TPMN: Texture Prior-Aware Multi-Level Feature Fusion Network for Corrugated Cardboard Parcels Defect Detection

Xing He[1*], Haoxiang Fan[2], Cuifeng Du[3], Xingyu Zhu[4✉], Yuyu Zhou[5], Renzhang Chen[6✉], Zhefu Li[7], Guihua Zheng[8*], Yuansheng Zhong[9], Changjiang Liu[10], Jiandan Yang[11*], Quanlong Guan[12]
College of Information Science and Technology, Jinan University, China[1,12]
Sun Yat-sen University, Guangdong, China[2]
Cete Potevio Science Technology Co. Ltd, Guangdong, China[3]
Office of Scientific R&D, Guangdong, Jinan University[4]
Guangdong Institute of Smart Education, Jinan University, Guangdong, China[5,12]
Guangdong-Macao Advanced Intelligent Computing Joint Laboratory, Guangdong, China[5,12]
Modern Educational Technology Center of Zhuhai Campus, Jinan University, Guangdong, China.[6,8,11]
Network and Education Technology Center, Jinan University, Guangdong, China[7]
Key Laboratory of Safety of Intelligent Robots for State Market Regulation, Guangdong Testing Institute of Product Quality Supervision, Guangdong, China[9,10]
Guangdong Key Laboratory of Data Security and Privacy Preserving, Guangdong, China[12]

***Abstract*—Surface defect detection is the task of identifying and localizing defects on the surface of an object, which is a widely applied task in various industries. In the logistics industry, logistics companies need to monitor the condition of goods for potential defects throughout the entire logistics process for effective logistics quality control. However, effective defect detection methods are still lacking for courier packages using corrugated cardboard boxes, which rely on judging whether deformation and leakage have occurred by examining areas on their surface with abundant texture. Specifically, the defect rate and supporting structure of the packages are influenced by temperature and humidity, and the openings and bends of defects are inconsistent. This results in defective packages having rich and non-uniform texture features. Moreover, convolutional neural networks struggle to effectively extract low-level semantic texture features of defects and perceive multi-level image features of packages. Considering the above challenges, we propose a novel texture prior-aware multi-level feature fusion network (TPMN). We first introduce prior knowledge and attention mechanisms to enable the neural network to focus on extracting low-level texture features from the image in the early stages. We also design a multi-level feature fusion method to integrate features from different levels, avoiding the gradual loss of low-level semantic information in CNN and enabling comprehensive perception of multi-level image features. To support further research, we contribute the cardboard-boxes-dataset, comprising 1210 images of packages. Experiments on this dataset showcase the superior performance of TPMN, even in few-shot learning scenarios, demonstrating its effectiveness in surface defect detection within the logistics and supply chain domains.**

***Keywords*—logistics; surface defect detection; multi-level feature fusion; prior attention; corrugated cardboard boxes**

*These authors contributed equally to this work.
✉Corresponding authors.

I. Introduction

Surface defect detection is a widely applied task in various industries, the goal of which is to identify and locate defects on the surface of objects. Nowadays, an increasing number of surface defect detection methods based on deep learning are being proposed. Lv et al. proposed a single shot multiBox detector-based end-to-end defect detection network for defects on metal surfaces [1]. Huang et al. proposed a method for defect detection in micro-nozzles using canny edge detection and evaluating the texture features of the regions [2]. Many mature methods have also been proposed for applications in other materials, such as steel strips [3], fabric [4], and solar cells [5].

Nevertheless, the logistics industry, which is rapidly developing alongside e-commerce, still lacks reliable methods for surface defect detection. Reliable courier packaging is crucial for logistics quality, especially for fragile items, and tracking courier parcel defect helps logistics companies determine responsibility and improve logistics quality control. However, the corrugated cardboard boxes used for courier packaging differ from other industrial materials as they have limited waterproof properties and compressive strength. It may suffer different degrees of defect in the logistics environment. Corrugated cardboard boxes are highly sensitive to atmospheric conditions, and the defect rate and structural support of the packages can be significantly affected by temperature and humidity [6], [7]. Meanwhile, under the pressure of other goods, corrugated cardboard boxes may also develop inconsistent sizes of openings [8] and bends [9], leading to damage to the cargo. This complex defect scenario renders traditional surface defect detection methods unsuitable for courier parcels. Therefore, logistics companies urgently need a comprehensive and reliable defect detection method to achieve precise detection of

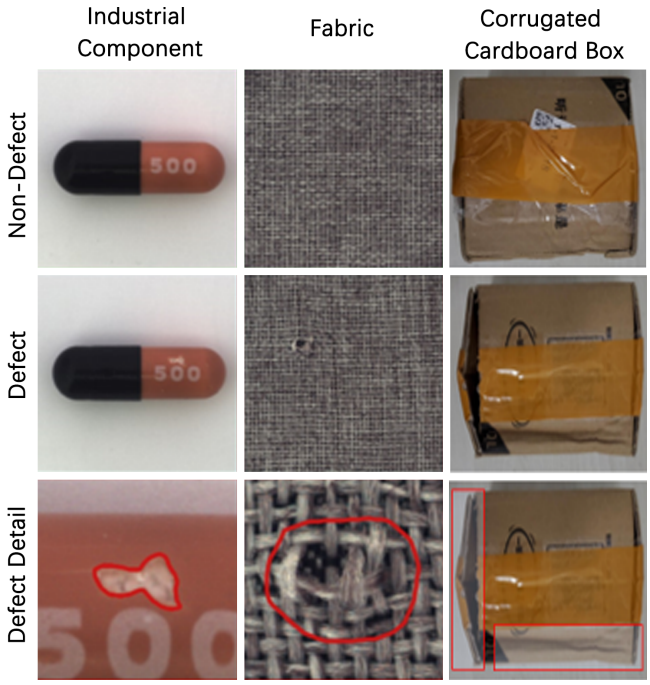


Fig. 1. Surface defects in different materials. In human judgment of whether a corrugated cardboard box has deformed or is leaking, it is typically reliant on image regions rich in textures, such as uneven package edges, folds extending around due to indentation, and edges at defect locations.

courier parcels, especially in cases with significant differences in defect sizes and overall structural. However, we still face the following challenges.

