DOI: 10.19884/j.1672-5220.202404009

Improvement of High-Speed Detection Algorithm for Nonwoven Material Defects Based on Machine Vision

LI Chengzu#, WEI Kehan#, ZHAO Yingbo, TIAN Xuehui, QIAN Yang, ZHANG Lu, WANG Rongwu*
College of Textiles, Donghua University, Shanghai 201620, China

Abstract: Defect detection is vital in the nonwoven material industry, ensuring surface quality before producing finished products. Recently, deep learning and computer vision advancements have revolutionized defect detection, making it a widely adopted approach in various industrial fields. This paper mainly studied the defect detection method for nonwoven materials based on the improved NanoDet-Plus model. Using the constructed samples of defects in nonwoven materials as the research objects, transfer learning experiments were conducted based on the NanoDet-Plus object detection framework. Within this framework, the Backbone, path aggregation feature pyramid network (PAFPN) and Head network models were compared and trained through a process of freezing, with the ultimate aim of bolstering the model's feature extraction abilities and elevating detection accuracy. The half-precision quantization method was used to optimize the model after transfer learning experiments, reducing model weights and computational complexity to improve the detection speed. Performance comparisons were conducted between the improved model and the original NanoDet-Plus model, YOLO, SSD and other common industrial defect detection algorithms, validating that the improved methods based on transfer learning and semi-precision quantization enabled the model to meet the practical requirements of industrial production.

Key words: defect detection; nonwoven materials; deep learning; object detection algorithm; transfer learning; half-precision quantization

CLC number: TP391 **Document code:** A

Article ID: 1672-5220(2024)04-0416-12

Open Science Identity (OSID)

0 Introduction

The nonwoven industry is a key industry focused on the health and safety of the population, playing a crucial role in areas such as epidemic prevention and smog protection[1]. Nonwoven production lines are characterized by their multi-equipment, continuous, refined, and highly collaborative nature. There is a significant demand in the forefront of technology and crucial industries that affect the livelihood of the populace[2]. The importance of high-speed defect detection in nonwoven material production lines for quality control and production efficiency is increasingly prominent. The construction of automated production lines with robotics in the nonwoven industry represents a major trend. To achieve higher levels of industrial chain sophistication and high-end application of production line systems, the application of computer vision technology for high-speed defect detection in nonwoven material production lines can enhance industrial production efficiency, reduce labor costs, and ensure the quality of nonwoven products. Significant achievements have been made in the field of nonwoven material defect detection, gradually moving from traditional manual detection methods to automated and intelligent directions[3]. Traditional defect detection methods often rely on manual visual inspection or simple image processing techniques, suffering from low efficiency and low accuracy[4]. Hence, there is an urgent need for more efficient and precise detection methods to improve production efficiency and product quality. In recent years, with the rapid development of deep learning technology, the defect detection industry has embraced new opportunities.

Deep learning, as a powerful machine learning technology, constructs multi-layer neural network models to learn and extract features from large datasets, enabling efficient processing and analysis of images, videos, and other data types[5-7]. This technology has achieved tremendous success across various fields, including defect detection[8]. Research on fabric defect detection based on deep learning, including techniques like back propagation

Received date: 2024-04-15
Foundation items: National Key Research and Development Program of China (Nos. 2022YFB4700600 and 2022YFB4700605); National Natural Science Foundation of China (Nos. 61771123 and 62171116); Fundamental Research Funds for the Central Universities and Graduate Student Innovation Fund of Donghua University, China (No. CUSF-DH-D-2022044)
* Correspondence should be addressed to WANG Rongwu, email: wrw@dhu.edu.cn
# LI Chengzu and WEI Kehan contribute equally to this work
Citation: LI C Z, WEI K H, ZHAO Y B, et al. Improvement of high-speed detection algorithm for nonwoven material defects based on machine vision[J]. *Journal of Donghua University (English Edition)*, 2024, 41(4): 416-427.

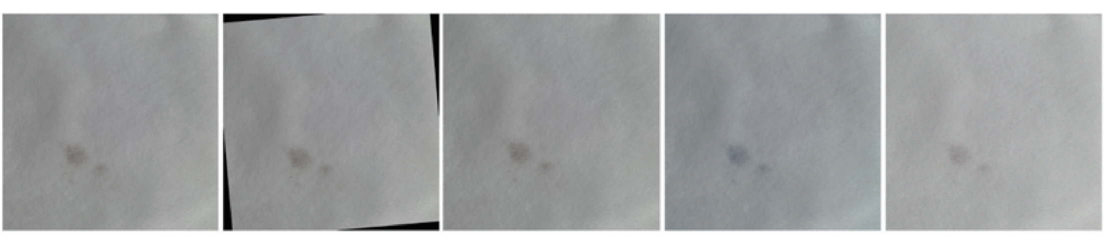
(BP) neural networks[9], autoencoders[10], and convolutional neural networks[11], has been a focus in recent years. For instance, Zhang et al.[12] used Gabor filters to filter the textures of curtain fabric defects and then utilized maximum entropy threshold segmentation to extract feature values for input into BP neural networks, achieving a recognition rate of 94%. Li et al.[13] proposed a stacked autoencoder based on the Fisher criterion (FCSAE) to predict the probability of each pixel belonging to a defect area based on contextual information centered around the pixel, enhancing feature recognition. Jing et al.[14] introduced an end-to-end automatic detection based on convolutional neural network (CNN), using a fast region-based CNN (fast R-CNN) for defect detection in fabric images, which accurately located fabric defect areas with improved adaptability. By constructing deep neural network models, efficient processing and analysis of complex image data could be achieved, thus enhancing the accuracy and efficiency of defect detection. The application of deep learning models like CNN has brought breakthroughs to defect detection, making it more intelligent and reliable.

Among many deep learning models, the NanoDet-Plus model has become a research hotspot in the field of nonwoven material defect detection due to its efficient target detection performance and adaptability to various scenarios[15]. NanoDet-Plus combines the advantages of object detection and instance segmentation, accurately detecting and locating defects in nonwoven materials. Through improvements and optimizations of the NanoDet-Plus model[16], detection accuracy and speed can be further enhanced, bringing new opportunities and challenges to the development and application of the nonwoven material defect detection field. This research will focus on the NanoDet-Plus target detection model, enhanced with the improvement techniques of transfer learning and half-precision quantization, to achieve fast and accurate detection of defects in nonwoven materials, and provide a more reliable quality control method for industrial production.

