM), and classifier that work together to improve recognition accuracy and system adaptability according to Figure 1.

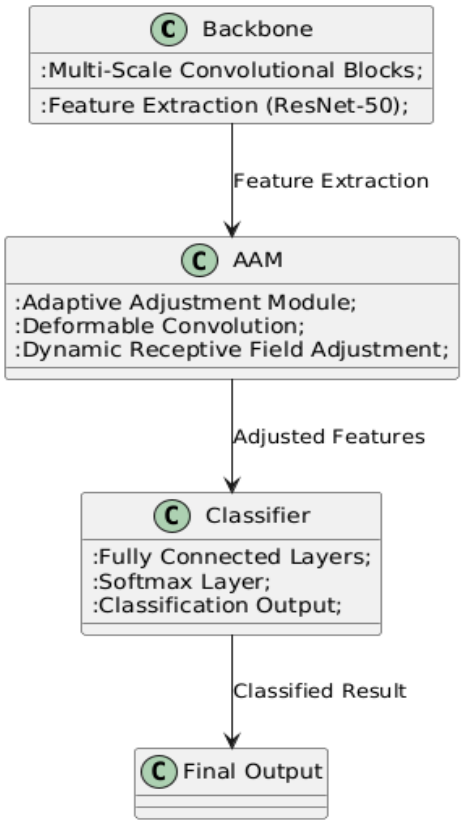

Figure 1: Overall scheme architecture diagram

Figure 2 illustrates the comprehensive pipeline of the proposed model, encompassing data preprocessing, hybrid augmentation procedures, and essential components like the feature extraction backbone, adaptive adjustment mechanism, and classification layers. This comprehensive flow delineates the complete recognition procedure tailored for industrial packaging photos.

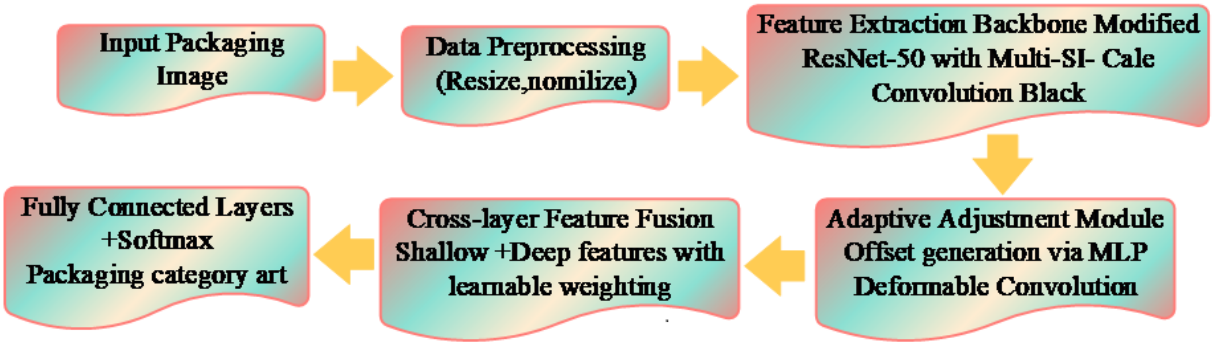

Figure 2: End-to-end workflow of the multi-scale adaptive CNN for packaging image recognition

The foundational Backbone component serves to identify essential data characteristics during input processing. The Backbone network structure receives ResNet-50 design elements, which lead to performance enhancements. ResNet-50 operates as a deep residual network, which was developed to solve the gradient vanishing issue in deep neural network training by using residual connections. The network contains several residual modules that create connections to merge previous layer input with current layer output, thus supporting effective gradient flow in deep networks. The model version improves the feature extraction capability of ResNet-50 through its transformation into a Multi-Scale Convolutional Block (MSCB). The network

improves its feature information processing capacity by implementing multiple convolutional layers, which function at different scales to identify distinct image elements. The network's multi-scale design enables it to detect targets of different sizes and shapes while also being able to handle various image input dimensions. A mathematical representation of the ResNet-50 residual structure exists through this formula [16]:

$$y = F(x, W_i) + x \tag{1}$$

The equation uses the variable x to represent input data while $F(x, W_i)$ stands for the transformation which consists of convolutional layers along with activation functions and other operations, and W_i indicates layer weights that produce the output y. It explains how the network is structured, combining the input data with its modified version using skip connections to tackle the issue of gradient vanishing in deep learning networks.

The key feature of the model consists of the Adaptive Adjustment Module (AAM) that modifies its receptive field automatically based on changing target conditions. Each specific neuron in the network has a particular limit on how much input data it can process simultaneously. The AAM module modifies the convolution kernel structure and size through deformable convolution to create automatic adjustments that enable the network to better handle different target scales and shapes. By using offsets, deformable convolution allows convolution operations to flexibly handle both regular grid patterns and local input data patterns. The approach demonstrates significant performance improvements when applied to complicated situations. The mathematical expression for deformable convolution operations is represented by the following formula [16]:

$$y(x) = \sum_{i,j} w(i,j)x(x + \Delta x(i,j), y + \Delta y(i,j)) \tag{2}$$

The mathematical expression utilizes the following elements for its operation: $x(x, y)$ as the input image, $w(i, j)$ functions as the convolution kernel weight, and $\Delta x(i, j)$, $\Delta y(i, j)$ represent the learned offsets that modify the convolution field. The formula elucidates the methodology for changing the convolutional receptive field to capture intricate feature information effectively.

The classifier component evaluates the classification for extracted feature information through its analysis. The system includes a series of Fully Connected Layers along with SoftMax layers, which compose this specific component. Fully connected layers execute two major functions, which are to map nonlinear relations between features, produce activation scores for class labels, and generate probabilities for class membership. The SoftMax layer transforms activation scores into distribution probabilities, which enables the model to generate predictions for every category. The SoftMax function is defined as follows [16]:

$$P(y = c|\mathbf{x}) = \frac{\exp(z_c)}{\sum_c \exp(z_c)} \tag{3}$$

The activation value of category c becomes z_c in Equation (3), while the function $\exp(z_c)$ calculates the sum of all category activation values. A probability value for each category results from the SoftMax function, which guarantees that all category probabilities combine to equal 1. This process enables the classifier to output the prediction probability of each category based on the feature information of the input data, thereby achieving the final classification.