Firstly, the features of datasets for different defect detection tasks vary significantly, and the optimal solutions also differ. For corrugated cardboard boxes, humans judge whether deformation and leakage have occurred by examining areas on the surface with abundant texture (e.g., uneven package edges, depressed folds, defect edges). However, these texture-related low-level semantic features are often overlooked by Convolutional Neural Network (CNN) during the feature extraction process [10]. Additionally, low-level texture features suffer from semantic ambiguity due to their small receptive fields [11], [12]. Therefore, when analyzing the overall image and semantics of corrugated cardboard boxes, it is difficult to extract low-level texture image features.

Secondly, traditional methods often use the last convolutional feature map [13], resulting in insufficient semantic information and the loss of local information in the image. Specifically, courier parcels vary in size, and there is inconsistency in the texture sizes of corrugated cardboard boxes. When employing CNN with multiple convolution layers, local texture information may gradually be lost [14]. Moreover, in large-scale images, CNNs pay more attention to the high-level semantic information of the image [15], such as the overall structure and shape of corrugated cardboard boxes. Therefore, it is difficult to capture multi-level image features, which limits the task of surface defect detection on corrugated cardboard boxes.

In this paper, considering the above challenges, we propose a texture prior-aware multi-level feature fusion network (TPMN). Our method aims to accurately detect defect courier parcels, meeting the logistics company's need to track packaging defect status and providing crucial information for determine responsibility and improving logistics processes. Specifically, we first introduce prior knowledge and attention mechanisms to enable the neural network to focus on extracting low-level texture features from the image in the early stages. Then, we designed a multi-level feature fusion method to integrate features from different levels, avoiding the gradual loss of low-level semantic information in CNN and enabling comprehensive perception of multi-level image features. Additionally, we have contributed a dataset that comprises 1210 images of packages, known as the cardboard-boxes-dataset. On this dataset, we conducted basic experiments, ablation experiments, and few-shot learning experiments, among others. The experimental results demonstrate the superior performance of the TPMN.

To summarize, the primary contributions of this paper are as follows:

- We design the Texture Prior-Aware Multi-Level Feature Fusion Network, which integrates ResNet-18 [16] with multi-scale feature fusion and a prior attention mechanism. This framework enables precise defect classification and localization.
- The proposed TPMN is model-agnostic, allowing for effective extraction and fusion of low-level texture features while comprehensively perceiving multiscale image information.
- We released the Cardboard-Boxes-Dataset, which can be used for the task of detecting express packaging defects and promote further research in this field. The dataset is publicly available at https://github.com/chanllon/corrugated-cardboard-boxes-dataset.

II. Related Work

This section provides an overview of related work in three key fields: surface defect detection, prior attention, and multi-level feature fusion.

A. Surface Defect Detection

Surface defect detection is a widely applied task in various industries, with the main goal of identifying and locating defects or flaws on the surface of objects. Before the development of deep learning, surface defect detection primarily utilized traditional image processing techniques to extract features, employing machine learning for classification. Sun et al. utilized learning vector quantization networks and backpropagation networks for classification after segmenting the images [17]. Borwankar et al. introduced an k-nearest neighbors based algorithm for cast iron rocker arm inspection using frequency domain image processing [18]. However, these methods fall short in achieving superior detection accuracy. With the development of deep learning (DL), there are also many DL-based methods utilized for surface defect detection in the industry. Schlüter et al. introduce a simple and intuitive self-supervised method for sub-image anomaly detection and localization [19]. Lv et al. proposed a single shot multiBox detector-based end-to-end defect detection network for defects on the metal surface [1]. Huang et al. proposed a method using canny edge detection and evaluating region texture features for the defect detection of micro-nozzles [2]. Fang et al. designed

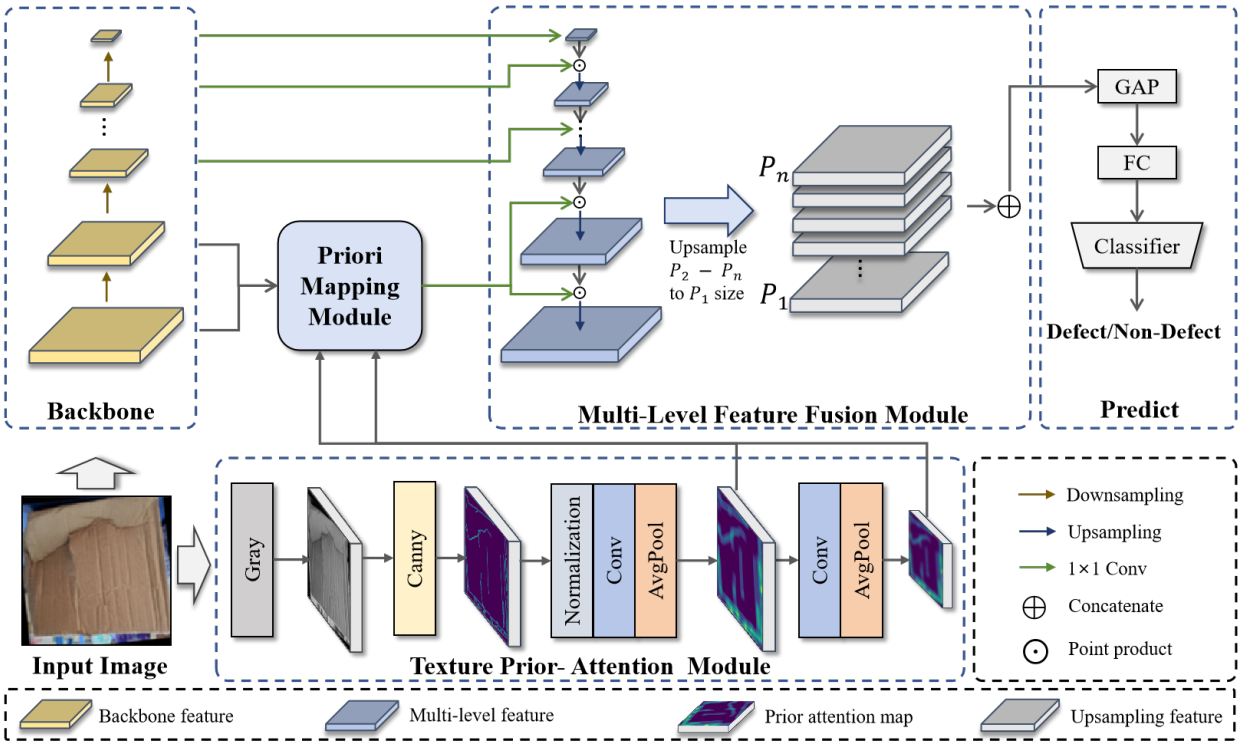


Fig. 2. Texture Prior-Aware Multi-Level feature fusion network.