1 Image Acquisition and Dataset Creation

During the image acquisition process, thermally bonded nonwoven materials with a surface density of approximately 30 g/m^2 were used as raw materials. Image acquisition was conducted using an MV-CS200-10GC camera (Hikvision, China) with a resolution of 1 024 pixels×1 024 pixels. To simulate the actual defect detection environment on a factory production line, different types of defects were captured under the same lighting condition, with adjustments made to the brightness of the light source using a circular light source regulator provided by Hikvision. The computer was equipped with the industrial camera adapter client for the machine vision system (MVS), and connected to the camera using a gigabit Ethernet cable.

According to surveys, damage, stains, clumps and wrinkles occur most frequently in thermally bonded nonwoven materials, accounting for over 80% of all types of defects. This project would focus on capturing images of damaged, stained, clumped and wrinkled defects. To enhance the generalization ability and detection performance of the model, and to avoid overfitting caused by insufficient samples of nonwoven material defects, the original images were subjected to affine transformation image processing to augment the original samples. Rotational transformation, scaling variation, changing image color channels and adding random noise methods were used to process the original images, as shown in Fig. 1. After data augmentation, a total of 6 346 samples were collected. The collected samples were classified and statistically analyzed into four types of defect samples as shown in Table 1. There are 1 247 images of damaged defects, 1 278 images of stained defects, 1 224 images of clumped defects, 1 235 images of wrinkled defects and 1 362 images of normal samples. The LabelImg image annotation tool was used to annotate, name and classify the created defect dataset. Each annotated image data was saved to an independent folder and stored in a .xml file format.



(a) (b) (c) (d) (e)

Fig. 1 Comparison of effects before and after affine transformation: (a) original; (b) rotational transformation; (c) scaling variation; (d) changing image color channels; (e) adding random noise

Table 1 Four defect types of sample image classification statistics

Defect type	Label	Extended data	Training set	Test set
Normal	normal	1 362	1 090	272
Damage	posun	1 247	998	249
Stain	wuzi	1 278	1 022	256
Wrinkle	zhezhou	1 235	988	247
Clump	tuankuai	1 224	979	245
Total		6 346	5 077	1 269

2 NanoDet-Plus Object Detection Algorithm Model Framework

NanoDet-Plus is a one-stage anchor-free object detection model based on CNN. It integrates the design principles of fully convolutional one-stage (FCOS) object detection and aims to provide an efficient and accurate solution for lightweight object detection. The algorithmic model framework of NanoDet-Plus is illustrated in Fig. 2. Backbone, path aggregation feature pyramid network (PAFPN) and Head network are used. In Fig. 2, L_{cls} corresponds to the quality focal loss (QFL) in the generalized focal loss[17]; L_{dist} corresponds to distribution focal loss (DFL) in the generalized focal loss; L_{iou} is the generalized intersection over union (IoU)[18] loss. $L_{cls,A}$, $L_{dist,A}$ and $L_{iou,A}$ respectively represent the losses corresponding to the auxiliary generative module (AGM). This model is meticulously designed and optimized through feature extraction, feature fusion and detection modules. It also introduces lightweight training auxiliary modules and innovative soft label assignment strategies, making the model easier to train and deploy. Evaluated over a range of IoU threshold from 0.50 to 0.95, NanoDet-Plus, a lightweight object detection model with a size of less than 5 MB, achieves a mean average precision (mAP) of 30.4% on the common objects in context (COCO) dataset. Moreover, it can achieve a detection speed of 97 frames per second (10.31 ms/frame) on mobile devices. This performance not only rivals models from the YOLO series but also has significant advantages in model lightweight and inference speed.

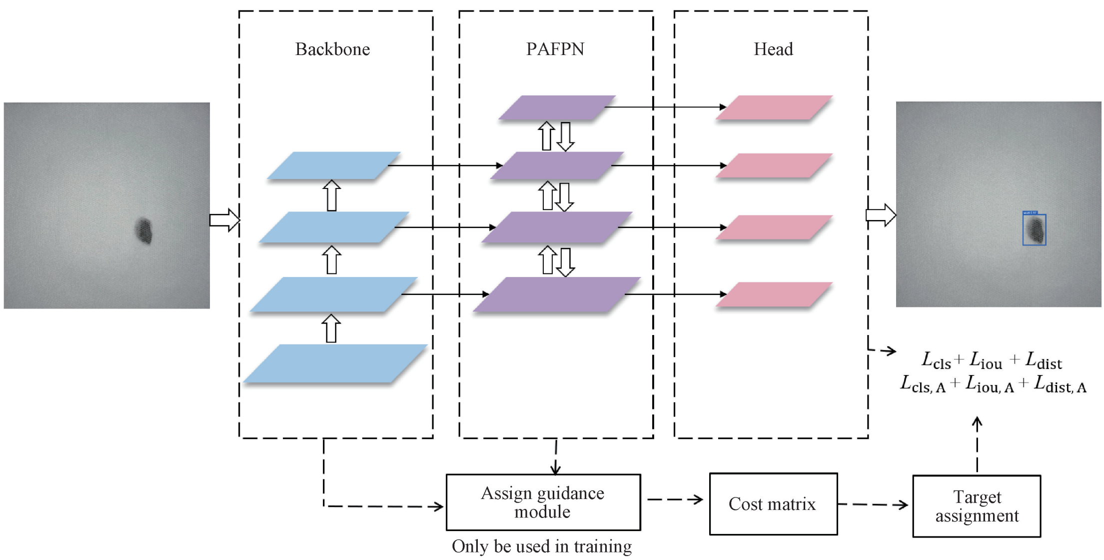


Fig. 2 Diagram of NanoDet-Plus object detection model framework

3 Application of Transfer Learning in NanoDet-Plus

In deep learning-based object detection and classification tasks, the success of model training largely depends on two important prerequisites: an adequate quantity of various class image samples in the dataset; the need for these sample data to be evenly distributed among different classes. However, in the nonwoven material production process, meeting both requirements simultaneously has significant challenges. On one hand, there is a scarcity of high-quality open-source datasets for nonwoven material defects, as many existing datasets are not widely shared, and most factories face difficulties in collecting samples due to the need to ensure work efficiency and the decrease in defect rates of material products with the improvement of production equipment performance. On the other hand, the complexity of nonwoven material defect types and the scarcity of corresponding samples lead to an uneven distribution of samples in the training dataset, thereby affecting the accuracy and speed of the model in real-time detection tasks[19].