Throughout the entire training process of the model, all modules collaborate to optimize the model parameters using the backpropagation algorithm. During the training process, errors are propagated layer by layer from the classifier to the Backbone network and AAM module via the gradient descent algorithm. The weights and offsets in each module are adjusted to optimize the model's performance on the training set. The loss function usually uses the Cross Entropy Loss function to measure the difference between the predicted category and the correct category. The formula for Cross Entropy Loss is [16]:

$$\mathcal{L} = -\sum_c y_c \log(\hat{y}_c) \tag{4}$$

In Equation (4), y_c is the binary encoding of the real label, $\hat{y}_c$ It is the category probability predicted by the model. By minimizing the loss function, the model can continuously adjust its parameters to improve classification performance.

To guarantee reproducibility and practical application, the mathematical models outlined in Eqs. (1) to (9) have been comprehensively implemented utilizing the PyTorch deep learning framework. The residual block structure in Eq. (1), based on ResNet-50, is executed by sequential convolutional modules incorporating batch normalization and ReLU activations, complemented with skip connections to facilitate efficient gradient propagation. The deformable convolution operation described in Eqs. (2), (7), and (9) are implemented using the Deformable Convolution v2 (DCNv2) module from the MMCV library, which facilitates learnable offset generation and adaptable sample sites. The Adaptive Adjustment Module (AAM) dynamically alters the receptive field by employing a lightweight multilayer perceptron (MLP) architecture to correlate feature maps with learnt offsets, as delineated in Eq. (8). This MLP is constructed using a series of 1×1 convolutional layers succeeded by ReLU activations, enabling it to produce offset tensors utilized in the deformable convolution layers effectively. The classification layer described in Eq. (3) employs a fully connected layer succeeded by a SoftMax activation, whilst the loss function in Eq. (4) is executed using the cross-entropy criterion with label smoothing (smoothing factor = 0.1). All parameters, encompassing convolutional weights and bias tensors, are concurrently tuned during training utilizing the Adam optimizer. The learning rate is adjusted by cosine annealing, and early stopping is used when validation loss fails to improve over five consecutive epochs. The systematic incorporation of theoretical frameworks into the training process guarantees clarity, coherence, and reproducibility in the model's execution.

3.2 Multi-scale feature fusion strategy design

In deep learning systems, image processing requirements often necessitate feature aggregation across multiple scales to enhance the model's understanding of targets with diverse dimensions. The model employs a Multi-Scale Feature Fusion Strategy (MSCB) to improve its capacity for handling image features at various scales. The MSCB architecture performs image feature extraction through parallel convolution kernels of varying sizes, including 3x3 and 5x5 structures. Different sizes of convolution kernels produce varying receptive fields, which enables 3x3 kernels to detect precise local elements and 5x5 kernels to collect extended relationship data. The network operates by simultaneously evaluating multiple image scales, which allows the collection of more comprehensive feature sets.

The feature extraction strategy at multiple scales combines features by merging features of different scales through channel connections. The proposed multi-scale feature fusion approach combines output feature maps from convolutional kernels of varying sizes ($3x3$ and $5x5$) via channel-wise concatenation. This process concatenates the feature maps along the channel dimension, yielding a more comprehensive and varied representation while maintaining the spatial resolution of each input. In contrast to channel addition, which necessitates uniform channel lengths and executes element-wise summing, channel concatenation augments the number of channels, enabling the network to preserve unique information derived from each kernel size. The combined feature map is then subjected to a 1×1 convolutional layer to diminish dimensionality and improve feature integration efficacy. The updated fusion process is mathematically expressed as: $F_{fusion} = F_{3x3}, F_{5x5}$ where F_{3x3} and F_{5x5} represent the feature maps from the $3x3$ and $5x5$ convolutional branches, respectively [16]:

$$F_{fusion} = F_{3x3}, F_{5x5} \quad (5)$$

A mathematical equation expressed as Equation (5) uses the term F_{fusion} to represent concatenation in channel dimensions. Network performance improves when the method for feature fusion combines local and global information.

The approach focuses on integrating multi-scale features at the same processing level while constructing effective relationships between different layers. The system acknowledges that basic features detect exact edge data while advanced features recognize complex semantic concepts. The system enhances its understanding of image patterns through the integration of superficial and complex data, which establishes connections between local parts and broad-scale content. The system now uses a cross-layer connection approach. In this particular method, shallow features F_{low} and deep features F_{high} combine with a certain ratio while following a specific formula [16]:

$$F_{fusion} = \alpha \cdot F_{low} + (1 - \alpha) \cdot F_{high} \quad (6)$$

Equation (6) contains a weight coefficient α which the network modifies between 0 and 1 to control the deep and shallow feature fusion ratio. The superior analysis performance of the network occurs because it makes real-time changes to the ratio of shallow and deep features for task completion.

To enhance operational performance, this article incorporates deformable convolution technology within feature design strategies to improve model adaptability for different target scale and shape conditions. Deformable convolution uses an offset mechanism to transform the convolution kernel shape, which enables the network to automatically adjust its receptive field based on input image features. The network performs better at recognizing complex image modifications because of its ability to automatically adapt to changing input conditions. The output of the multi-scale feature fusion block undergoes deformable convolution for flexible spatial adaptation. The deformable convolution procedure adheres to the mathematical formulation presented in Equation (2), wherein learnt offsets modify the kernel's sample positions to more effectively capture intricate geometric variations.

The development of a multi-scale feature fusion method represents a crucial element in image processing because it strengthens the model against difficult situations that include intricate scenes and various target scales, together with highly deformed objects. The model demonstrates superior image comprehension by utilizing scale-based feature fusion alongside cross-layer connections and deformable convolutions to achieve better flexibility and adaptability. The innovative feature fusion techniques, when united, produce better model outcomes, which demonstrate enhanced accuracy in multiple operational tasks.

4 Optimization of adaptive convolutional kernel adjustment mechanism

A method employs dynamic convolution kernel adjustment by creating image-based kernel offsets that optimize convolutional network performance for complex geometric images and deformed objects. The standard convolution mechanism uses a fixed-size kernel structure, which cannot modify its receptive field according to different target areas containing significant deformations and scale changes, thus causing feature extraction errors. The solution to this problem involves the successful implementation of dynamic kernel shape and position variations, which result in enhanced model flexibility and accuracy.