a convolutional neural network (CNN) integrated with an attention mechanism to enhance training stability and detection accuracy in tactile methods for fabric structural defect detection [20]. However, the methods designed based on the specified material features mentioned above are not applicable to the complex defect scenarios in corrugated cardboard box defect detection [12], [21], [22]. Specifically, as illustrated in Fig. 1, the datasets for various defect detection tasks exhibit significant differences in features. Given the unique features of express packaging materials, we have devised an innovative model for surface defect detection.

B. Prior Attention

Prior attention enables the model to focus on important regions in the image, thereby improving the accuracy of object detection or image classification. Specifically, attention mechanisms allocate different weights to different parts of the input, allowing the neural network to concentrate on specific regions. SENet introduces a structure called the "Squeeze-and-Excitation" block, enabling the model to adaptively learn relationships between input feature channels [23]. DANet incorporates parallel global and local attention modules, focusing on global context and local details, respectively [24]. The role of prior knowledge in attention mechanisms is to introduce previous experience or assumptions to guide the neural network in focusing on specific information during the learning process. Cai et al. introduced image noise and edges as prior knowledge into the neural network, significantly enhancing the detection performance [25]. Wang et al. generate prior attention maps through a binary classifier to enhance lesion detection in COVID-19 CT screenings [26]. Zhang et al. assigns different weights to positions based on the prior that objects are near the image center and perceives object context information through different receptive fields [26]. However, existing prior attention methods are not applicable to the detection of defects in corrugated cardboard boxes, which focus more on regions rich in surface textures.

C. Multi-Level Feature Fusion

The task of multi-level feature fusion aims to effectively integrate feature information from different levels to enhance the performance of deep learning models when handling multi-level input data. HyperNet achieves effective multi-level feature fusion by aggregating hierarchical features and compressing them into a uniform space, enabling superior object detection performance across various levels [13]. Single shot multiBox detector achieves multi-level feature fusion by predicting category scores and box offsets for default bounding boxes using small convolutional filters [27]. Feature pyramid networks utilize a top-down architecture with lateral connections to facilitate effective multi-level object detection by integrating contextual information [28]. This method also finds extensive applications in other fields, such as remote sensing images [29], [30], classification of agricultural pests [31], and medical applications [32], [33], and so on. However, the above methods do not fully consider the texture information of low-level images, making it difficult to effectively integrate texture features on corrugated cardboard boxes at different levels and overall structures.

III. Texture Prior-Aware Multi-Level Feature Fusion Network

The overall architecture of the texture prior-aware multi-level feature fusion network is depicted in Fig. 2. Our network mainly consists of four parts: backbone, texture prior attention module, priori mapping module, and multi-level feature fusion module. We employ data augmentation techniques including mirroring, scaling, rotation, and translation to boost the diversity and complexity of the samples in light of the small number of samples in the dataset. This helps reduce the overfitting problem of the model. Enhanced images are created by randomly augmenting the original images, which are then sent into the backbone and texture prior attention module. Afterwards, we introduce each module of the network separately.

A. Backbone

ResNet-18 [16] provides strong feature learning capabilities for image features at various levels and abstraction levels. It is critical for detecting package defects, which typically manifest as local detail changes in the image, and ResNet-18 is capable of capturing these subtle features. Therefore, we designed the backbone based on ResNet-18. The ResNet-18 was constructed from residual blocks. ResNet-18 has N residual blocks, with each residual block's input set to x_i. The first block's input is an enhanced image, and the inputs of subsequent blocks are drawn from the previous block's output. The calculation for each residual block in ResNet-18 is as follows.

$$F_{RB}^{1} = ReLU(Conv(BN(Conv(x)))) \tag{1}$$

$$F_{RB}^{i} = ReLU(Conv(BN(Conv(F_{RB}^{i-1})))), i = \{2, \ldots, n\} \tag{2}$$

where F_{RB}^{i} represents the output of the i-th residual block. $Conv$ stands for convolution, BN for batch normalization, and $ReLU$ for rectified linear unit.

B. Texture Prior Attention Module

Humans frequently rely on "textured" parts, such as zigzag edges and depressed creases, to determine whether corrugated boxes are distorted or leaking. Moreover, the value of each pixel in an RGB image is determined by the richness of the surrounding texture. We propose a canny-based prior attention method for texture recognition that extracts wrapped texture features as priori knowledge, allowing the model to pay more attention to essential texture areas. Experiments have shown that the prior attention map improves model performance significantly.

We first convert the image to grayscale before using the canny algorithm to extract the edges. The texture feature map is then min-max normalized to ensure that the value of each pixel is between 0 and 1, in order to obtain the prior knowledge map. The calculation process is as follows:

$$F_C = Canny(Gray(x), C_{lower}, C_{upper}) \tag{3}$$

$$T = \frac{F_C – min(E_C)}{max(F_C) – min(F_C)} \tag{4}$$

where, $Gray$ represents converting the input image x to a grayscale image, $Canny$ represents the Canny edge detection algorithm, and C_{lower} and C_{upper} indicate the lower and upper thresholds. T denotes the feature map after Max-Min normalization.

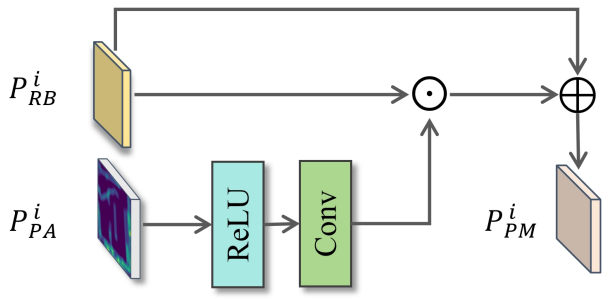


Fig. 3. The priori mapping module.