Therefore, to make deep learning more widely applicable to nonwoven material defect detection tasks, it is necessary to overcome the challenges of training data availability and the uneven distribution of sample types in training and test data. Faced with these challenges, transfer learning[20] can be used to address the unmet target data requirements in defect detection tasks. Through transfer learning, knowledge learned by pre-trained models on rich datasets can be utilized to adapt to nonwoven material defect detection tasks with few samples and uneven distribution of sample types, thereby improving the model's generalization ability and detection performance[21].

3.1 Basic concepts of transfer learning

Transfer learning, as a method in machine learning, is as the name implies: it involves saving the parameters of a trained source task model and transferring them to a new task algorithm model to improve the learning efficiency and the performance of the new model. It can be understood as reusing knowledge learned in a mature field for a new field, achieving significant benefits with less effort. This method is based on the premise that there may be shared knowledge or patterns among different tasks, and by transferring this knowledge, the amount of data required to learn the new task can be reduced, speeding up the training process, and even improving the model's generalization ability. It is generally used to address the dependency of deep learning on data volume and label categories.

Transfer learning typically involves two main concepts namely domain and task[22]. The domain could be divided into the source domain and the target domain. In the context of transfer learning, each domain (source and target) is associated with a specific task, which dictates the learning objectives and activities within those domains. The source task represents the task that the model is originally trained to solve, and the knowledge gained by the model on this task will be used for another task; while the target task is the new task that the model is intended to learn and perform well on. The target task may be highly related to the source task, or it may be only similar in certain aspects, as shown in Fig. 3.

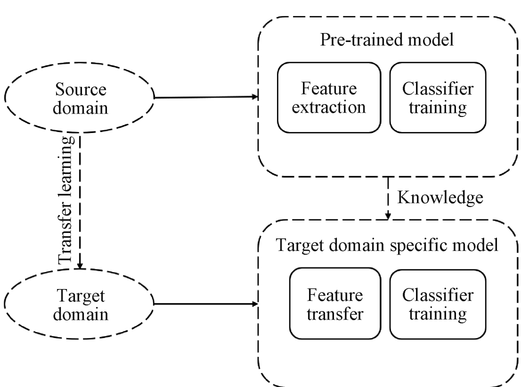


Fig. 3 Conceptual diagram of transfer learning

In the process of transfer learning, a domain $\mathcal{D}$ consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(x)$:

$$\mathcal{D} = \{\mathcal{X}, P(x)\}, x \in \mathcal{X}, \tag{1}$$

where x is a sample in the feature space $\mathcal{X}$. A task $\mathcal{T}$ consists of a label space $\mathcal{Y}$ and a target prediction function $f(x)$:

$$\mathcal{T} = \{\mathcal{Y}, f(x)\}, y \in \mathcal{Y}, \tag{2}$$

where y is a sample in the label space $\mathcal{Y}$. From a probabilistic perspective, $f(x)$ is equivalent to $P(Y \mid X)$, where X represents a given feature, and Y represents the label. Thus, Eq. (2) can also be represented as

$$\mathcal{T} = \{\mathcal{Y}, f(x)\} = \{\mathcal{Y}, P(Y \mid X)\}, y \in \mathcal{Y}. \tag{3}$$

For a given source domain $\mathcal{D}_s$ and learning task $\mathcal{T}_s$, and a given target domain $\mathcal{D}_t$ and learning task $\mathcal{T}_t$, the goal of transfer learning is to learn and transfer knowledge from the source domain $\mathcal{D}_s$ to improve the learning effect of the target domain task $\mathcal{T}_t$, where $\mathcal{D}_s \neq \mathcal{D}_t$ or $\mathcal{T}_s \neq \mathcal{T}_t$.

The core advantage of transfer learning lies in its ability to significantly improve learning efficiency and model performance on the target task when there is limited data or high acquisition costs, by leveraging a large amount of data and learned features from the source task. Model-based transfer learning methods are relatively simple yet effective in this process. The core ideas of these methods are to use pre-trained network models on one or more source tasks as the starting points for the new target tasks, and further fine-tune the models to adapt and optimize them for the target tasks. This is also the most classic pre-training-fine-tuning method in model-based transfer learning methods.

3.2 Application of transfer learning in NanoDet-Plus

Considering the practical application scenario of transfer learning-based nonwoven material defect detection algorithms, and the insufficient number of samples in defect datasets, we use the classic COCO dataset in the object detection domain as the source data for pre-training models. This would yield a NanoDet-Plus object detection model that has been pre-trained and achieved good results. Next, we need to train and identify nonwoven material defect datasets collected and preprocessed.

In this scenario, the specific steps of the nonwoven material defect detection algorithm based on the NanoDet-Plus model and transfer learning are as follows. Firstly, use the training set from the nonwoven material defect dataset to fine-tune the model pre-trained on the COCO dataset. Then, use a portion of the test set from the defect dataset to test the newly fine-tuned NanoDet-Plus model and compare the performance parameters. Figure 4 represents the algorithm training process under this pre-training-fine-tuning method.

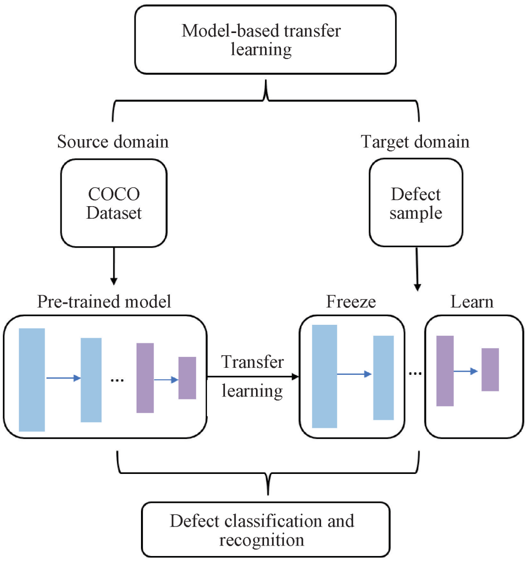


Fig. 4 Conceptual diagram of model-based transfer learning algorithm improvement

From the theoretical perspective of deep learning and transfer learning, the main advantages of using this method are twofold. On the one hand, model-based transfer learning methods reuse knowledge originally belonging to the COCO object detection dataset in this defect detection task, thereby improving the classification and recognition accuracy of the new version of the NanoDet-Plus model. On the other hand, the pre-training-fine-tuning method of transfer learning does not require retraining the original network, and only partial parameters need to be fine-tuned, greatly reducing the number of network parameters that need to be trained.

