IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 October 2023, accepted 3 December 2023, date of publication 12 December 2023,
date of current version 20 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3341928

RESEARCH ARTICLE

Automated Surface Defect Detection for Hot-Pressed Light Guide Plates Based on GDA-YOLOv7

ZHENYU LI[1] **AND JUNFENG LI**[1,2,3]
[1]School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2]Tongxiang Research Institute, Zhejiang Sci-Tech University, Jiaxing, Zhejiang 314599, China
[3]Changshan Research Institute, Zhejiang Sci-Tech University, Quzhou, Zhejiang 324299, China

Corresponding author: Junfeng Li (ljf2003@zstu.edu.cn)

This work was supported in part by the Key Research and Development Program of Zhejiang under Grant 2023C01062, and in part by the Basic Public Welfare Research Program of Zhejiang Province under Grant LGF22F030001 and Grant LGG19F030014.

ABSTRACT In this paper, a high-precision hot-pressed light guide plate defect detection model based on improved YOLOv7 is proposed. The model strengthens the spatial correlation between background and foreground by fusing global context information. A densely connected convolutional network is used to enhance the feature extraction capability and mitigate problems such as gradient vanishing while ensuring the maximum information flow in the network. Further, adaptive spatial feature fusion is used in the feature fusion structure of the model; the adaptive spatial feature fusion structure compensates for the small targets that are difficult to extract in high dimensions from low dimensions, thus solving the problem of detecting small targets that are easy to lose. Finally, a self-constructed dataset is built using images of hot-pressed light guide plates collected from industrial sites, and a large number of experiments are conducted. Experimental results show that the defect detection model has a mean average precision (mAP) of 99.1% and a detection speed of 127 FPS. Compared with the mainstream surface defect target detection algorithms, while ensuring the detection speed, the accuracy rate has been significantly improved, and the accuracy rate and real-time can meet the requirements of the industrial field inspection of hot-pressed light guide plate.

INDEX TERMS Hot-pressed light guide plate, defect detection, deep learning, YOLOv7.

I. INTRODUCTION

The Light Guide Plate(LGP) is the primary component of a backlight module and can convert a point light source and a line light source from a Light Emitting Diode (LED) into a uniform surface light source, and the structure of the resulting Liquid Crystal Display (LCD) system is shown in FIGURE 1. Due to the advantages of ultrathin, high transparency, high reflection, uniform and bright light guides, LGPs are commonly used in cell phones, tablets, computers, car navigation and other LCD screens. The quality of LGPs directly affects the quality of the LCD screen. However, when producing LGPs, due to the raw material composition, the use of equipment, processing technology and manual operation and other factors, their surfaces will inevitably exhibit bright spots, line scratches, dark shadows and other defects. To avoid the assembly of defective LGPs into an LCD screen, which would waste more resources, a factory must test for defects in LGPs before they leave the factory to remove them from future production steps.

Traditional defect detection methods include manual inspection and machine vision inspection. Manual inspection is affected by operation time, human eye accuracy, endurance and worker emotion. Traditional machine vision detection often has to go through image preprocessing [1], threshold processing [2], feature selection [3] and other steps, which are easily disturbed by environmental factors such as light and dust. The resulting generalizability is weak, and algorithm stability and versatility is poor, making it difficult to meet the online defect detection of LGPs. Compared with traditional

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

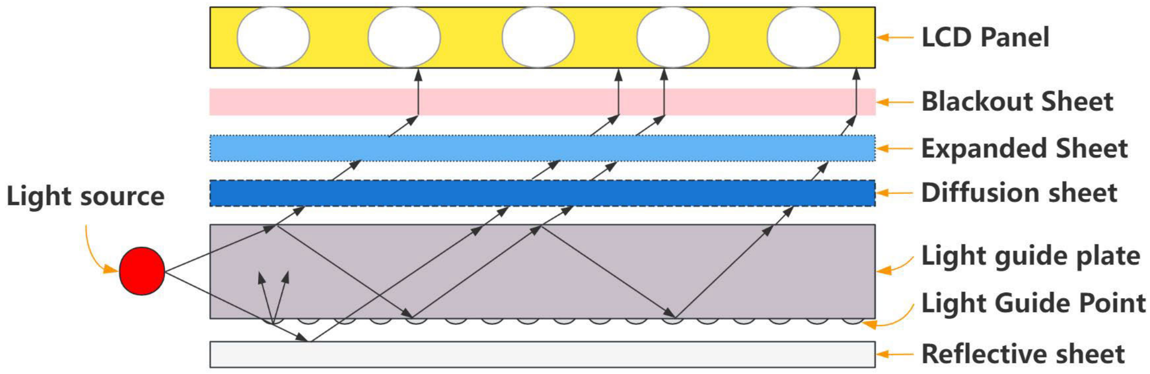

FIGURE 1. Liquid crystal display system structure.

machine vision methods, with the development of deep learning theory and the improvement of computer performance, deep learning-based target detection algorithms are widely used due to their powerful feature expression, generalization and cross-scene capabilities, and the primary fields of study include insulator detection [4], [5], [6], fruit detection [7], [8], [9], face recognition [10], [11], [12], etc.