Included in the texture feature should be the surrounding texture-rich area as well as the texture itself. We extend the attention space even more by using downsampling and average pooling layers. Specifically, through the local perceptiveness of convolution and the information integration of average pooling, the expansion of texture edges can be achieved. This process allows the final Prior Attention Map to cover both the texture edges and the nearby texture information. The size of the output image after average pooling is the same as the size of the RGB branch feature map. Each feature downsampling module includes a 3×3 convolutional layer with 16 channels, followed by a 7×7 average pooling layer. The preceding method is as follows:

$$F_{PA}^{1} = AvgPool(Conv(T)) \tag{5}$$

$$F_{PA}^{2} = AvgPool(Conv(F_{PA}^{1})) \tag{6}$$

where F_{PA}^{i} represents the prior attention map at the i-th layer. $AvgPool$ is an acronym for average pooling.

C. Priori Mapping Module

The primary function of the Prior Attention Mapping module is to map the prior attention map to the enhanced feature map of the image obtained from the backbone. We consider fusing texture features and image features in the shallow layer of the network since downsampling the prior attention map leads to an erroneous attention range. The overview of this module is illustrated in Fig. 3.

This module gets the backbone feature F_{RB} of size $H \times W \times C$ from the encoder shallow layer and the texture prior attention module's prior attention map F_{PA} of size $H \times W \times 16$ as input and outputs the fusion feature F_{PM}. The process is as follows:

$$K = ReLU(Conv(F_{PA}^{i})), i = \{1, 2\} \tag{7}$$

$$F_{PM}^{i} = F_{RB}^{i} \odot K + F_{RB}^{i} \tag{8}$$

where, $\odot$ stands for dot-product. To prevent the above procedure from producing too small values and causing the gradient to vanish, we use the residual technique to let another F_{RB}^{i} skip the priori mapping module and add it to $F_{RB}^{i} \odot K$. This ensures that the performance of the network with the attention map will not be poorer than the original performance. Specifically, this

module is applied to the output of block1 and block2 of the encoder.

D. Multi-Level Feature Fusion Module

In order to better incorporate high-level semantic information from images and prevent low-level semantic information, including texture features, from vanishing during the training process. We design the multi-level feature fusion module based on the feature pyramid network mechanism. Consider that there are n blocks in the multi-level feature fusion module. Defined the input consists of the priori mapping map F_{PM}^1, F_{PM}^2 obtained in the priori mapping module and the backbone feature F_{RB}^i of various sizes obtained in the backbone. The output of each block as F_{MF}^i, where, $i \in \{1, .., n\}$. The module's ultimate output is the final fusion map P, which integrates feature maps from all levels. The final output of this module is the fusion of feature maps from all levels, denoted as P. Later, we will explain the specific details.

In order to preserve the rich texture information contained in low-level semantics, the lowest-level multi-level features need to be fused with the priori mapping map F_{PM}^i and the upper-level multi-level features F_{MF}^{i+1}. The process is as follows:

$$F_{MF}^1 = Conv(F_{PM}^1) \odot F_{MF}^2 \tag{9}$$

$$F_{MF}^2 = Conv(F_{PM}^2) \odot F_{MF}^3 \tag{10}$$

where $Conv$ stands for the convolution. $\odot$ stands for dot-product. F_{MF}^1 and F_{MF}^2 represent the priori mapping maps for the first and second layers, respectively. F_{MF}^i denotes the multi-level features for the i-th layer.

In multi-level features from the third layer and above, the model focuses more on the high-level semantic information of corrugated cardboard boxes. Therefore, the fusion of multi-level features from the third layer and above involves the backbone feature F_{RB}^i from the backbone and the upper-level multi-level features F_{MF}^{i+1}. The computation is as follows:

$$F_{MF}^i = Conv(F_{RB}^i) \odot F_{MF}^{i+1}, i = \{3, ..., n-1\} \tag{11}$$

$$F_{MF}^n = Conv(F_{RB}^n) \tag{12}$$

where F_{RB}^i represents the backbone feature for the i-th layer. For the multi-level feature F_{MF}^n in the nth layer, which doesn't have upper-level features, we only need to consider the nth layer's backbone feature F_{RB}^n.

In order to further fuse maps of different levels, we upsample each F_{MF}^i to the same size as F_{MF}^1 using different factors, denoted as p_i. After that, we process the features with the following:

$$P = Concate\{F_{MF}^1, p^2, \ldots, P^n\} \tag{13}$$

where $Concate$ represents connecting features by channel dimension.

E. Predict and Loss Function

In addition to the aforementioned modules, it is essential to incorporate an additional classifier that takes the final fusion map as input to detect whether the corrugated cardboard box has defects. Specifically, the features that the multi-level feature fusion module outputs are average pooled and input to a fully connected network classifier for classification. The training goal is to minimize the cross-entropy loss function, aiming to make the predicted probability $\hat{y}$ closely match the true label y:

$$L = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \tag{14}$$

where N denotes the total number of samples in the test set. y_i represents the ground truth label of the i-th sample, taking binary values (0 or 1). $\hat{y}_i$ represents the predicted probability of the i-th sample.

IV. Experiment

In this section, we first introduce the data we collected, known as the cardboard-boxes-dataset. Then, we detail a series of experiments we conducted to test on cardboard-boxes-dataset. Additionally, we analyze the critical roles of several key modules designed by us in neural network learning through visualization.

A. Dataset

We provide a detailed overview of the data collection and feature labeling processes, as well as the specific details of dataset split.

1) Data Collection and Feature Labeling: In cardboard-boxes-dataset, we collected a total of 1210 images of packages. Among them, 761 images are from packages that have undergone express delivery and were actually delivered. The remaining images were obtained by purchasing new corrugated cardboard boxes of different sizes and manually simulating various types of defects that might occur to packages before capturing images.

We used LabelMe* to annotate detailed auxiliary information for 990 images, which can be used for tasks such as package localization and defect detection.

The annotation process involved two experts and was conducted in two rounds due to slight differences in their psychological expectations regarding whether the packages were defective. In the first round, each expert independently labeled each package image without any communication. Images that both experts found ambiguous or impossible to categorize were eliminated (225 images). In the second round, the experts continued to label the images without communication. The consistency of the labeling results was assessed using Cohen's Kappa:

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{15}$$

where p_0 is the actual probability of agreement between the two experts, and p_e is the probability of agreement due to chance.