After obtaining the pre-trained model, for the fine-tuning strategy, the three key nodes in the NanoDet-Plus model structure, Backbone, PAFPN and Head are used as the baseline for transfer learning. Their network parameters are sequentially frozen and other parts are trained. This fine-tuning strategy is then compared with the strategy of not freezing any layers' network parameters, specifically comparing the performance differences of the following four strategies: freeze Backbone's fine-tuning strategy (Backbone-frozen), freeze PAFPN's fine-tuning strategy (PAFPN-frozen), freeze Head's fine-tuning strategy (Head-frozen), and fine-tuning strategy without freezing any layers (None-frozen) in this defect detection task.

3.3 Performance comparison of different transfer methods

The experimental platform hardware is based on GPU-RTX2080Ti, with the Windows 10 operating system. The software environment is based on the programming language Python 3.7 and uses PyTorch 1.7.1 deep learning framework. During the experiment, the training parameters of the neural network training framework are set as follows: input image dimension is 3×1 024 pixels ×1 024 pixels, batch size is 24, initial learning rate is 0.001, and epoch is set to 300.

As demonstrated in Section 3.2 regarding experimental methods, the difference between these four fine-tuning strategies based on transfer learning lies in the freezing of network parameters for the Backbone, PAFPN and Head structures in the NanoDet-Plus model structure. The model structure is shown in Fig. 5. These fine-tuning strategies are important factors affecting the final model accuracy.

In Fig. 5, C3, C4 and C5 of the Backbone module represent the feature map of the backbone network, that is, the feature map obtained by the backbone network after feature extraction of input images. PAFPN module's P3, P4, P5, P6 and P7 are the feature levels used for final prediction. H and W are the height and the width of the feature graph, respectively. For example, the input resolution for a defect image is 1 024 pixels × 1 024 pixels.

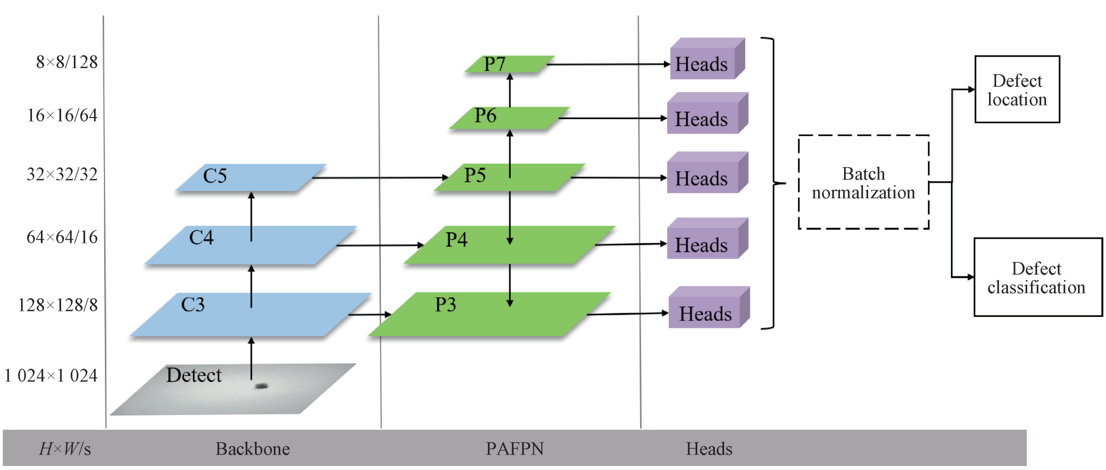


Fig. 5 Diagram of Backbone, PAFPN and Head structures in NanoDet-Plus model

Below, we will compare in detail the defect recognition performance of these four different fine-tuning strategies for nonwoven material defect detection.

1) Backbone-frozen. The Backbone in the NanoDet-Plus model is based on the same backbone network ShuffleNetV2 as the previous generation NanoDet model. This method freezes and retains the weight parameters of the pre-trained Backbone in the model and conducts the secondary training on the other modules/layers of the model.

2) PAFPN-frozen. Compared to the previous generation, the PAFPN in the NanoDet-Plus model has been redesigned to a lightweight and high-performance PAN, namely Ghost-PAN. It uses GhostBlock as the module for handling feature fusion between multiple layers. This freezing method retains the weight parameters of the pre-trained PAFPN in the model and conducts secondary training on the other modules/layers of the model.

3) Head-frozen. The NanoDet-Plus model changes the normalization method of the detection head from group normalization in the previous NanoDet model to batch normalization. This method directly integrates the normalized model structure parameters into the convolution and the kernel size of the depthwise convolution in the detection head changes from 3 pixels× 3 pixels to 5 pixels ×5 pixels. This freezing method retains the weight parameters of the pre-trained Head in the model and conducts secondary training on the other modules/layers of the model.

4) None-frozen. This method does not freeze any network structures of the NanoDet-Plus model pre-trained on the original COCO dataset. After obtaining the pre-trained model, it directly utilizes the nonwoven defect dataset constructed in this paper for secondary training based on all convolutional layers and fully connected layers pre-trained on the COCO dataset.

After training models with four different fine-tuning strategies on the dataset, the experimental results are shown in Table 2, and the recognition accuracy is characterised by mAP.