The defect detection requirements of hot-pressed LGPs are relatively high, and most are smaller than 100-μm white spots, line scratches, white stains, pressure wounds, etc. It is also necessary to provide accurate location information of each defect to optimize the production process and equipment parameters. Deep learning classification network-based defect detection can only obtain coarse localization, the localization accuracy is related to the sliding window size and network classification performance, and the overall detection speed is slow. The target detection network, which can obtain both precise location and classification information of the target, is the closest network to the defect detection task and is generally divided into single-stage and dual-stage networks. The two-stage network first finds the location of the target object to obtain the suggestion frame to ensure sufficient accuracy and recall. Then, this network classifies the suggestion frame to find more accurate locations. The two-stage algorithms with higher accuracy but slower speed primarily include R-CNN [13], SPP-Net [14], FastR-CNN and FasterR-CNN. The single-stage network does not need to obtain the suggestion frame stage and directly generates the class probability and position coordinate values of the object, and the final detection result can be directly obtained by a single detection. Single-stage networks are generally faster than two-stage algorithms with a small loss of accuracy, primarily including SSD, YOLOv3 [15], YOLOv4 [16], YOLOv5, YOLOv6 and YOLOv7 series. In July 2022, the official YOLO team launched YOLOv7, a collection of primarily existing tricks, module re-referencing and dynamic tag assignment strategies that outperform some other known target detectors such as YOLOR,YOLOX,Scaled-YOLOv4, YOLOv5,DETR,DeformableDETR,DINO-5scale-R50 and ViT-Adapter-B in speed and accuracy in the 5FPS to 160FPS range.

Currently, YOLOv7 demonstrates robust detection capabilities in the realm of single-stage object detection networks. However, the surface texture of hot-pressed LGPs is characterized by complexity, with a diverse array of defect types and shapes, coupled with relatively small defect sizes. Utilizing YOLOv7 directly for detecting surface defects on hot-pressed LGPs may result in the loss of semantic information for small target defects. This can lead to lower accuracy or missed detection of small target defects. Therefore, there is a need to enhance YOLOv7 to make it suitable for defect detection on hot-pressed LGPs. Thus, we propose an improved YOLOv7 defect detection network based on global context information, densely connected convolutional networks and adaptive spatial feature fusion. The hot-pressed LGP defect intelligent detection system is successfully applied to industrial sites. The primary contributions of this study are as follows:

(1) The backbone network introduces the Global Context Block (GCBlock), which integrates global contextual information by establishing long-range dependencies among all feature pixels. This enhances spatial correlation between the background and foreground, strengthening the recognition of targets in complex backgrounds and improving the perceptual capability for small target defects.

(2) The neck network incorporates the Densely Connected Convolutional Network (DenseNet), enabling information to flow more freely across different layers of the network. This facilitates the capture of features at various levels, mitigates the issue of vanishing gradients, and is particularly beneficial for training deeper network architectures.

(3) Adding the Adaptive Spatial Feature Fusion (ASFF) structure to the feature fusion mechanism allows for compensating small targets that are challenging to extract from high-dimensional information. This addresses the issue of potential loss of small targets and effectively resolves prediction conflicts across different dimensions. It enhances the adaptability of the network, which is particularly effective in identifying complex and diverse defects in hot-pressed LGPs.

II. RELATED WORK

Currently, the YOLO family of networks and its improved networks are more widely used in target detection. Chen et al. [17] proposed an improved YOLOv5 network model to identify rubber tree diseases. In the backbone network, the bottleneck module in the C3 module was improved so that it is beneficial to capture the long-range information in the spatial range. The attention mechanism SE module was added in the last layer of the backbone network to increase the weight of effective image features. The loss function was changed from GIOU to EIOU to accelerate the convergence of the network model. Although the detection of this network achieves better results, the mAP only reaches 70%, and the detection accuracy still must be improved. Cai et al. [18] proposed a network model based on improved YOLOv4 to achieve the best trade-off between accuracy and speed for autonomous driving. The last output layer of the backbone network was replaced with deformable convolution to enhance the network feature extraction capability. A new feature fusion module PAN++ was then designed to enhance the feature fusion capability in the neck network. A sparse scaling factor method was used to improve the existing channel pruning algorithm to reduce computational resources, and the network achieved marked improvements in accuracy and speed compared to YOLOv4, particularly enhancing the detection of small objects. Jiang et al. [19] proposed an improved attention mechanism YOLOv7 algorithm with three CBAM modules added to the backbone network to improve the network's ability to extract features for counting tasks in dense hemp duck flocks. Su et al. [20] proposed an improved YOLOv5-based algorithm for defect detection in rail fasteners, analyzed the size of fastener defect target boxes using the K-mean algorithm, and analyzed small objects of rail fasteners by combining an attention mechanism and multiscale fusion. Yang et al. [21] proposed an improved YOLOv3-based algorithm for insulator defect detection, changed the FPN to bidirectional fusion to improve the perceptual field of small targets, and added EIoU and Smooth-EIoU loss functions to significantly improve the overlap between the predicted box and the calibrated box and speed up the convergence speed. Lu et al. [22] proposed an improved YOLOv5-based algorithm for integrated circuit (IC) defect detection by adding a prediction head to detect objects at different scales and integrating squeeze-and-excitation layers to improve the feature extraction capability of the network in dense scenes. Chen et al. [23] proposed an improved YOLOv4 algorithm for detecting and counting bayberry trees in drone images. The Leaky ReLU activation function was used to accelerate the model extraction speed, and the DIoU NMS and K-Means clustering methods were used to retain the most accurate prediction boxes. The detection accuracy reached 97.78%. Xu et al. [24] proposed an improved YOLO-v5 algorithm for defect recognition in weld radiographic images. The Coordinate Attention module, SIOU loss function, and FReLU activation function were added to improve the ability to detect small targets, capture low-sensitivity spatial information, and perform global optimization. Li et al. [25] proposed an improved YOLOv4-tiny algorithm for real-time detection of non-motorized vehicles. Dilated convolution and depthwise separable convolution were added to increase the model's receptive field. An improved Spatial Pyramid Pooling module was added to enhance the network's feature extraction capabilities. The improved network's mAP increased by 2.01% compared to the original YOLOv4-tiny network. Wu et al. [26] proposed an improved YOLOv4 algorithm for identifying small target weeds. By modifying the backbone feature extraction and feature pyramid structure, the network's feature expression and small target extraction capabilities were enhanced. A depthwise separable convolution block with a residual structure was introduced to reduce the number of network parameters. The improved network's mAP increased by 4.2% compared to the original YOLOv4 network.