*https://github.com/wkentaro/labelme

TABLE I. COMPARISON EXPERIMENTS OF ATTENTION NETWORKS

Times	CE				Acc			
	SENet	DANet	AttNet	Our Method	SENet	DANet	AttNet	Our Method
1	0.5202	0.516	0.5342	0.4625	0.8365	0.8428	0.8365	0.8805
2	0.5193	0.4779	0.5353	0.4753	0.8616	0.8679	0.8176	0.8742
3	0.5214	0.4925	0.5289	0.5114	0.8491	0.8742	0.8239	0.8553
4	0.5017	0.5098	0.5109	0.4856	0.8742	0.8553	0.8428	0.8742
5	0.4979	0.4726	0.5241	0.4932	0.8428	0.8679	0.8302	0.8742
6	0.5353	0.5096	0.5107	0.4986	0.8365	0.8491	0.8491	0.8491
7	0.4771	0.4985	0.5234	0.4953	0.8553	0.8742	0.8239	0.8679
8	0.5008	0.5322	0.5607	0.4770	0.8491	0.8302	0.8239	0.8679
9	0.5002	0.4878	0.5227	0.5323	0.8679	0.8428	0.805	0.8176
10	0.5364	0.4879	0.5342	0.4641	0.8553	0.8616	0.8113	0.8868
Average	0.5110	0.4984	0.5285	**0.4895**	0.8528	0.8566	0.8264	**0.8647**
Min CE/Max Acc	0.4771	0.4726	0.5107	**0.4625**	0.8742	0.8742	0.8491	**0.8868**
variance	0.0186	0.0184	**0.0143**	0.0215	**0.0126**	0.0150	0.0136	0.0199

TABLE II. DATASET SPLIT

	Positive Samples	Negative Samples	Total
Training Set	191	287	478
Validation Set	80	80	160
Test Set	79	80	159
Total	350	447	797

TABLE III. PREDICTION EXPERIMENTAL RESULTS OF TPMN AND MN-TPMN ON THE TEST SET

Times	CE		Acc	
	TPMN	MN-TPMN	TPMN	MN-TPMN
1	0.4625	0.4746	0.8805	0.8994
2	0.4753	0.5361	0.8742	0.8113
3	0.5114	0.4742	0.8553	0.8742
4	0.4856	0.5126	0.8742	0.8239
5	0.4932	0.4924	0.8742	0.8553
6	0.4986	0.5043	0.8491	0.8742
7	0.4953	0.4776	0.8679	0.8805
8	0.4770	0.4979	0.8679	0.8679
9	0.5323	0.5148	0.8176	0.8428
10	0.4641	0.5646	0.8868	0.7925
Average	**0.4895**	0.5049	**0.8647**	0.8522
Min CE/Max Acc	**0.4625**	0.4742	0.8868	**0.8994**
variance	**0.0215**	0.0288	**0.0199**	0.0339

The final result is 0.6179, indicating a high level of consistency ($\geq$ 0.61 and $<$ 0.8) in the labeling results between the two experts. This implies a highly unified standard regarding whether the packages are defective. After removing inconsistent images labeled by both experts, there are a total of 350 images of non-defective packages and 447 images of defective packages.

2) Dataset Split: The dataset is randomly split into training, validation, and test sets with a ratio of 6:2:2, ensuring a balanced distribution of positive and negative samples in the validation and test sets to avoid data imbalance during validation and testing. Details are shown in Table II.

B. Hyperparameter Setting

TPMN is implemented by TensorFlow 2.8.The GPU used for training is the NVIDIA GeForce RTX 3090 24G. The input size of the backbone is 224×224×3, and the input size of the texture prior attention module is 224×224×1. The canny's upper threshold is 140 and its lower threshold is 80.

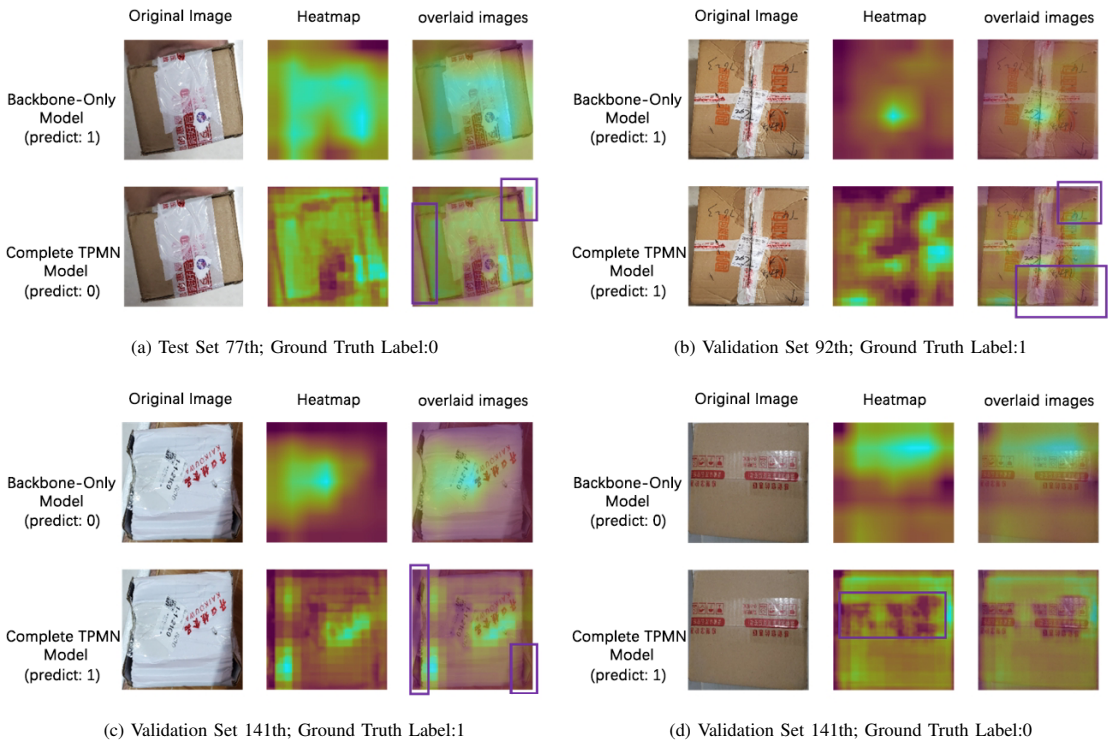
The chosen optimizer is Adam, with a learning rate of 0.001 and a decay of 0.002. The batch size is set to 128. Early stopping is implemented with a patience of 10, monitored by the cross entropy. The fully connected layers have 32 and 64 units, and the channels of multi-level feature fusion module are set to 32. Label smoothing is applied with a coefficient of 0.2.