The network generates convolution kernel offsets from image features to adapt the receptive field dynamically. To generate convolution kernel offsets, an MLP model uses the input feature map to produce output results. An offset ΔW transforms the convolution kernel's spatial properties through processing of the input feature map F_{input} within the MLP module. The mathematical formula for this process can be expressed as: The

network needs to use image features to develop convolution kernel offsets, which allow adaptive adjustment of the kernels. Through the use of an MLP model, the input image feature map generates convolution kernel offsets as its output. The MLP module processes the input feature map F_{input} to calculate an offset ΔW that adjusts the spatial properties of the convolution kernel. The mathematical operation is:

$$\Delta W = f_{MLP}(F_{input}) \quad (7)$$

The network employs an MLP to convert input feature maps into spatial sampling offsets for convolutional kernels. The offset ΔW is calculated as: $\Delta W = f_{MLP}(F_{input})$. where F is the input feature map and f_{MLP} denotes the offset prediction function. The MLP consists of two consecutive 1×1 convolutional layers with 64 and $2k^2$ output channels respectively, where K is the kernel size (e.g., 3). ReLU activation is utilized between layers, and batch normalization is incorporated to enhance training stability. The final result presents horizontal and vertical offsets for each sampling position within the deformable kernel. This approach allows the network to dynamically learn adaptable receptive fields while maintaining a minimal parameter count. The MLP and deformable convolution layers are trained concurrently by backpropagation.

Equation (7) contains the function f_{MLP} which conducts mapping transformations through the multi-layer perceptron while F_{input} depicts the input feature map and ΔW demonstrates the convolution kernel offset. Through dynamic adjustments of its convolution kernel, the network adapts to different deformation types and local structures to achieve improved performance.

As defined in Equation (7), the convolutional kernel offset $\Delta W = f_{MLP}(F_{input})$ is generated by passing the input feature map F_{input} through a learnable MLP. This MLP is implemented using two stacked 1×1 convolutional layers with ReLU activation in between. For a kernel of size $k \times k$, the output $\Delta W \in \mathbb{R}^{2k^2} \times H \times W$ contains pixel-wise horizontal and vertical offsets for each sampling location in the kernel. The offsets facilitate the adjustment of the sample grid in the deformable convolution procedure, allowing the convolutional kernel to dynamically modify its receptive field by the geometric attributes of the input image. Bilinear interpolation is utilized at fractional places to preserve differentiability. The parameters of the convolutional kernel and the MLP are concurrently trained using backpropagation during model optimization.

Convolution relies on a calculated offset value to adjust the convolution kernel, changing its receptive field properties. This process is transformed by Deformable Convolution. The fixed position of each convolution kernel element in standard convolution operations contrasts with the dynamic sampling position of the convolution kernel elements in deformable convolution, which enables flexible shape adaptation. The mechanism works effectively to process scale changes along with rotations, translations, and other geometric deformations.

The specific implementation serving deformable convolution refers to DCNv2 (Deformable Convolution v2), which allows the convolution operation to modify its receptive field through offset-based adjustments. The adaptive adjustment technique employs deformable convolution, allowing the receptive field to dynamically adapt to alterations in shape and structural distortions in packing pictures.

The equation contains $x(x, y)$ for image pixel values alongside the convolution kernel weight $w(i, j)$ and the offset changes $\Delta x(i, j)$ and $\Delta y(i, j)$ on the x-axis and y-axis. During training, the offset learns the proper values. The convolutional kernel demonstrates flexibility in adjusting its receptive field in response to changes in image features, thereby enhancing the model's performance on deformed objects.

The research paper enhances deformable convolution modeling by introducing a joint optimization method for the convolutional layer and offset in DCNv2.The simultaneous optimization of the convolution kernel offset and weight inside the network enhances the combination between feature extraction requirements and geometric deformation modeling. The network requires learning to both perform effective feature extraction and modify the convolution kernel positions, along with their shapes, based on feature information. Through joint training, the model can automatically adjust the receptive field in each layer, thereby better adapting to geometric changes in the image, while also introducing multi-scale feature fusion. Different scales of features carry different information in images, and combining multi-scale features can help networks better understand the various levels and details of objects. In the adaptive convolution kernel adjustment mechanism, through multi-scale convolution kernel adjustment, the network can dynamically adjust the receptive field at different scales, further enhancing its adaptability to geometric deformation of images.

5 Experimental and simulation analysis

5.1 Experimental design and dataset construction

To verify the effectiveness of the proposed DCNN model in packaging image recognition tasks, this study conducted systematic experiments from three aspects: dataset construction, experimental environment configuration, and comparative scheme design. Firstly, in response to the scarcity of data in the packaging industry, a multi-scenario packaging image dataset, PackNet-10K, was constructed in collaboration with a specific intelligent packaging equipment manufacturer. This dataset comprises 10,000 high-resolution (1920 × 1080) images, covering six major categories of packaging, including food, medicine, and daily chemical products, with each category containing 1,200 to 2,000 samples. The data collection scenario simulates a real industrial

environment, covering four typical interference conditions:

1) Lighting changes (low light, high exposure, dynamic fill light);

2) Complex background (conveyor belt reflection, product stacking obstruction);

3) Deformation interference (packaging bag wrinkles, box body compression deformation);

4) Motion blur (image blur caused by high-speed movement of the production line). The dataset is divided into a training set (7,000 images), a validation set (2,000 images), and a testing set (1,000 images), with a ratio of 7:2:1, ensuring class distribution balance through random stratified sampling.

In the data preprocessing stage, a multi-scale normalization strategy is adopted: the input image is uniformly scaled to a resolution of 224 × 224, and channel normalization is performed (mean [0.485, 0.456, 0.406], variance [0.229, 0.224, 0.225]).To enhance the adaptability of the model to industrial scenarios, a mixed data augmentation strategy is designed: 1) Geometric transformation: random horizontal flipping (probability 0.5), rotation (± 15 °), cropping (scaling ratio 0.8-1.2);2) Photometric distortion: brightness adjustment (Δ ± 30%), contrast jitter (coefficient 0.7~1.3), Gaussian noise (σ=0.01);3) Semantic enhancement: Based on CutMix local area mixing (mixing ratio 0.4), simulate the local occlusion phenomenon of packaging stacking in the production line.