At present, some progress has also been made in the detection of defects in LGPs based on deep learning. Ming et al. [27] proposed a combined classifier with dynamic weights (CCDW)-based LGP defect detection method. Considering the diversity of the underlying classifiers, the proposed CCDW selects the best combination of features to distinguish between defective and defect-free LGP samples. Although the network improved the diversity of feature extraction and the accuracy of the classifier, the detection speed was slow and the accuracy was low. Li and Li [28] proposed an end-to-end multitask learning network architecture for cell phone LGP defect detection. The encoder part uses a similar U-Net encoder, which makes full use of redundant features while increasing the network perception field. The feature fusion part uses feature fusion and multiscale feature interaction, and the segmentation head performs defect segmentation and classification tasks. Although the network improves the positioning accuracy of defects, the structure is relatively complex, and pixel-by-pixel labeling is required when adding a segmentation branch again. Li et al. [29] proposed a two-stage multiscale residual attention network based on "segmentation+decision" for LGP defect detection. The segmentation subnetwork was constructed using a U-shaped structure and designed a multiscale residual attention unit (MRAU) to achieve precise defect location. The segmentation subnetwork was used to extract features, and the decision subnetwork was used to achieve accurate determination of LGP images. A detection accuracy of up to 99% is achieved on the self-built defect detection dataset. Although the network has high detection accuracy, it cannot achieve precise positioning of defects. Hong et al. [30] proposed a dense bilinear convolutional neural network. The introduction of dense blocks, bilinear feature layers, and SE modules improved the network's ability to classify and discriminate defective textures. Although the network model has small parameters and low training

TABLE 1. Summary of defect detection of the LGP.

article	methods	advantages	weaknesses
Ming et al. [27]	A combined classifier with dynamic weights (CCDW)	Improved diversity of feature extraction and classifier accuracy.	Slow detection speed and low accuracy.
Li and Li [28]	An end-to-end multitask learning network architecture	Improvement of defect localization accuracy	The network structure is complex and needs to be labeled pixel by pixel when adding segmentation branches again.
Li et al. [29]	A two-stage multiscale residual attention network based on "segmentation+decision"	High detection accuracy	Accurate positioning of defects cannot be accomplished.
Hong et al. [30]	A dense bilinear convolutional neural network	Small model parameters and low training costs	Low detection accuracy
Yao and Li [31]	AYOLOv3-Tiny-based defect detection network	High detection accuracy and small model parameters	High false detection rate
Li and Yang [32]	An improved YOLOv5 network-based defect detection method	Fast detection speed	High missed detection rate
Li and Wang [33]	An improved RetinaNet network-based defect detection method	Faster training and inference of models	The model structure is more complex.

cost, the detection accuracy is low. Yao and Li [31] proposed an AYOLOv3-Tiny-based defect detection network for PAD light guides. It combined overlap pooling and a spatial attention module to construct an overlap pooling spatial attention module (OSM) to replace the traditional convolution of the backbone network, which can improve the accuracy and enhance the feature extraction ability of defect regions and prevent overfitting. And it used a residual block structure to construct a dilated convolution module (DCM) to improve the detection capability of large defects. Although the network has high detection accuracy and small model parameters, the false detection rate is high. Li and Yang [32] proposed an improved YOLOv5 network-based defect detection method for hot-pressed LGPs. The HAM module combining a spatial attention mechanism and channel attention mechanism enables the network to have higher recognition capability for targets. The perceptual field is also enhanced, and feature extraction capability is improved using dilated convolution with different expansion rates. Although the network has a fast detection speed, the rate of missed detection is high. Li and Wang [33] proposed a visual detection method for light guide plate defects based on an improved RetinaNet. The improved ResNeXt50 with the lightweight Ghost Module was used as the backbone network, reducing resource parameters and consumption, and improving training and inference speed. The feature pyramid network module was improved to more effectively fuse shallow and high-level semantic information, further enhancing the detection ability of small target defects. Although the model training and inference speed is fast, the structural model is relatively complex. There is a summary table provided, as shown in TABLE 1.