C. Experimental Results

The model evaluation metrics include cross entropy (CE) and accuracy (Acc). Additionally, to better assess the model's performance, we will also separately consider metrics such as average cross entropy, average accuracy, minimum cross entropy, maximum accuracy, as well as the variance of cross entropy and accuracy. We additionally designed the Backbone of the TPMN based on MobileNet, known as MN-TPMN. The model was trained and tested a total of 10 times. All models were trained and tested in the cardboard-boxes dataset for a total of 10 times, and predicted whether the package was defective in the validation set.

The basic experimental prediction results, as shown in Table.III, demonstrate that the TPMN performs exceptionally well, exhibiting lower CE values and a higher average Acc. Compared to MN-TPMN, TPMN has a 1.5% lower average CE and a 1.2% higher average Acc. Each variance indicates that TPMN shows better stability.

We also compared our method with other attention mechanism networks: AttNet, SENet [16], and DANet [24]. The backbone network for these three models is ResNet-18, and the other training parameters are kept consistent with our training approach. AttNet is a spatial attention network that we re-implemented based on ResNet-18 [16]. The results, as shown in Table I, demonstrate that TPMN achieves optimal performance in terms of both CE and Acc compared to other



(a) Test Set 77th; Ground Truth Label:0

(b) Validation Set 92th; Ground Truth Label:1

(c) Validation Set 141th; Ground Truth Label:1

(d) Validation Set 141th; Ground Truth Label:0

Fig. 4. Visualization of the attention map of the convolutional layer. Each subfigure includes the original image in the first column, the heatmap of attention from the convolutional layer extracted by GradCAM in the second column, and the overlay of the first two images in the third column.

TABLE IV. ABLATION STUDY

	Backbone	MFF	TPA	Average	Min CE Max Acc	variance
CE	✓			0.5220	0.4869	0.0226
	✓	✓		0.4929	0.4837	**0.0066**
	✓	✓	✓	**0.4895**	**0.4625**	0.0215
Acc	✓			0.8327	0.8553	0.0209
	✓	✓		0.8528	0.8679	**0.0122**
	✓	✓	✓	**0.8647**	**0.8868**	0.0199

attention mechanism networks. The higher variance of 1.2% is attributed to the complexity of texture patterns, leading to frequent changes in attention weights. In summary, TPMN exhibits substantial advantages in the logistics package defect detection task, proving the efficacy of our designed Texture Prior Attention Module.

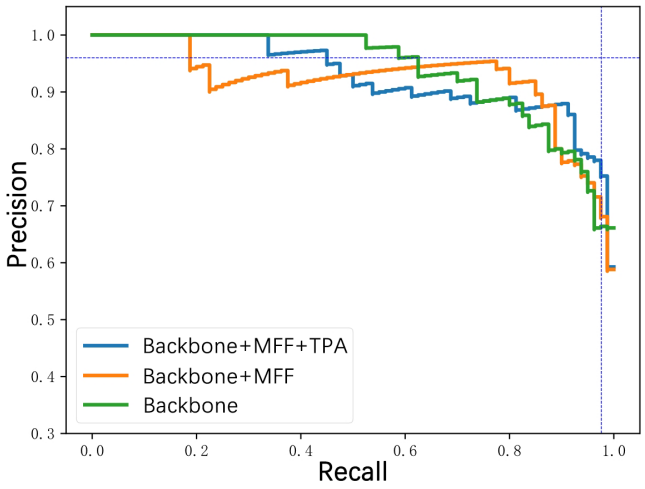


Fig. 5. PR curves from ablation experiments.

D. Ablation Study

To demonstrate the effectiveness of each module of TPMN, we conducted the following ablation studies: 1) The complete network, TPMN; 1) backbone whitout the texture prior attention module (TPA) from our proposed model; 2) backbone whitout the texture prior attention module(TPA) and the multi-level feature fusion module (MFF). As shown in Table.IV , compared to the backbone, we can observe that the utilization of MFF leds to an decrease in the average CE from 0.522 to 0.489, and a increase in the average ACC from 0.832 to 0.852. Considering the optimal values the model can achieve, the maximum accuracy increase by 1.26%. Furthermore, compared to the backbone, the combined use of TPA and MFF shows a significant decrease of 5% in the average CE, and decrease of 2.3% int the minimum CE. Meanwhile, Backbone-only model exhibits a pattern where its average ACC is 2.01% lower than backbone without TPA and 3.2% lower than the complete network. This indicates that, with the TPA and MFF, the model's average performance is much better than both backbone and backbone without TPA. Experiments prove the effectiveness of the two modules we designed.