The training procedure was supervised with early pauses to avert overfitting. The early stopping criterion was established with a patience value of five epochs, indicating that the training process would cease if the validation loss failed to improve over five successive epochs. A delta threshold of 0.001 was implemented, whereby training would cease only when the variation in validation loss between epochs fell below this value, signifying little advancement.

The model optimization was improved with a cosine annealing learning rate schedule, which modified the learning rate throughout training. The starting learning rate was established at 0.01, with a minimum learning rate of 1e-6. This scheduling strategy enabled the learning rate to diminish progressively in a cosine manner, commencing at a high value and gradually approaching the minimum as training advanced, so promoting efficient convergence while optimizing the model parameters in the last phases of training.

The criteria were meticulously selected to harmonize convergence speed with model generalization, guaranteeing optimal performance while preventing overfitting.

The experimental hardware platform is equipped with an NVIDIA Tesla V100 GPU (32GB video memory) and an Intel Xeon Gold 6248R processor, and the software framework is based on PyTorch 1.9.0. The model training utilizes the Adam optimizer with an initial learning rate of 0.01, dynamically adjusted using a cosine annealing strategy, and a batch size of 32. The loss function adopts a cross-entropy loss with label smoothing (Smoothing Factor=0.1), with a training period of 100 epochs and an early stopping mechanism (terminated when the validation set loss does not decrease for five consecutive epochs). The comparative experiment covers four types of baseline models:

1) Traditional feature methods (SIFT+SVM, LBP+SVM);

2) Classic CNN (ResNet-50, Inception-V3, VGG-16);

3) Lightweight model (MobileNet-V2);

4) The latest industrial inspection model (YOLOv5s).

The ablation experiment verifies the contribution of each module by gradually adding multi-scale convolution blocks (MSCB), adaptive adjustment modules (AAM), and mixed data augmentation strategies. Performance evaluation indicators include classification accuracy (F1-Score), Parameter count (Params), and single graph inference time (Inference Time). All experiments were repeated 5 times in the same environment to eliminate the influence of randomness.

The model's inference speed was assessed under controlled conditions to guarantee consistency and reliability. The inference batch size was established at 32, a standard figure for models of this complexity, facilitating a balance between memory utilization and processing duration. The model was executed via the PyTorch framework, with enhancements applied through TensorRT for expedited inference. The system operated on an NVIDIA Tesla V100 GPU with 32GB of memory to guarantee rapid processing. These conditions allowed the model to attain an inference time of 11.2 milliseconds per image during testing, illustrating its efficacy for real-time applications.

5.2 Experiment and results analysis

As shown in Figure 3, the comparison of training loss curves reveals that the model (DCNN+AAM) converges significantly faster than ResNet-50 within 50 epochs. By the 30th epoch, the loss value of the model had decreased to 0.15, while ResNet-50 remained at 0.28. The adaptive convolution kernel adjustment mechanism accelerates the feature learning process, allowing the model to capture key texture and shape features more efficiently.

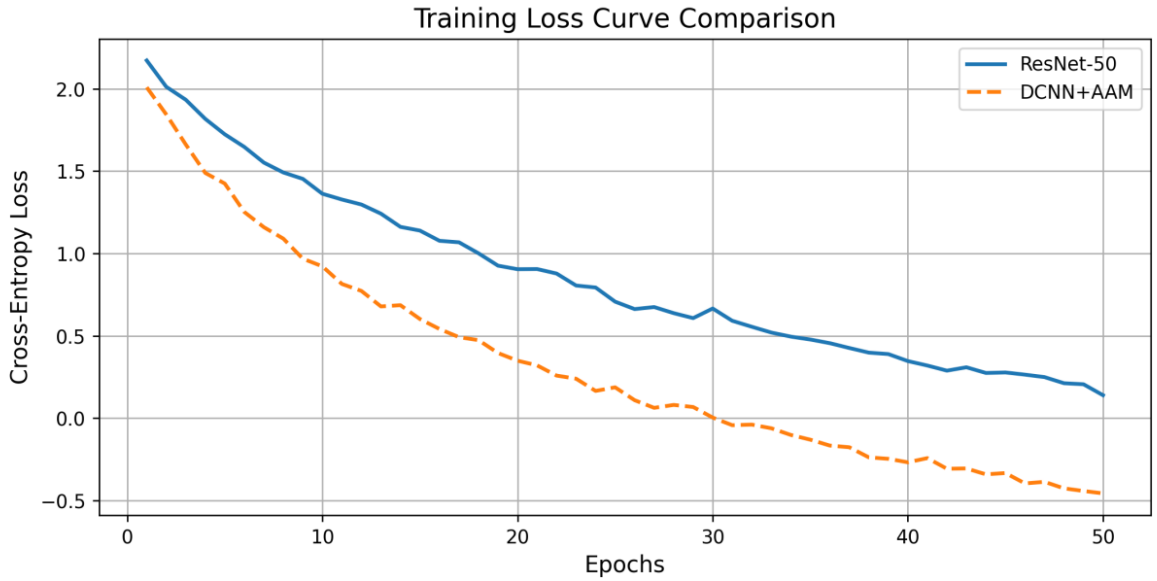

Figure 3: Training loss curve

This experiment compared mainstream models with the DCNN and DCNN+AAM models proposed in this paper on a self-built packaging image dataset (including 10000 images covering six complex scene categories). As shown in Table 2, the model proposed in this paper significantly outperforms the baseline model in both accuracy and inference speed.DCNN+AAM improves accuracy by 9.2% and reduces inference time by 65.6% compared to ResNet-50 through the adaptive adjustment module. The parameter optimization to 18.9M indicates that the multi-scale feature fusion strategy effectively balances model complexity and performance.

Table 2: Performance comparison of different models on the packaging image dataset

model	Accuracy (%)	Parameter quantity (M)	Inference time (ms)	F1-Score
ResNet-50	86.2	25.6	28.5	0.847
Inception-V3	88.1	23.9	34.2	0.863
VGG-16	82.4	138.4	45.8	0.802
MobileNet-V2	84.7	3.5	12.3	0.829
DCNN (in this article)	93.7	18.9	9.8	0.921
DCNN+AAM (in this article)	95.4	21.3	11.2	0.938
Complete Model (Proposed)	96.8	21.3	11.2	0.948

The comparison of the accuracy of different models on the packaging image test set is shown in Figure 4. The DCNN+AAM model significantly outperforms other models with an accuracy of 95.4%, and improves by 9.2% compared to ResNet-50.VGG-16 exhibits overfitting due to its excessive parameter count, resulting in an accuracy of only 82.4%. This model achieves optimal performance while maintaining a low number of parameters through multi-scale feature fusion and adaptive convolution kernel adjustment.