III. HOT-PRESSED LGP DEFECT DETECTION SYSTEM

The defect detection device for hot-pressed LGPs primarily consists of a transmission device, image acquisition device and image processing device, as shown in FIGURE 2. The transmission device primarily consists of a conveyor belt and conveyor belt rollers. The image acquisition device is primarily composed of a light source and a high-resolution camera. The image processing device is primarily composed of high-performance computers and image processing software. First, the transmission device transports the hot-pressed LGP to the inspection station, and then, the image acquisition device captures the image of the hot-pressed LGP. Finally, the image processing device analyzes and detects the captured image.

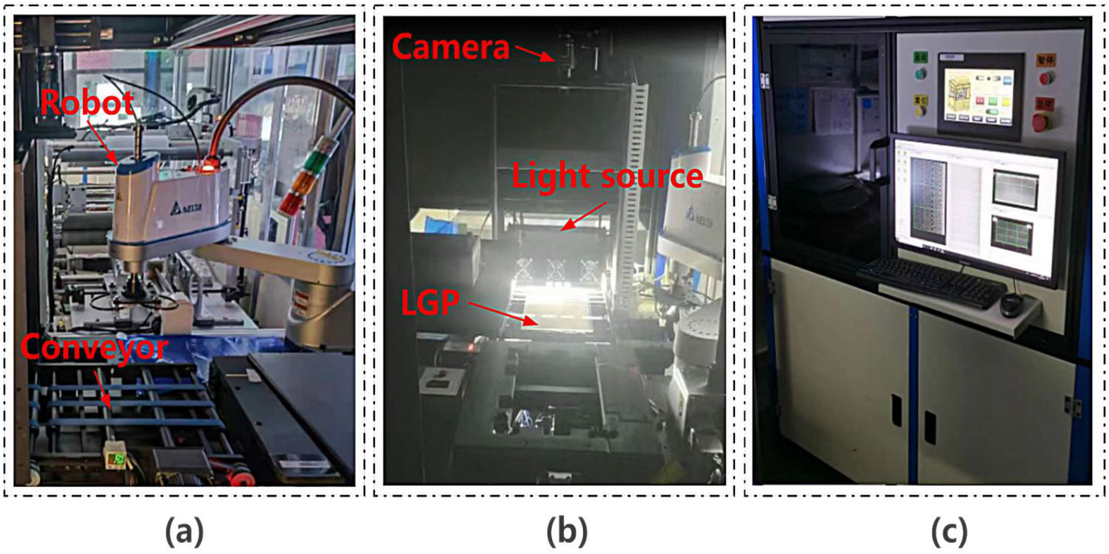

FIGURE 2. Hot-pressed LGP defect detection device. (a) Transmission device; (b) Image acquisition device; (c) Image processing device.

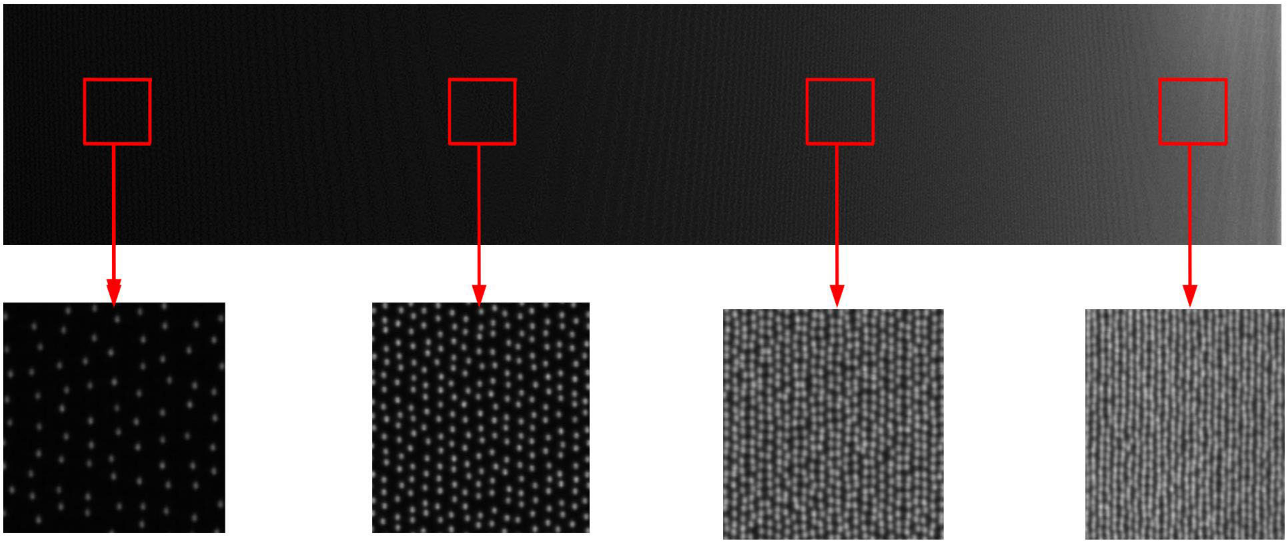

FIGURE 3. Hot-pressed LGP image.

The quality inspection accuracy of the hot-pressed LGPs is relatively high, and it is difficult for the industrial surface camera to meet such requirements. Thus, the inspection system in this study used a 16k line matrix camera to collect clear images of hot-pressed LGPs, as shown in FIGURE 3. The top image is a partial image intercepted from an original hot-pressed LGP image, and the bottom image is intercepted from the corresponding position on the top image. As shown in the four windows at the bottom, the LGP image has dense light guide points and a complex textured background, the light guide points become increasingly dense from left to right, and the image gradually blurs.