Table 2 Experimental results of models using four different fine-tuning strategies

Different fine-tuning strategies for application transfer learning	Label	mAP/%
Original NanoDet-Plus model	Original	82.9
Freeze Backbone's fine-tuning strategy	Backbone-frozen	87.0
Freeze PAFPN's fine-tuning strategy	PAFPN-frozen	89.1
Freeze Head's fine-tuning strategy	Head-frozen	87.5
Fine-tuning strategies without freezing	None-frozen	88.1

The accuracy curves and loss curves during the training process of different models are shown in Fig. 6. The confusion matrices of different defect categories when the IoU threshold is 0.45 and the confidence threshold is 0.25 during model training are shown in Fig. 7. The results show that under the constructed nonwoven defect dataset, the combination of the pre-trained model and the dataset results in a new model with reused knowledge from transfer learning, which improves performance compared to the original model. This indicates that the knowledge learned from the COCO dataset is helpful for the classification and recognition of nonwoven material defects through transfer learning optimization. Among them, PAFPN-frozen has a higher stable value than other fine-tuning strategies after 300 rounds of training, and the fluctuation of training loss is also close to None-frozen. This shows that the new NanoDet-Plus model PAFPN-frozen has better recognition performance and does not need to retrain all network parameters. Next, PAFPN-frozen is used as the improved new model for subsequent optimization of detection speed performance.

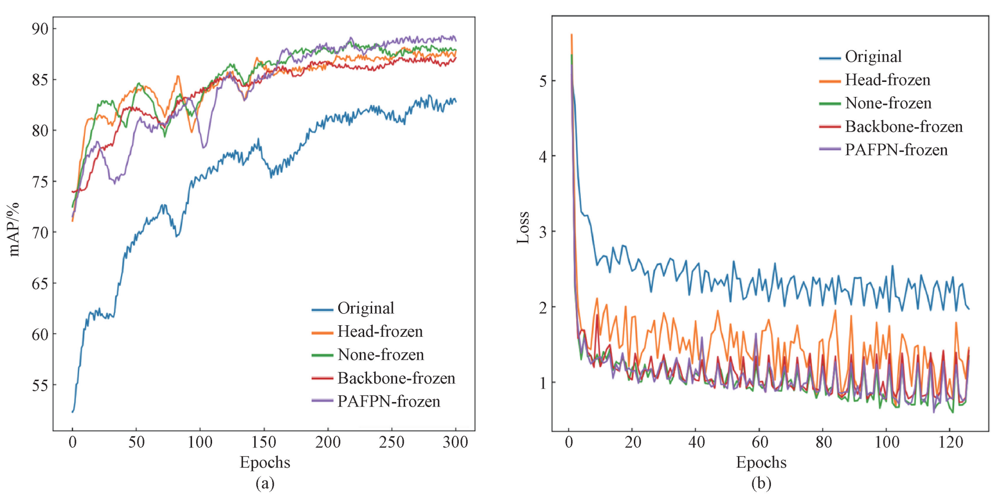


Fig. 6 Training curves of models with different fine-tuning strategies: (a) accuracy curves; (b) loss curves

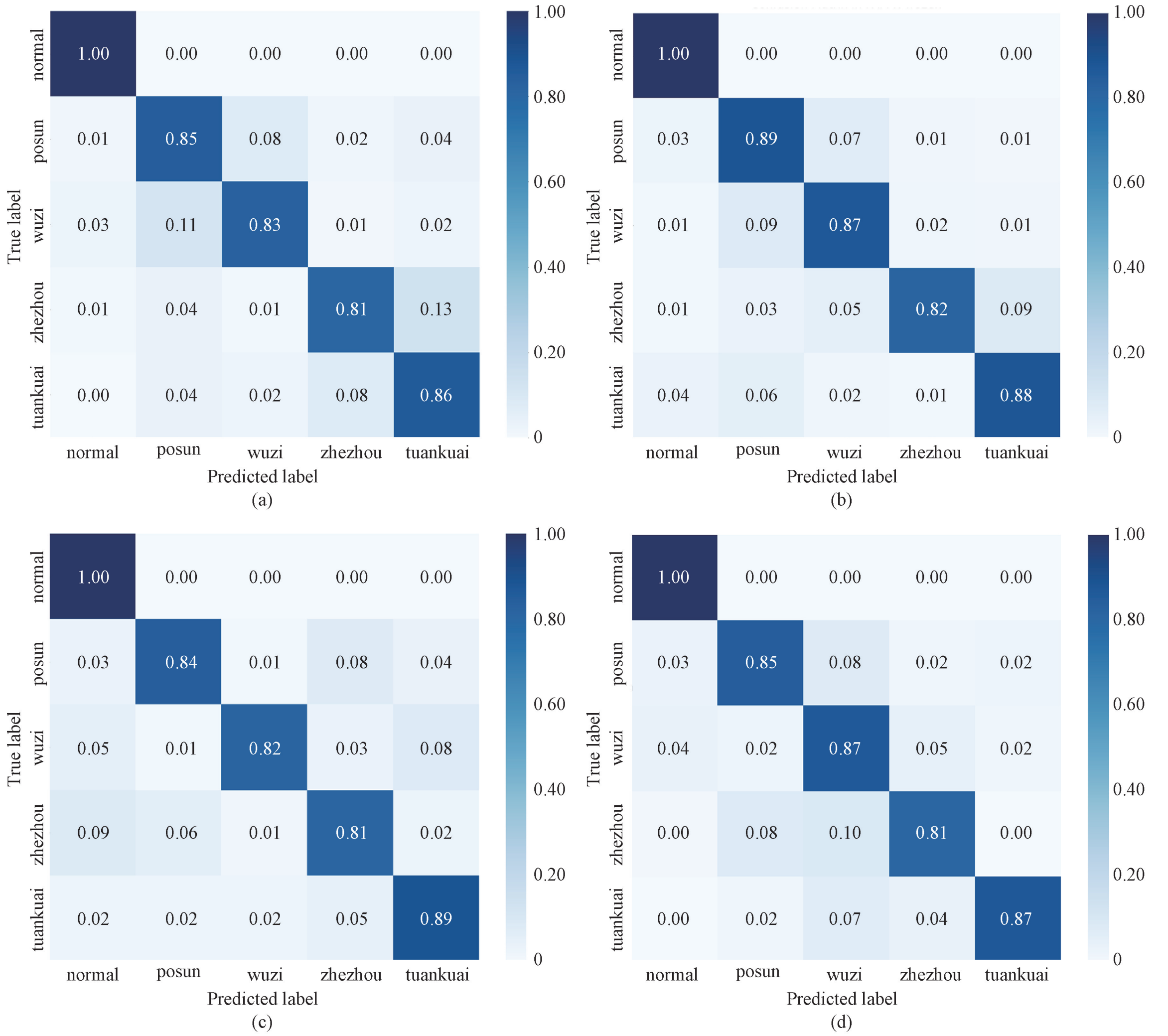


Fig. 7 Confusion matrix of defect categories for models with different fine-tuning strategies: (a) Backbone-frozen; (b) PAFPN-frozen; (c) Head-frozen; (d) None-frozen