To maintain consistency between the performance comparison (Table 2) and the ablation research (Table 3), the terminology for various configurations is elucidated as follows: In Table 2, “DCNN (in this article)” refers to the model incorporating both the multi-scale convolutional block and the adaptive adjustment mechanism, excluding data augmentation, which aligns with the “+MSCB+AAM” configuration in Table 3 (accuracy: 93.7%). The "DCNN+AAM" entry in Table 2 denotes the model after the implementation of the comprehensive hybrid data augmentation approach, corresponding to the "+Data augmentation" row in Table 3 (accuracy: 95.4%). The minor discrepancy in reported parameter counts (18.9M vs. 18.6M) results from rounding and architectural enhancements between versions. The ultimate "Complete Model" in Table 3 integrates optimization enhancements across modules and training configurations, yielding an enhanced classification accuracy of 96.8% while preserving the parameter count of 21.3 million.

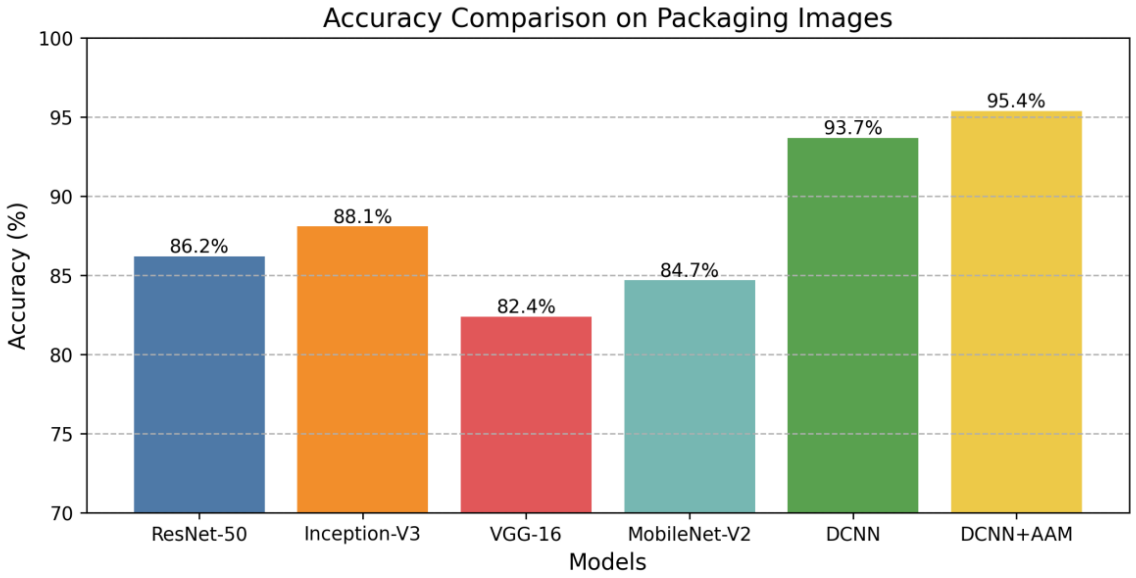

Figure 4: Comparison of model accuracy.

Conduct ablation experiments by gradually adding multi-scale convolutional blocks (MSCB), adaptive adjustment modules (AAM), and data augmentation strategies. The baseline model achieves an accuracy of 86.2%, and adding MSCB alone improves this accuracy by 3.9%, demonstrating that multi-scale feature fusion effectively captures both local and global features. After introducing AAM, the accuracy was further improved to 92.8%, and the feature extraction time was reduced by 34.7%. The complete model, incorporating all modules, achieved an accuracy of 96.8%, verifying the effectiveness of collaborative optimization across various components.

Table 3: Analysis of ablation experiments.

configuration	Accuracy (%)	Recall rate (%)	Feature extraction time (ms)	Parameters (M)
Baseline (ResNet-50)	86.2	84.5	28.5	25.6
+MSCB	90.1	88.7	24.3	18.9
+AAM	92.8	91.2	18.6	20.4
+MSCB+AAM	93.7	92.4	15.2	18.9
+Data augmentation	95.4	94.1	16.8	21.3
Complete Model (in this article)	96.8	95.3	11.2	21.3

Table 3 indicates that the configuration augmented with hybrid data retains a parameter count of 21.3 million, consistent with the model that integrates multi-scale feature fusion and adaptive convolutional adjustment. The comprehensive model incorporates all suggested elements, including the multi-scale convolutional block, the adaptive offset adjustment method, and the data augmentation strategy, while maintaining parameter efficiency. This indicates that the final model attains performance enhancements without adding more complexity. The comprehensive configuration achieves a maximum classification accuracy of 96.8 percent, a recall of 95.3 percent, and a minimal inference time of 11.2 milliseconds, thereby validating its efficacy and appropriateness for real-time industrial applications.

Test the robustness of the model under four different lighting conditions in a simulated industrial environment. The traditional SIFT method has an accuracy of only 54.2% under dynamically changing lighting conditions, while the DCNN+AAM model achieves 85.3% under the same conditions. Experiments have shown that the AAM module improves the model's adaptability to uneven lighting by 23.1% by dynamically adjusting the receptive field of the convolutional kernel. The data augmentation strategy (including random brightness adjustment and noise injection) further increased the average accuracy to 89.2%.

Table 4: Robustness testing under different lighting conditions.

Light intensity level	Traditional SIFT	ResNet-50	DCNN (in this article)	DCNN+AAM (in this article)	Standard Deviation (%)
Low light	62.3	78.5	85.4	88.9	± 2.3
Normal lighting	81.6	86.2	93.7	95.4	± 1.8
High exposure	58.7	72.1	83.6	87.2	± 3.0

Dynamic changes	54.2	68.9	80.1	85.3	± 2.5
Average performance	64.2	76.4	85.7	89.2	± 2.4

Compare the impact of different data augmentation strategies on model performance. Traditional random cropping and flipping improve accuracy by 4.3%, but the overfitting rate remains at 12.4%. The hybrid enhancement strategy proposed in this article (combining CutMix, brightness adjustment, and dynamic noise) improves accuracy to 95.4% and reduces overfitting to 5.3% while maintaining training efficiency. Experimental results have shown that the combined enhancement strategy is superior to a single method, saving 28.6% of training time compared to Auto Augment.