During production, according to the manufacturer's technical requirements and the imaging characteristics of the LGP, the defects of the Hot-pressed LGPs are divided into four categories: white dot defects, bright line defects, dark line defects and area defects, which are shown in FIGURE 4.

IV. METHODOLOGY

A. YOLOv7 NETWORK STRUCTURE

YOLOv7 [34], a detector using the YOLO architecture, is a single-stage target detection network with fast detection speed, high accuracy, and easy training and deployment. The entire network model structure of YOLOv7 can be divided

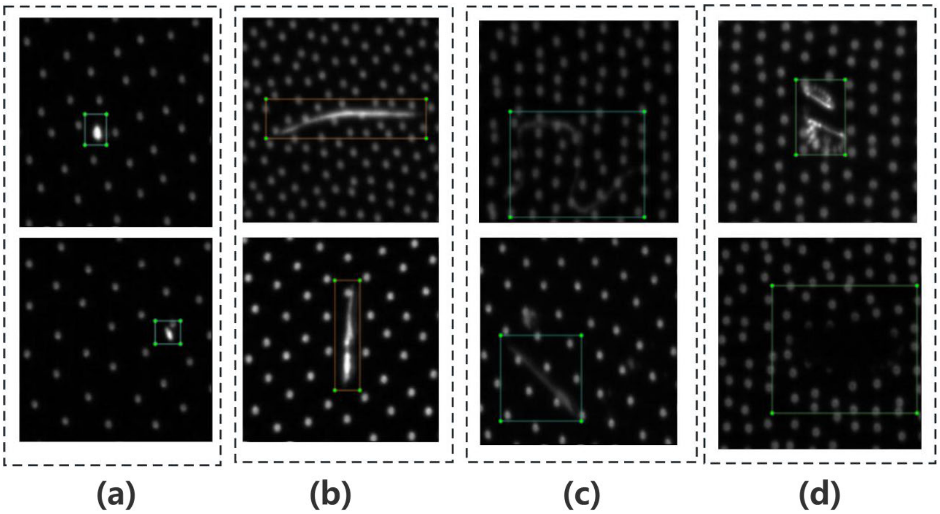

FIGURE 4. Types of defects in hot-pressed LGPs. (a) White dot defects; (b) Bright line defects; (c) Dark line defects; (d) Area defects.

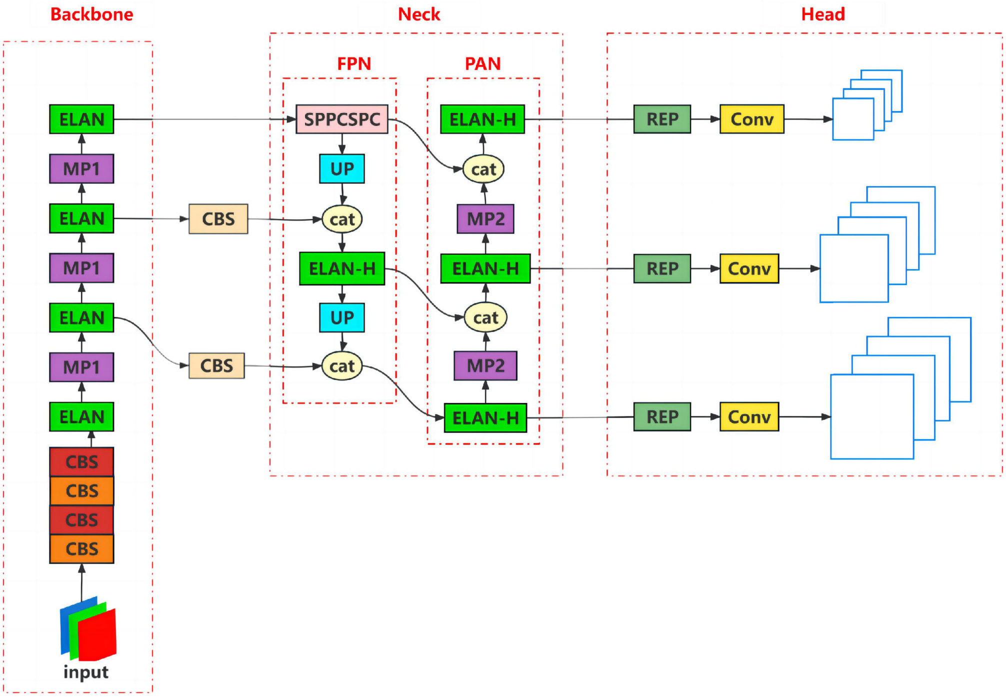

FIGURE 5. YOLOv7 network structure diagram.

into three parts: backbone, neck and head, and the network structure (see FIGURE 5). The backbone network extracts the multiscale features of the input image and outputs the multiscale feature map to the neck network as the input. The primary role of the neck network is feature fusion, where the FPN structure [35] transfers stronger semantic features in the deeper layers to the shallower layers, augmenting the entire pyramid and thus enhancing semantic representation at multiple scales. The PAN structure [36] of the neck network transmits stronger positional information from the shallow layers to the deep layers, enhancing localization at multiple scales. The FPN and the PAN enhance the expressive power of the network and assign the multiscale learning task to 3 different sized detection networks. The head integrates the new feature information and performs target detection and classification.