The P-R curves for the three models on the validation set

TABLE V. FEW-SHOT LEARNING EXPERIMENT

	Model	Size of Training Set					
		All	200	100	50	20	10
CE	Backbone	0.4869	0.5837	0.6043	0.6627	0.8467	0.8332
	Backbone+MFF	0.4837	0.5694	0.5522	0.6180	**0.7193**	0.7607
	Backbone+MFF+TPA	**0.4625**	**0.4896**	**0.5363**	**0.5872**	0.752	**0.6612**
Acc	Backbone	0.8553	0.7610	0.7673	0.6918	0.6352	0.5723
	Backbone+MFF	0.8679	0.8302	0.7925	0.7421	**0.6667**	0.5912
	Backbone+MFF+TPA	**0.8868**	**0.8616**	**0.8365**	**0.7484**	**0.6667**	**0.7044**

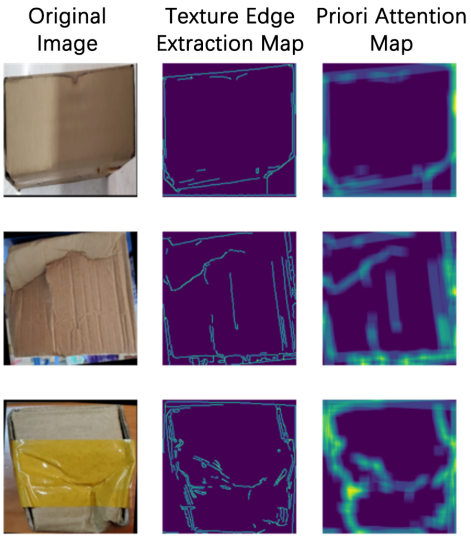


Fig. 6. Visualization of texture feature extraction.

are shown in Fig. 5, providing a preliminary insight into the workings of the two main modules. In situations where high precision is emphasized, the TPMN performs relatively worse compared to backbone with MFF, potentially misclassifying more intact packages as defective. Nevertheless, when Recall is greater than 0.85, TPMN outperforms the other control group, and backbone with MFF performs better than backbone-only model when Recall is greater than 0.7. This indicates that the TPMN is highly sensitive to defective packages. The texture prior attention module highlights the importance of texture, yielding better results for packages with rich textures. However, when the packages are non-defect, the model is less affected by effective textures, leading the attention towards patterns and text on the packages, which interferes with model training.

E. Visual Analysis

We first visualized the extraction of prior information from texture features. Then, we conducted a visual analysis of the impact of the prior attention map on the model training.

1) Visualization of Texture Extraction: We conducted an analysis of the effectiveness of the Texture Prior-Attention Module in extracting texture features. In Fig. 6, after Canny edge detection, the texture edges of the corrugated cardboard were successfully extracted, but the nearby texture information was not included. The final Prior Attention Map, obtained through convolution and average pooling, expands the attention range along the texture edges, thereby accommodating richer texture information. Through the local perceptiveness of convolution and the information integration of average pooling, the expansion of the attention range on the texture can be achieved.

2) Visual Analysis of Model Effectiveness: To provide additional evidence of the effectiveness of our approach, we conduct visual analysis on both the backbone-only model and the complete TPMN model. We analyze the model and classification results from the perspective of original images and convolutional layer attention. The attention map of the convolutional layer is obtained using the GradCAM [34] applied to the model's output and its last convolutional layer. Fig. 4a Fig. 4b and Fig. 4c demonstrate the advantages of the two main modules. TPMN can more accurately locate and focus on edge features, leading to accurate predictions. Without the two modules, backbone can only make predictions by broadly attending to various regions of the image, making it difficult to achieve the same performance. Fig. 4d, illustrates another extreme. When there are clear patterns or text on non-defect packages, the performance of the TPMN tends to decline. Due to the attention mechanism, the model's attention is forced to concentrate on information unrelated to the defect of package, resulting in prediction errors.

F. Few-Shot Learning Experiment

This section aims to investigate whether the model can maintain excellent performance with a reduced amount of data. Five sets of experiments were designed, each trained on different-sized training sets, and their final performance was measured. Despite variations in training set sizes, all datasets were processed using the data augmentation methods mentioned earlier.

The experimental results are shown in Table V. Regardless of the dataset size, TPMN has the highest Acc among all models, and CE also reaches the lowest in all five experiments. This indicates that the TPMN is able to maintain great performance even under conditions of limited data.

Attention mechanisms based models on often require larger datasets for training [35]. However, the TPMN excels on smaller datasets, because our designed prior attention and multi-level feature fusion methods effectively extract low-level texture features and Fusion multi-level features.

V. Conclusions and Future Work

In this study, we proposed a novel approach named the Texture Prior-Aware Multi-Level Feature Fusion Network to address the challenges in surface defect detection for corrugated cardboard boxes used in the logistics industry. Our method integrates a multi-level feature fusion technique that preserves and utilizes information from different semantic levels, overcoming the limitations of traditional Convolutional Neural Network (CNN)-based methods, which often suffer from the loss of local information and insufficient semantic details. The introduction of a prior attention mechanism enables the neural network to focus on extracting low-level texture features from the images in the early stages. The TPMN model, being model-agnostic, effectively extracts and fuses low-level texture features while comprehensively perceiving multiscale image information.

We conducted extensive experiments on our newly contributed Cardboard-Boxes-Dataset, which comprises 1210 images of packages. The results consistently demonstrated the superior performance of the TPMN model in precise defect classification and localization compared to traditional methods. The integration of ResNet-18, multi-scale feature fusion, and a prior attention mechanism proved effective in addressing the challenges unique to the logistics industry, where courier parcels, especially those made of corrugated cardboard, can exhibit varying defect sizes and structural complexities.

In the Future, there are several promising directions for further research and enhancement of TPMN. Firstly, expanding the dataset by incorporating diverse samples under varying environmental conditions would strengthen the model's robustness and generalization capabilities. Additionally, extending the application of the TPMN model to detect defects in different packaging materials commonly encountered in logistics, such as plastic or composite materials, could enhance its versatility. Exploring optimization strategies for real-time deployment, considering computational efficiency and resource constraints, is crucial for practical implementation in logistics settings. Integration with robotic systems for automated surface defect detection in logistics warehouses represents a potential avenue to improve efficiency and reduce manual intervention. In summary, the TPMN model lays a foundation for effective surface defect detection, and future research endeavors can capitalize on these insights to address evolving challenges in the dynamic logistics industry.

Acknowledgment

This work was supported by the Science and Technology Planning Project of Guangdong (2021B0101420003, 2023ZZ03, 2023A0505030013), the Science and Technology Planning Project of Guangzhou (202206030007, Nansha District: 2023ZD001), Guangdong Key Laboratory of Data Security and Privacy Preserving (2023B1212060036), Guangdong-Macao Advanced Intelligent Computing Joint Laboratory (2020B1212030003), the Opening Project of Key Laboratory of Safety of Intelligent Robots for State Market Regulation (GQI-KFKT202205).