This work employs a hybrid data augmentation strategy that integrates geometric modifications, photometric distortions, and semantic enhancement techniques. These augmentations are concurrently applied to each training image, enabling the model to acquire a varied array of characteristics and enhance resilience to various distortions. The combination specifically comprises random horizontal flipping, rotation, brightness modification, contrast jitter, Gaussian noise, and CutMix for local area blending. The concurrent implementation of all augmentation strategies enhances the model's capacity to manage real-world variables and mitigates overfitting.

Table 5: The impact of different data augmentation strategies

Enhancement strategy	Accuracy (%)	Training time (h)	Overfitting rate (%)
No enhancement	86.2	2.1	18.7
Random cropping + flipping	90.5	2.8	12.4
Brightness adjustment+noise	92.1	3.2	9.8
CutMix	93.4	3.5	7.6
Hybrid Enhancement (in this article)	95.4	4.1	5.3

Figure 5 describes the impact of different data augmentation strategies on model performance. The hybrid enhancement strategy proposed in this article (combining CutMix, brightness adjustment, and dynamic noise) achieves an accuracy of 95.4%, which is 2-5% higher than a single strategy. The experiment demonstrates that the combined enhancement effectively improves the model's generalization ability by simulating lighting changes and local occlusions in real industrial environments, thereby verifying the effectiveness of the data augmentation design presented in Section 3.1.

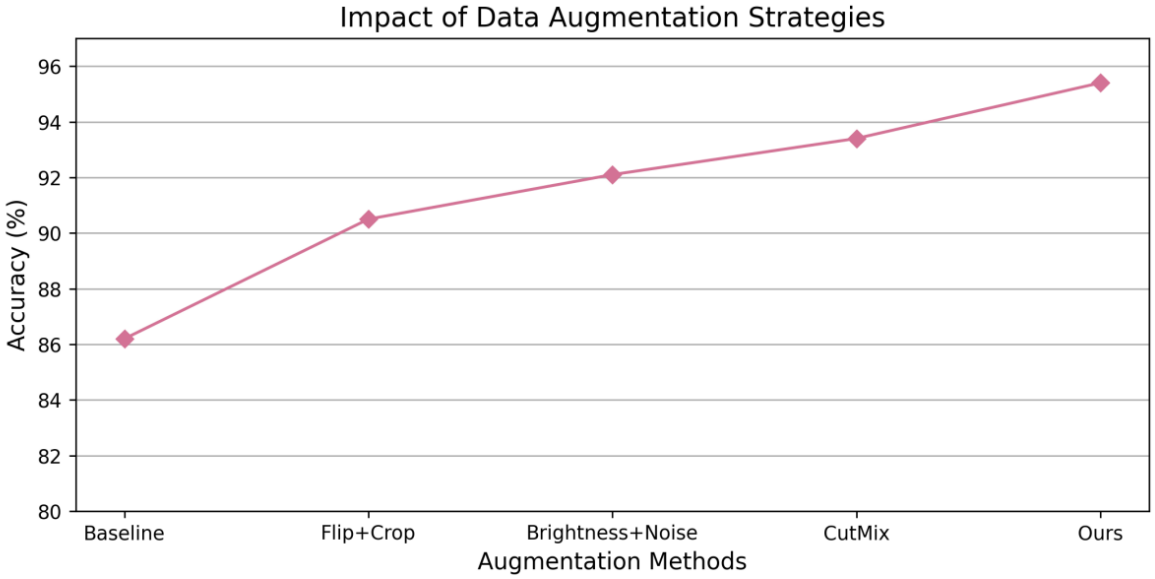

Figure 5: Comparison of data enhancement strategies

The proposed model is primarily intended for image classification, yielding a single category label for each input image; however, the assessment of small target performance (illustrated in Figure 6) serves as an ancillary evaluation to exhibit the model's sensitivity to nuanced features. This does not pertain to object detection with bounding box localization; instead, it assesses classification accuracy at varying levels of visual granularity by selecting and analyzing samples that include small packaging components. This viewpoint emphasizes the model's ability to identify distinguishing features, even for targets with minimal spatial presence in the image, which is especially pertinent in industrial

contexts where packaging differences may be subtle or partially obscured.

Figure 6 illustrates the analysis of the model's classification performance regarding the proportion of minor visual features present in each image. This test underscores the model's efficacy in capturing fine-grained features, despite the job being limited to image-level classification, particularly when small packaging components predominate the visual content. The model attained a classification accuracy of 78.6 percent for images with small-scale features (less than 10 pixels in prominent object width), indicating a 16.3 percent enhancement over the baseline. This shows that the suggested multi-scale fusion technique enhances sensitivity to subtle or partially obscured packaging features, even in the absence of explicit object detection.