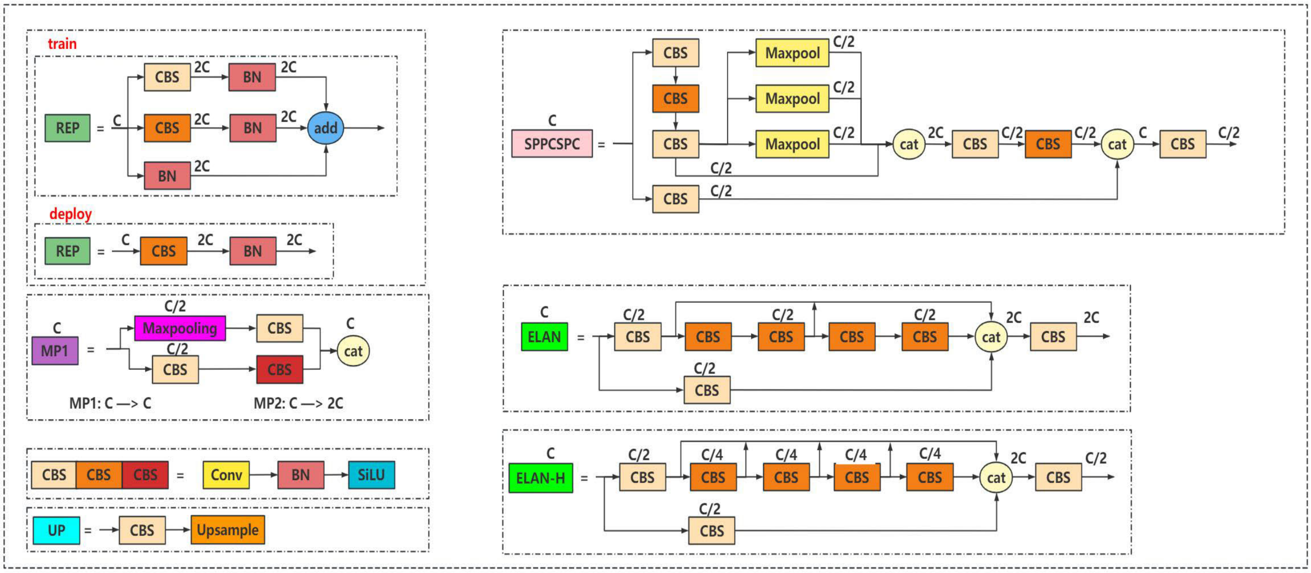

FIGURE 6. Structure diagram of each module.

YOLOv7 is composed of a CBS module, MP module, ELAN module, ELAN-H module, UPSample module, SPPCSPC module and RepConv module. The structure of each module is shown in FIGURE 6. The CBS module consists of regular convolution, batch normalization, and SiLU activation functions. The MP module is a downsampling module that consists of a maximum pooling layer and a CBS module. The ELAN module is an efficient aggregation network with two branches, which allows the network to learn more features with greater robustness by controlling the shortest and longest gradient paths. The role of the ELAN-H module is similar to that of the ELAN module. The difference is that ELAN stitches 4 convolutional layers for output and ELAN-H stitches 6 convolutional layers for output. The UPSample module is an upsampling module that uses the upsampling method of nearest neighbor interpolation. The SPPCSPC module can increase the perceptual field, allowing the algorithm to adapt to different resolution images and is obtained via maximum pooling for different perceptual fields. The RepConv module is somewhat different in training and deployment. The training has the summed output of three branches, and deployment reparameterizes the parameters of these three branches to the master branch.

B. GDA-YOLOv7 NETWORK STRUCTURE

The backbone network of YOLOv7 downsamples the input five times. The downsampling is done by ELAN and MP1 modules, and the backbone network may lose the semantic information of small target defects during the downsampling process. The original feature extraction module of the backbone network, ELAN, cannot effectively use the global contextual features in the image, and the feature extraction capability is weak, which may eventually result in missed detection of small target defects or low detection accuracy. For the aforementioned reasons, in the modified network GDA-YOLOv7 proposed in this paper, the second and third ELAN modules in the original YOLOv7 network backbone are replaced with Global Context Blocks (GCBlocks) [37]. This substitution aims to more effectively capture global contextual information in the image, establishing long-range dependencies among all feature pixels. The integration of GCBlocks enhances spatial correlation between the background and foreground, thereby improving the recognition of targets in complex backgrounds and the perceptual capability for small target defects.

The original neck network of YOLOv7 has problems such as low feature extraction ability and disappearance of back propagation gradients, which may cause the deeper layers in the neck network to learn extremely slowly or not at all, thus causing the neurons to enter a stagnant state and stop learning new things. This will cause the head network to receive no valid prediction information and the network to have poorer prediction results. Building upon the aforementioned rationale, the enhanced GDA-YOLOv7 network, as proposed in this study, replaces the first and second ELAN-H modules in the original YOLOv7 network neck with Densely Connected Convolutional Network (DenseNet) [38]. This substitution promotes a more unrestricted flow of information across different layers of the network, facilitating the capture of features at various levels. Consequently, this addresses the challenge of gradient vanishing and enhances the training of deeper network architectures.