4 Half-Precision Quantization Application in NanoDet-Plus

The NanoDet-Plus model improved through transfer learning has shown a certain increase in accuracy for defect detection tasks. However, in the actual production process of nonwoven materials, due to the improvement of industrial performance, the deployment speed of various machines has greatly increased. This imposes higher demands on the real-time detection speed for defect identification tasks, requiring detection devices to process more real-time input images in shorter time frames. Essentially, it challenges the core algorithm performance of the detection devices. Therefore, in this section, the method of half-precision quantization is introduced into the NanoDet-Plus model which has already been improved through transfer learning. This aims to verify that, while minimizing the loss of model accuracy, significant improvements in the detection speed can be achieved, thus meeting the requirements of industrial production in the future.

4.1 Basic principles and advantages of half-precision quantization

Quantization is a technique for optimizing deep learning models, aiming to reduce the precision of model parameters or activation values. Half-precision quantization is a form of quantization that converts parameters and activation values in the model from 32-bit floating-point (FP32) to 16-bit floating-point (FP16). This conversion accelerates the training and inference processes of the model by reducing the required storage space and computational resources while minimizing the precision loss.

Without quantization, a model parameter (such as weight) can be represented by an FP32 number W_{FP32}:

$$W_{FP32} = 1.23456789. \tag{4}$$

Applying the method of half-precision quantization to this weight, simplifies it to an FP16 number W_{FP16}:

$$W_{FP16} = 1.2345. \tag{5}$$

Through this conversion, the storage requirements of the data are halved, thereby reducing memory usage and improving the processing speed, especially on hardware that supports FP16 computation.

4.2 Method and effects of applying half-precision quantization in NanoDet-Plus

The transfer learning fine-tuning strategy of PAFPN-frozen was used as the improved new model for subsequent optimization of detection speed performance. In the PyTorch deep learning framework, an application programming interface (API) is provided to implement half-precision quantization of models and data: "model .half()". This method call internally converts the data types of the model and data from single-precision floating-point "torch.float32" to half-precision floating-point "torch.float16". The implementation of the NanoDet-Plus model involves the following key steps.

1) Model loading and parameter conversion. Load the NanoDet-Plus model optimized through transfer learning. This model consists of several layers, each containing specific weights and configuration parameters. By invoking the model .half() method, all layer parameters are converted from single-precision floating-point (FP32, torch.float32) to half-precision floating-point (FP16, torch.float16). This conversion reduces the model's memory footprint and enhances the computational efficiency.

2) Training set image data preprocessing and model training. For the input defect training set image data, it needs to be converted to half-precision type using the .half() method. After converting the input data from FP32 to FP16, the model training begins, resulting in a new model improved through half-precision quantization. In the half-precision data format, the model learns from these training data, optimizing its parameters to adapt to the precision changes brought by quantization.

3) Test set image data preprocessing. Similarly, the input defect test set image data also needs to be converted to the half-precision type. After conversion, the improved model is used for testing.

4) Performance evaluation. Test the input defect image data using the improved NanoDet-Plus model converted to half-precision format. During this phase, both the forward propagation (i.e., the model's prediction process) and the error backpropagation (used to optimize model weights) are conducted in a half-precision environment. The generated experimental results include metrics such as classification accuracy and detection speed, which are used to evaluate the impact of half-precision quantization on model performance and further analyze the cost-effectiveness of half-precision quantization in actual model deployment.

After conducting training experiments on the improved model, the results are shown in Table 3. NanoDet-Plus-TL is the model improved through transfer learning, and NanoDet-Plus-TL & HP is the new model improved through both transfer learning and half-precision quantization. The experimental results show that NanoDet-Plus-TL & HP, with only a 1.8% loss in accuracy, reduced the average time required to process each image (computational latency) by 18.05 ms. Moreover, compared to the original NanoDet-Plus model before improvements, there is a comprehensive enhancement in the defect detection accuracy and speed. This indicates that the new NanoDet-Plus model improved through the application of half-precision quantization, possesses faster recognition speed, meets the requirements for defect detection performance in nonwoven material industrial production, and balances the recognition speed and accuracy, demonstrating strong algorithm applicability. The defect detection results of the original model and the improved model are shown in Fig. 8. Next, it will be compared with common industrial defect detection algorithms such as Yolo and SSD, to evaluate the application characteristics of this newly optimized and improved model.

Table 3 Comparison of training experiment results before and after improvement

Model	mAP/%	Detection speed/(ms/frame)
NanoDet-Plus	82.9	29.02
NanoDet-Plus-TL	89.1	29.02
NanoDet-Plus-TL & HP	87.9	10.97