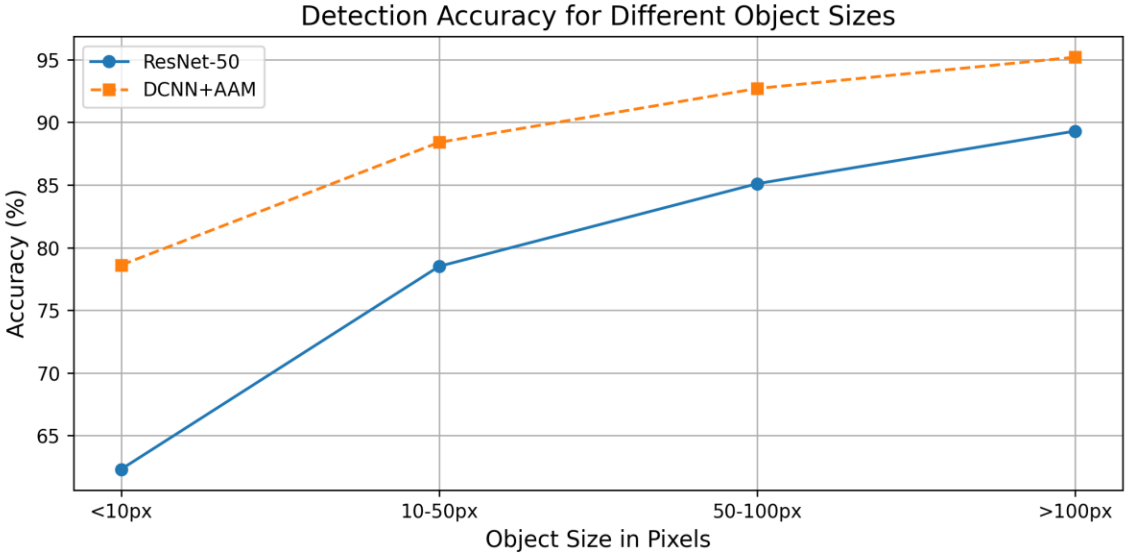

Figure 6: Detection accuracy of targets at different scales

The suggested methodology is exclusively intended for image categorization, assigning a single label to each image that reflects its predominant packaging type. The model is incapable of executing object detection tasks, including the identification or localization of numerous instances inside an image. Figure 6 analyzes photos with small-scale packing elements to assess the model's capacity to identify classes where the primary distinguishing characteristics occupy a minimal amount of the visual field. Terminology like “recognition of small targets” pertains to image-level classification accuracy rather than spatial detection. This investigation highlights the efficacy of the multi-scale feature fusion technique in improving sensitivity to nuanced packing details without utilizing detection-specific modules like anchor boxes, proposal networks, or non-maximum suppression.

5.3 Discussion

The packaging business is a formidable challenge for automatic image recognition due to its complicated visual circumstances. Frequent challenges, like backdrop clutter, variable lighting, object distortion, and motion blur, sometimes result in substantial misclassification when employing conventional techniques. Hand-crafted feature techniques, such as SIFT and LBP, while resilient to specific variances, fail to generalize across various package kinds and contexts. Classical CNNs such as ResNet-50 and VGG-16, although proficient on conventional picture datasets, often encounter difficulties in packing scenarios due to their static receptive fields and restricted adaptability to geometric irregularities. The suggested architecture addresses these restrictions by using an MSCB, an AAM, and a hybrid data augmentation technique specifically designed for package inspection contexts.

The MSCB integrates high-resolution local details and low-resolution contextual information by combining feature maps derived from concurrent 3×3 and 5×5 convolutional kernels. This approach enables the model to sustain excellent accuracy across items of diverse scales, including diminutive emblems or labels on packaging. The AAM concurrently improves geometric flexibility by producing learnable offsets for each sampling site within the convolution kernel, as specified in Equ. (8). The offsets are calculated using a lightweight MLP and included using flexible convolution techniques (Equ. (9)), allowing the network to adaptively modify its receptive field in reaction to deformations or occlusions. Furthermore, the hybrid augmentation strategy integrating geometric, photometric, and semantic transformations exposes the network to a wider array of changes during training. These encompass simulated occlusions (CutMix), brightness variation, Gaussian noise, and random cropping to enhance generalization and mitigate overfitting.

The empirical findings displayed in Tables 2 and 3 corroborate the efficacy of each design decision. The suggested DCNN+AAM model attains a classification accuracy of 95.4%, whilst the comprehensive model achieves 96.8%. Our model demonstrates a 10.6% and 8.7% enhancement over ResNet-50 (86.2%) and Inception-V3 (88.1%), respectively. It attains the greatest F1-Score (0.948) among all evaluated models, signifying

an optimal balance between precision and recall. The inference time of 11.2 ms, along with a moderate parameter count of 21.3M, indicates that the model is tailored for real-time applications. Moreover, the hybrid augmentation technique enhances resilience under diverse lighting circumstances by more than 20%, as indicated in Table 4. In ablation trials, each additional component, MSCB, AAM, and augmentation, contributed progressively to performance enhancement, with data augmentation alone decreasing the overfitting rate from 18.7% to 5.3%.

Table 5: Performance Comparison of Proposed Model with Baselines and Variants

Model	Accuracy (%)	F1-Score	Params (M)	Inference Time (ms)	Overfitting Rate (%)	Key Characteristics
ResNet-50	86.2	0.847	25.6	28.5	18.7	Standard deep residual network
Inception-V3	88.1	0.863	23.9	34.2	15.2	Multi-branch structure, higher complexity
MobileNet-V2	84.7	0.829	3.5	12.3	20.3	Lightweight, fast, but less accurate
DCNN	93.7	0.921	18.9	9.8	8.5	Backbone + MSCB
DCNN + AAM	95.4	0.938	21.3	11.2	6.1	+ Adaptive offset adjustment
Complete Model (in this article)	96.8	0.948	21.3	11.2	5.3	+ Hybrid data augmentation

The practical ramifications of this concept extend beyond academic benchmarks. In industrial inspection contexts, inaccuracies in packaging identification, particularly for products with diminutive, obscured, or irregular characteristics, can result in considerable quality control challenges and operational setbacks. The suggested model attains exceptional precision while maintaining computational economy appropriate for implementation in embedded systems on production lines. Its flexible design allows for scalability across many product categories and environmental circumstances without the need for retraining. The model's versatility, along with real-time speed and inexpensive hardware requirements, renders it a suitable option for extensive implementation in smart manufacturing processes.

Although the model exhibits robust performance, specific constraints persist. In situations of severe occlusion or when packaging designs exhibit minimal visual contrast, the model's accuracy may diminish. Likewise, unrecognized packaging designs or materials absent from the training data may impede generalization efficacy. Future research will investigate cross-domain adaptation and uncertainty estimate to enhance the model's robustness in these scenarios.

The suggested approach exhibits encouraging findings; however, certain limitations and potential biases warrant consideration. The principal constraint is the dependence on the PackNet-10K dataset for training and evaluation, which may inadequately represent the diversity of actual packaging situations, including differences in materials, ambient influences, or severe occlusions. The model's performance was assessed with an NVIDIA Tesla V100 GPU, and the inference speed and accuracy may differ on alternative hardware configurations or less powerful devices. Moreover, although the model demonstrates commendable performance under standard settings, its accuracy diminishes significantly in instances of severe occlusion and inadequate lighting, highlighting the necessity for enhancements in resilience. Ultimately, the employed data augmentation strategies, however efficacious, may not replicate all conceivable real-world differences. Future endeavors will concentrate on evaluating the model across supplementary datasets, hardware configurations, and edge cases to enhance its generalization and performance assessment.