References

[1] X. Lv, F. Duan, J.-j. Jiang, X. Fu, and L. Gan, "Deep metallic surface defect detection: The new benchmark and detection network," *Sensors*, vol. 20, no. 6, p. 1562, 2020.

[2] K.-Y. Huang and Y.-T. Ye, "A novel machine vision system for the inspection of micro-spray nozzle," *Sensors*, vol. 15, no. 7, pp. 15 326–15 338, 2015.

[3] Y. Song, Z. Liu, J. Wang, R. Tang, G. Duan, and J. Tan, "Multiscale adversarial and weighted gradient domain adaptive network for data scarcity surface defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–10, 2021.

[4] B. Su, H. Chen, P. Chen, G. Bian, K. Liu, and W. Liu, "Deep learning-based solar-cell manufacturing defect detection with complementary attention network," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4084–4095, 2020.

[5] M. Chen, L. Yu, C. Zhi, R. Sun, S. Zhu, Z. Gao, Z. Ke, M. Zhu, and Y. Zhang, "Improved faster r-cnn for fabric defect detection based on gabor filter with genetic algorithm optimization," *Computers in Industry*, vol. 134, p. 103551, 2022.

[6] S. Allaoui, Z. Aboura, and M. Benzeggagh, "Effects of the environmental conditions on the mechanical behaviour of the corrugated cardboard," *Composites Science and Technology*, vol. 69, no. 1, pp. 104–110, 2009.

[7] Z. Chen, C. Du, Y. Zhou, H. Guan, X. Huang, Z. Li, C. Liu, X. Zhuang, X. Zhu, and Q. Guan, "Ytcnet: A real-time algorithm for parcel damage detection with rich features and attention," in *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*.

[8] Garbowski , Tomasz and Gajewski, Tomasz and Grabski, Jakub Krzysztof, "Estimation of the compressive strength of corrugated cardboard boxes with various openings," *Energies*, vol. 14, no. 1, p. 155, 2020.

[9] Garbowski, Tomasz and Gajewski, Tomasz and Grabski, Jakub Krzysztof, "The role of buckling in the estimation of compressive strength of corrugated cardboard boxes," *Materials*, vol. 13, no. 20, p. 4578, 2020.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[11] W. Yu, X. Sun, K. Yang, Y. Rui, and H. Yao, "Hierarchical semantic image matching using cnn feature pyramid," *Computer Vision and Image Understanding*, vol. 169, pp. 40–51, 2018.

[12] Z. Chen, C. Du, Q. Guan, Y. Zhou, V. Hoo, X. Huang, Z. Li, S. Lv, X. Wu, and X. Zhuang, "Efficient parcel damage detection via faster r-cnn: A deep learning approach for logistical parcels'automated inspection," in *20th EAI International Conference, MobiQuitous 2023, Australia, November, 2023*.

[13] T. Kong, A. Yao, Y. Chen, and F. Sun, "Hypernet: Towards accurate region proposal generation and joint object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 845–853.

[14] Y. He, K. Song, Q. Meng, and Y. Yan, "An end-to-end steel surface defect detection approach via fusing multiple hierarchical features," *IEEE transactions on instrumentation and measurement*, vol. 69, no. 4, pp. 1493–1504, 2019.

[15] D. Guo, Z. Wu, J. Feng, and T. Zou, "Multi-scale semantic enhancement network for object detection," *Scientific Reports*, vol. 13, no. 1, p. 7178, 2023.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] T.-H. Sun, F.-C. Tien, F.-C. Tien, and R.-J. Kuo, "Automated thermal fuse inspection using machine vision and artificial neural networks," *Journal of Intelligent Manufacturing*, vol. 27, pp. 639–651, 2016.

[18] R. Borwankar and R. Ludwig, "An optical surface inspection and automatic classification technique using the rotated wavelet transform," *IEEE Transactions on Instrumentation and Measurement*, vol. 67, no. 3, pp. 690–697, 2018.

[19] H. M. Schlüter, J. Tan, B. Hou, and B. Kainz, "Natural synthetic anomalies for self-supervised anomaly detection and localization," in *European Conference on Computer Vision*. Springer, 2022, pp. 474–489.

[20] B. Fang, X. Long, F. Sun, H. Liu, S. Zhang, and C. Fang, "Tactile-based fabric defect detection using convolutional neural network with attention mechanism," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–9, 2022.

[21] Z. Chen, C. Du, X. Huang, Z. Lin, Y. Zhou, Q. Guan, Z. Li, S. Lv, X. Wu, and X. Zhuang, "Deformation and penetration hybrid detection-net for parcels inspection in industrial supply chain," in *ICASSP 2024, Korea (South)*.

[22] Z. Chen, Q. Guan, X. Duan, H. Zhong, Z. Li, S. Lv, J. Li, and Y. Zhou, "Few-shot learning for quality detection of logistical parcels," in *2023 11th International Conference on Information Systems and Computing Technology (ISCTech)*. IEEE, 2023.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[25] C. Yu, P. Chen, J. Dai, X. Wang, W. Zhang, J. Liu, and J. Han, "Focus by prior: Deepfake detection based on prior-attention," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[26] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, and D. Qian, "Prior-attention residual learning for more discriminative covid-19 screening in ct images," *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2572–2583, 2020.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[29] Y. Du, W. Song, Q. He, D. Huang, A. Liotta, and C. Su, "Deep learning with multi-scale feature fusion in remote sensing for automatic oceanic eddy detection," *Information Fusion*, vol. 49, pp. 89–99, 2019.

[30] C. Zhang, Y. Chen, X. Yang, S. Gao, F. Li, A. Kong, D. Zu, and L. Sun, "Improved remote sensing image classification based on multi-scale feature fusion," *Remote Sensing*, vol. 12, no. 2, p. 213, 2020.

[31] D. Wei, J. Chen, T. Luo, T. Long, and H. Wang, "Classification of crop pests based on multi-scale feature fusion," *Computers and Electronics in Agriculture*, vol. 194, p. 106736, 2022.

[32] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, and A. Li, "Hifuse: Hierarchical multi-scale feature fusion network for medical image classification," *Biomedical Signal Processing and Control*, vol. 87, p. 105534, 2024.

[33] X. Liu, L. Yang, J. Chen, S. Yu, and K. Li, "Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 71, p. 103165, 2022.

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[35] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022.