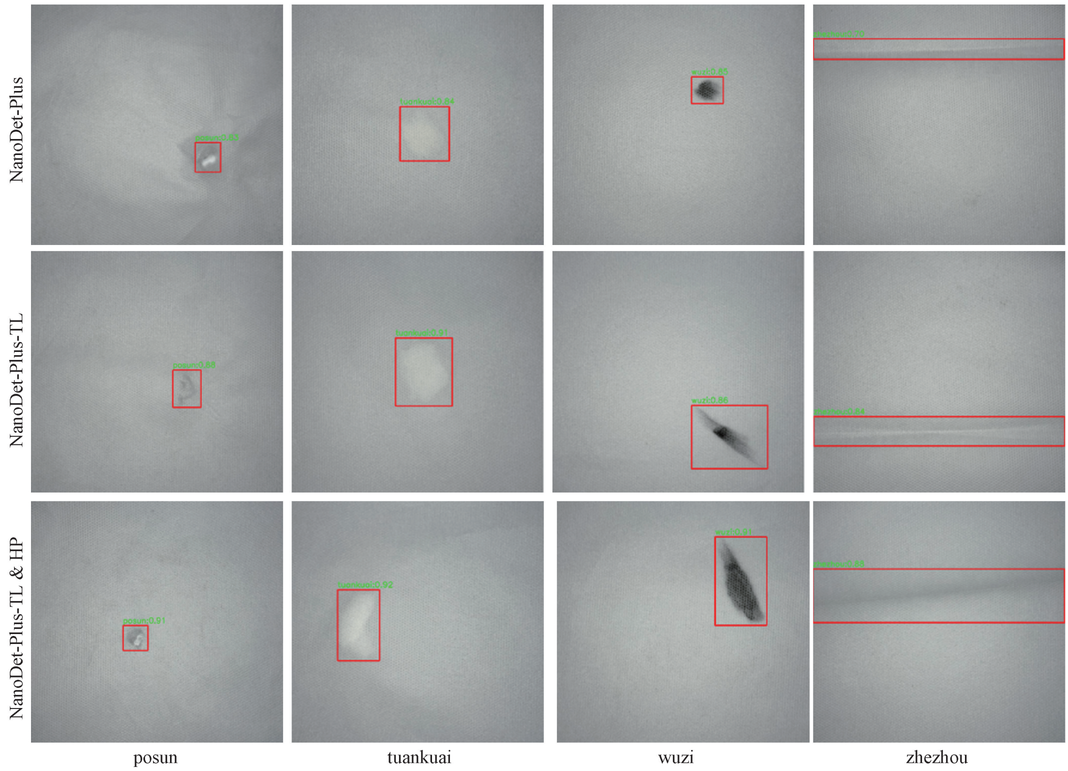


Fig. 8 Defect detection effect before and after model improvement

5 Comparative Experiment and Result Analysis

5.1 Model training parameters

The comparative experiments are based on training the network using the constructed nonwoven material defect dataset. The dataset is divided into training and test sets at a ratio of 8 : 2. The hardware platform, software environment, and network training parameter settings for the specific comparative experiments are detailed in Table 4.

Table 4 Parameters for comparative experiment

Environment configuration	Version	Training parameter	Setup
GPU	RTX2080Ti	Input image dimension	3×1 024×1 024
Operating system	Windows 10	Batch size	24
Programming language	Python 3.7	Initial learning rate	0.001
Deep learning framework	PyTorch 1.7.1	Epoch	300

5.2 Evaluation metrics

To measure the detection performance of the improved new model from multiple dimensions, the following indicators are introduced as dimensional indicators for analyzing experimental results.

1) mAP. After model training, mAP is calculated using the test set to validate the classification accuracy of the model. It measures the average precision of the model at different IoU thresholds, assessing the detection accuracy of the model on different target detection categories.

2) Latency. Latency is the time required for the model to process a single image during classification testing, usually measured in milliseconds. Lower latency indicates better real-time performance of the model. In the actual production environment of nonwoven materials, the defect detection task requires high real-time image processing capabilities, making the latency parameter crucial.

3) Floating point operations (Flops). The number of Flops required by the model during inference, usually measured in billions (G) or trillions (T) of operations. Lower Flops indicate higher computational efficiency of the model, representing fewer computational resources required during inference.

4) Parameters (Params). The number of parameters in each layer of the model, i. e., the total number of trainable parameters in the model. The parameter count reflects the complexity of the model. Generally, a higher parameter count implies stronger model representation

capability, but it also increases computational and storage costs.

5) Model size. Model size is the file space occupied by the model on disk, and is usually measured in megabytes (MB) or gigabytes (GB). Model size is influenced by the number of parameters and storage precision (e. g., single precision or half precision). A larger model size indicates a need for more storage space.

5.3 Comparative analysis of experimental results

After training the improved NanoDet-Plus model, the original NanoDet-Plus model, SSD, and the Yolo series of classic object detection models on a constructed dataset of nonwoven material defect data, the experimental indicators are shown in Table 5. The defect images have a resolution of 1 024 pixels × 1 024 pixels.

Table 5 Experimental results of each detection model

Model	mAP/%	Latency/ms	Flops/G	Params/M	Model size/MB
NanoDet-Plus	82.9	29.02	1.75	7.06	30.0
NanoDet-Plus-TL & HP	87.9	10.97	1.75	7.06	16.0
SSD	83.1	32.24	30.00	11.59	103.0
Yolov5s	89.8	32.38	16.40	7.26	54.2
Yolov7	91.3	129.52	104.60	36.90	296.3

Table 5 shows that compared to the original NanoDet-Plus model, the NanoDet-Plus-TL & HP optimized through transfer learning and half-precision quantization increases mAP by 5.0%, improves the inference latency by 18.05 ms, and reduces the model size by 46.7%. Compared to the NanoDet-Plus-TL & HP, the SSD model performs worse in terms of the detection accuracy and the inference latency on the defect dataset. The Yolov5s model, with only a 1.9% improvement in accuracy, falls behind by 21.41 ms in the detection speed. Meanwhile, the Yolov7 model, which has more parameters, is larger and more complex and improves the detection accuracy by 3.4% on the defect dataset but at the cost of increasing the inference latency, the computational load, the parameter count, and the model size by 11.8 times, 59.8 times, 5.2 times, and 18.5 times, respectively.

Space complexity involves the storage and memory usage of the model parameters. The space complexity of the NanoDet-Plus model is mainly determined by the number of parameters and the storage precision. The original NanoDet-Plus model has 7.06 M parameters, with a model size of 30.0 MB. After optimization with transfer learning and half-precision quantization, the NanoDet-Plus-TL & HP model retains the same number of parameters, but the model size is reduced to 16.0 MB.

Considering the need for cost-effective detection in actual production, where nonwoven material defect detection models may need to be deployed to mobile devices with lower computing capabilities, and the high demands of the production line for real-time detection speed, it can be concluded that in the field of nonwoven material defect detection, the NanoDet-Plus target detection model improved through transfer learning and half-precision quantization is more suitable. It balances the detection speed and the accuracy with strong algorithm applicability and meets the defect detection requirements in industrial production, and the algorithm's lower hardware requirements allow it to be applied cost-effectively in actual production lines.