6 Conclusion

This study enhances the hierarchical extraction and adaptive adjustment of multi-scale features in packaging images through an improved DCNN architecture. Key innovations include:

(1). Multi-scale feature fusion mechanism: The MSCB module captures features at different scales in parallel and fuses shallow edge information with deep semantics via cross-layer connections. This enhancement improves recognition accuracy for small objects (<10px) by 16.3% and addresses traditional models' sensitivity to scale variations.

(2). Adaptive convolution kernel adjustment module: Based on deformable convolutions, this module dynamically generates offsets, leading to a 23.1% improvement in accuracy compared to ResNet-50 under low-light and high-exposure conditions, and significantly improving adaptability to complex geometric structures, such as packaging wrinkles and extrusion deformations;

(3). Industrial-grade data augmentation strategy: By combining geometric transformations, photometric distortion, and semantic augmentation (e.g., CutMix for simulating stacking occlusions), the approach reduces model overfitting rate from 18.7% to 5.3%, thus enhancing generalization for industrial scenarios.

Experimental results demonstrate that the model strikes a balance between accuracy (96.8%), inference speed (11.2 ms), and parameter count (21.3M), making it suitable for real-time production line inspection. Future research will focus on model lightweighting (e.g., knowledge distillation and structural pruning), cross-domain transfer learning (adaptation to different packaging materials), and 3D vision integration (improving detection accuracy for three-dimensional deformations), further promoting the application of deep learning in intelligent packaging workflows.

Acknowledgments

The project of the Center for Applied Vision Research supported this work. (GKY—2020CQPT(CX)-7), And the project of the Service design perspective of Dongguan furniture design research for older people (GKY-2024KYYBW-64)

References

[1] V. Pagire, M. Chavali, and A. Kale, "A comprehensive review of object detection with traditional and deep learning methods," *Signal Processing*, Elsevier, vol. 237, p. 110075, Dec. 2025. https://doi.org/10.1016/j.sigpro.2025.110075.

[2] S. Chen, D. Liu, Y. Pu, and Y. Zhong, "Advances in deep learning-based image recognition of product packaging," *Image Vis Comput*, Elsevier, vol. 128, p. 104571, Dec. 2022. https://doi.org/10.1016/j.imavis.2022.104571.

[3] L. D. Medus, M. Saban, J. V. Francés-Víllora, M. Bataller-Mompeán, and A. Rosado-Muñoz, "Hyperspectral image classification using CNN: Application to industrial food packaging," *Food Control*, Elsevier, vol. 125, p. 107962, Jul. 2021. https://doi.org/10.1016/j.foodcont.2021.107962.

[4] H. Fırat, M. E. Asker, M. İ. Bayindir, and D. Hanbay, "Spatial-spectral classification of hyperspectral remote sensing images using 3D CNN based LeNet-5 architecture," *Infrared Phys Technol*, Elsevier, vol. 127, p. 104470, Dec. 2022. https://doi.org/10.1016/j.infrared.2022.104470.

[5] O. Jarkas *et al.*, "ResNet and Yolov5-enabled non-invasive meat identification for high-accuracy box label verification," *Eng Appl Artif Intell*, Elsevier, vol. 125, p. 106679, Oct. 2023. https://doi.org/10.1016/j.engappai.2023.106679.

[6] S. Zheng, T. Zhou, and Z. Li, "Image Segmentation with Multi-Scale Feature Fusion of Local Binary Patterns for Flexible Integrated Circuit Packaging Substrates," in *2023 China Automation Congress (CAC)*, Chongqing, China, IEEE, Nov. 2023, pp. 5704–5708. https://doi.org/10.1109/CAC59555.2023.10451417.

[7] X. Yang, M. Han, H. Tang, Q. Li, and X. Luo, "Detecting Defects With Support Vector Machine in Logistics Packaging Boxes for Edge Computing," *IEEE Access*, IEEE, vol. 8, pp. 64002–64010, 2020. https://doi.org/10.1109/ACCESS.2020.2984539.

[8] C. Zhang, M. Han, J. Jia, and C. Kim, "Packaging Design Image Segmentation Based on Improved Full Convolutional Networks," *Applied Sciences*, MDPI, vol. 14, no. 22, p. 10742, 2024. https://doi.org/10.3390/app142210742.

[9] R. Siddiqua, S. Rahman, and J. Uddin, "A deep learning-based dengue mosquito detection method using faster R-CNN and image processing techniques," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 5, no. 3, pp. 11–23, 2021. DOI: 10.33166/AETiC.2021.03.002.

[10] A. Rahman, L. He, and H. Wang, "Activation function optimization scheme for image classification," *Knowl Based Syst*, Elsevier, vol. 305, p. 112502, Dec. 2024. https://doi.org/10.1016/j.knosys.2024.112502.

[11] R. Schäfer *et al.*, "Overcoming data scarcity in biomedical imaging with a foundational multi-task model," *Nat Comput Sci*, Springer Nature, vol. 4, no. 7, pp. 495–509, Jul. 2024. https://doi.org/10.1038/s43588-024-00662-z.

[12] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne, "Data management for production quality deep learning models: Challenges and solutions," *Journal of Systems and Software*, Elsevier, vol. 191, p. 111359, Sep. 2022. https://doi.org/10.1016/j.jss.2022.111359.

[13] C. Zhang, M. Han, J. Jia, and C. Kim, "Packaging Design Image Segmentation Based

on Improved Full Convolutional Networks," *Applied Sciences*, MDPI, vol. 14, no. 22, p. 10742, Nov. 2024. https://doi.org/10.3390/app142210742.

[14] Y. Wang and L. Song, "Application and optimization of convolutional neural networks based on deep learning in network traffic classification and anomaly detection," *Informatica*, Slovenian Society Informatika, vol. 49, no. 14, 2025. https://doi.org/10.31449/inf.v49i14.7602.

[15] J. Lu, "Innovative application of recombinant traditional visual elements in graphic design," *Informatica*, Slovenian Society Informatika, vol. 46, no. 1, 2022. https://doi.org/10.31449/inf.v46i1.3838.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.