Because the size of defect targets is uncertain, small-sized defects dilute their semantic information more quickly as the number of layers in the network increases, thus easily resulting in the loss of small-sized targets. Although the original

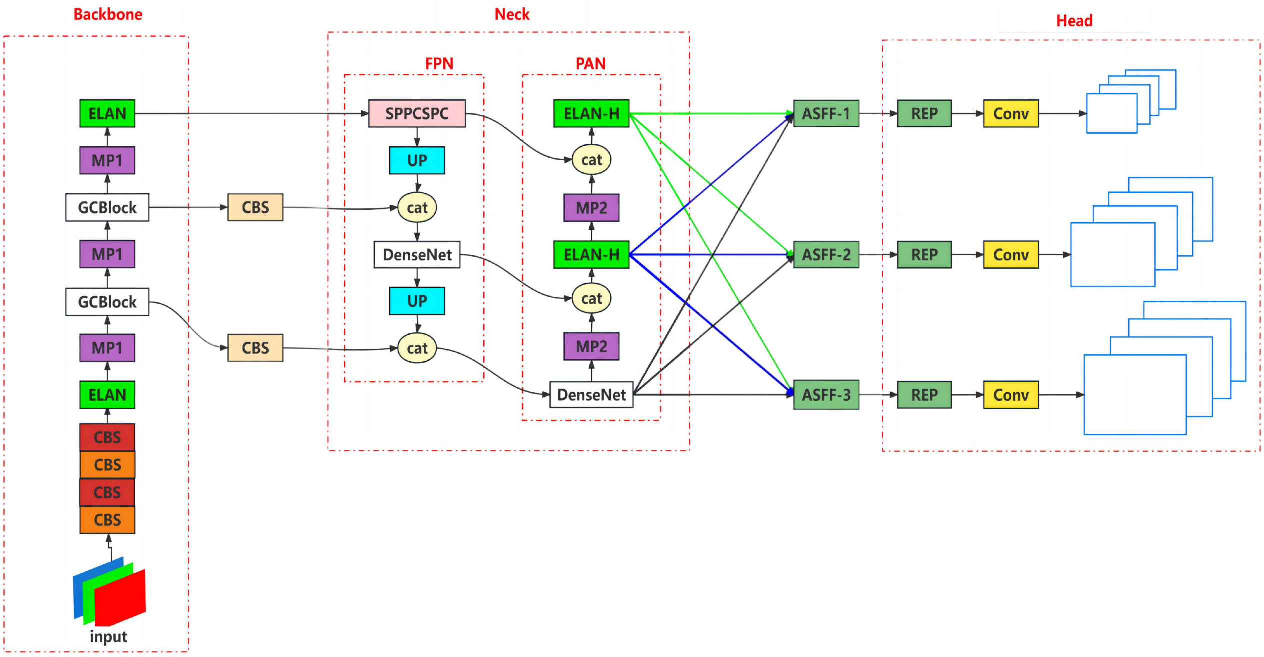

FIGURE 7. Improved YOLOv7 network structure diagram.

feature fusion method of YOLOv7 can enrich the overall semantic feature information, there are often prediction conflicts between different dimensions. For the aforementioned reasons, in the refined GDA-YOLOv7 network proposed in this study, the Adaptive Spatial Feature Fusion (ASFF) [39] is incorporated into the feature fusion structure of YOLOv7. This addition allows for compensating small targets that are challenging to extract from high dimensions in low dimensions, addressing the issue of potential loss of small targets. Moreover, it effectively resolves prediction conflicts across different dimensions, enhancing the adaptability of the network. This is particularly effective for identifying light guide plate defects with diverse and complex shapes. The improved YOLOv7 network is shown in FIGURE 7.

C. GLOBAL CONTEXT BLOCK(GCBLOCK)

Each pixel in an image is not isolated. One pixel has a certain relationship with the surrounding pixels, and the interconnection of a large number of pixels produces various objects in the image. In LGP detection, the contrast between some defects and the background is low, and fusing global contextual information [40] can help the network to enhance the spatial correlation between the background and defect targets, thus relying on this potential relational feature to allow the network to better detect defect targets. Thus, we add a global context block to the backbone network of YOLOv7 to obtain global context features in the image and let long-range dependencies be constructed among all feature pixels so that the network can focus on different regions and detect defective targets more effectively.

The global context block can assign different weights to the input elements from the spatial and channel dimensions to highlight useful information, and its structure is shown in FIGURE 8. The global context block is abstracted into three processes.

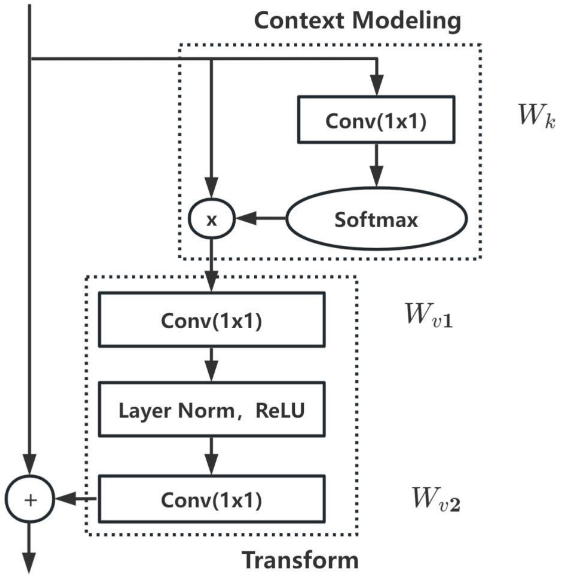

FIGURE 8. Structure of the global context block.