6 Conclusion

Defect detection is a crucial part of the nonwoven material industrial production process. Its purpose is to detect and locate defects on the surface of nonwoven materials before making finished products, ensuring the quality of nonwoven materials. In recent years, with the development of deep learning and computer vision technology, deep learning for defect detection has been widely applied in various fields of industrial production. The nonwoven material defect detection method based on the improved NanoDet-Plus model was studied using the constructed nonwoven material defect samples as the research object. Based on the NanoDet-Plus target detection model, combined with transfer learning experiments, the Backbone, PAFPN and Head network model structures in the model were compared and frozen to enhance the feature extraction capabilities of the model and improve the detection accuracy. Half-precision quantization method was used to optimize the model after transfer learning experiments, reducing the model weight and the computational complexity to improve the detection speed. Performance comparison of the improved model with the original NanoDet-Plus model, YOLO, SSD, and other common industrial defect target detection algorithms validated that the improvement method based on transfer learning and half-precision quantization enabled the model to meet the actual requirements of industrial production.

References

[1] ZHANG R T, Qi X N. A study of China's

nonwovens industry based on the SCP paradigm [J]. *Information Systems and Economics*, 2022, 3(4): 1-8.

[2] LIU S P. Research on the mechanism and path of the manufacturing industry upgrading with "Internet Plus" [D]. Changsha: Zhongnan University of Economics and Law, 2019. (in Chinese)

[3] XU W F, DU J. Discussion on green innovation and development of textile industry enabled by digital technology [J]. *Cotton Textile Technology*, 2023, 51 (8): 33-37. (in Chinese)

[4] NGAN H Y T, PANG G K H, YUNG N H C. Automated fabric defect detection: a review [J]. *Image and Vision Computing*, 2011, 29 (7): 442-458.

[5] BOU NASSIF A, ABU TALIB M, NASIR Q, et al. Machine learning for anomaly detection: a systematic review [J]. *IEEE Access*, 2021, 9: 78658-78700.

[6] ZHU R H, XIN B J, DENG N, et al. Review of fabric defect detection based on computer vision [J]. *Journal of Donghua University* (*English Edition*), 2023, 40(1): 18-26.

[7] LI F, PAN H S, SHENG S X, et al. Image retrieval based on vision transformer and masked learning [J]. *Journal of Donghua University* (*English Edition*), 2023, 40(5): 539-547.

[8] KHAN A, LAGHARI A, AWAN S. Machine learning in computer vision: a review [J]. *ICST Transactions on Scalable Information Systems*, 2021, 8(32): 169418.

[9] LI M, YANG S, HE R H, et al. Jacquard fabric defect detection technology combining context-awareness and convolutional neural network [J]. *Modern Textile Technology*, 2021, 29 (6): 62-66. (in Chinese)

[10] WEI W, DENG D X, ZENG L, et al. Real-time implementation of fabric defect detection based on variational automatic encoder with structure similarity [J]. *Journal of Real-Time Image Processing*, 2021, 18(3): 807-823.

[11] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]// Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. New York: ACM, 2014: 580-587.

[12] ZHANG W Y, YANG Y, LIN C, et al. Method of fabric defect detection based on Gabor filter and BP neural network [J]. *Journal of Zhongyuan University of Technology*, 2014, 25 (3): 1-6. (in Chinese)

[13] LI Y D, ZHANG J Y, LIN Y B. Combining fisher criterion and deep learning for patterned fabric defect inspection [J]. *IEICE Transactions on Information and Systems*, 2016, 99 (11): 2840-2842.

[14] JING J F, FAN X T, LI P F, et al. Yarn-dyed fabric defect detection based on deep-convolutional neural network [J]. *Journal of Textile Research*, 2017, 38 (2): 68-74. (in Chinese)

[15] LIU F J, LI Y. SAR remote sensing image ship detection method NanoDet based on visual saliency [J]. *Journal of Radars*, 2021, 10(6): 885-894. (in Chinese)

[16] XU Y T. Defect detection of automotive composite leather based on NanoDet-Plus [D]. Wuhan: Jianghan University, 2023. (in Chinese)

[17] QI P Y, WANG H Y, ZHANG J, et al. Crowded pedestrian detection algorithm based on improved FCOS [J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 811-818.

[18] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, IEEE, 2019: 658-666.

[19] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection [EB/OL]. (2020-10-23) [2024-4-15]. https://arxiv.org/abs/2004.10934v1.

[20] HUGHES D P, SALATHE M. An open access repository of images on plant health to enable the development of mobile disease diagnostics [EB/OL]. (2015-11-25) [2024-4-15]. https://arxiv.org/abs/1511.08060.

[21] LI X L, LI Z X, YANG X F, et al. Boosting the generalization ability of Vis-NIR-spectroscopy-based regression models through dimension reduction and transfer learning [J]. *Computers and Electronics in Agriculture*, 2021, 186: 106157.

[22] PAN S J, YANG Q. A survey on transfer learning [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22 (10): 1345-1359.

基于机器视觉的非织造材料疵点高速检测算法改进

李成族，位珂晗，赵英博，田雪慧，钱　洋，张　璐，王荣武*
东华大学 纺织学院，上海 201620

摘　要：疵点检测在非织造材料工业中至关重要。随着深度学习和计算机视觉技术的发展，深度学习已广泛应用于非织造材料表面疵点的检测和定位，以保证成品质量。该论文主要研究基于改进 NanoDet-Plus 模型的非织造材料疵点检测方法，以构建的非织造材料疵点样本为研究对象，在 NanoDet-Plus 目标检测模型的基础上，结合迁移学习实验对模型中的 Backbone、PAFPN 和 Head 网络模型结构进行对比冻结训练，增强模型特征提取能力以提升检测精度。使用半精度量化方法对迁移学习实验后的模型进行优化，降低模型权重与计算量从而提升检测速度。将改进后的模型与原 NanoDet-Plus 模型、YOLO 和 SSD 等常见的工业化疵点检测算法进行性能对比，验证结果表明，迁移学习与半精度量化相结合的改进方法可使模型满足工业生产的实际需求。

关键词：疵点检测；非织造材料；深度学习；目标检测算法；迁移学习；半精度量化