First, a 1 × 1 convolution $\mathrm{W_k}$ and softmax function are used to obtain spatial attention weights, and the weights are

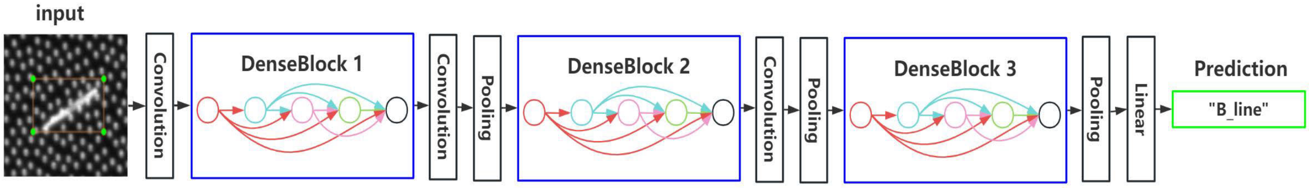

FIGURE 9. Structure of the densely connected convolutional network.

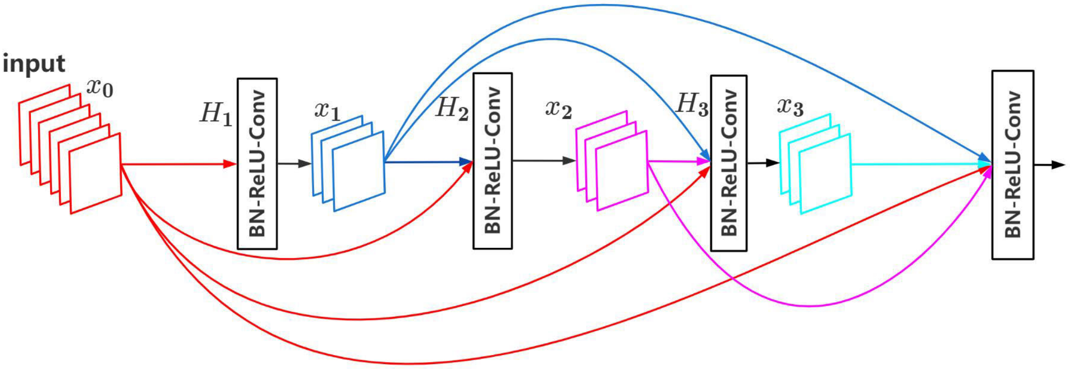

FIGURE 10. Structure diagram of DenseBlock.

then multiplied with the original graph. Thus, the features of all locations are aggregated together to form global context features, expressed as shown in Equation (1). Using the Softmax function ensures that the sum of probabilities for each prediction equals 1, thus ensuring that useful information remains within the appropriate range.

Second, feature transformation via 1 × 1 convolution W_{v1} is performed, and Layer Norm [41] and ReLU enhanced generalizations are used to capture the interdependencies on the channels, formulated as shown in Equation (2). Introducing the ReLU function between two convolutional layers enhances their nonlinearity and improves the model's representation capability. Due to the increased difficulty in optimization caused by the two-layer bottleneck transformation, Layer Norm is added to the bottleneck transformation, before the ReLU activation, to simplify the network and promote generalization.

Third, feature transformation via 1 × 1 convolution W_{v2} is performed, followed by feature aggregation (i.e., global contextual features are aggregated to features at each location using an additive method), as shown in Equation (3):

$$\alpha_i = \sum_{j=1}^{N_p} \frac{e^{W_k \times j}}{\sum_{m=1}^{N_p} e^{W_k \times m}} x_j \tag{1}$$

$$\delta(\cdot) = ReLU(LN(W_{v1}(\cdot))) \tag{2}$$

$$z_i = x_i + W_{v2}(ReLU(LN(W_{v1}(\sum_{j=1}^{N_p} \frac{e^{W_k \times j}}{\sum_{m=1}^{N_p} e^{W_k \times m}} x_j)))) \tag{3}$$

D. DENSELY CONNECTED CONVOLUTIONAL NETWORK(DENSENET)

In the YOLOv7 network, the neck network suffers from low feature extraction ability and vanishing back propagation gradients. Each layer in a densely connected convolutional network (DenseNet) can access the gradient from the loss function and the original input signal, thus prompting implicit deep supervision [42] and helping to train deeper network architectures. Thus, we add a densely connected convolutional network to the neck network of YOLOv7 to enhance the feature extraction ability and alleviate the gradient disappearance problem while ensuring the maximum information flow of the network.

The densely connected convolutional network (DenseNet) consists of the DenseBlock and the transition layer, as shown in FIGURE 9. DenseBlock is a unique module in the densely connected convolutional network, as shown in FIGURE 10. In the same DenseBlock, the width and height of the layer will not change, but the number of channels will change accordingly. Connecting all layers in DenseBlock directly to each other allows information to flow smoothly between layers in the network and improves feature propagation. This process also promotes feature reuse and fusion, enhances feature extraction, and solves problems such as gradient disappearance in deep neural networks. To preserve the feed-forward feature, the input of each layer in the network is the sum of the outputs of all the previous layers, and the output of each layer is also propagated backward and becomes part of the input of the later layer. Thus, the Lth layer in DenseBlock has L inputs, consisting of the feature maps of all the

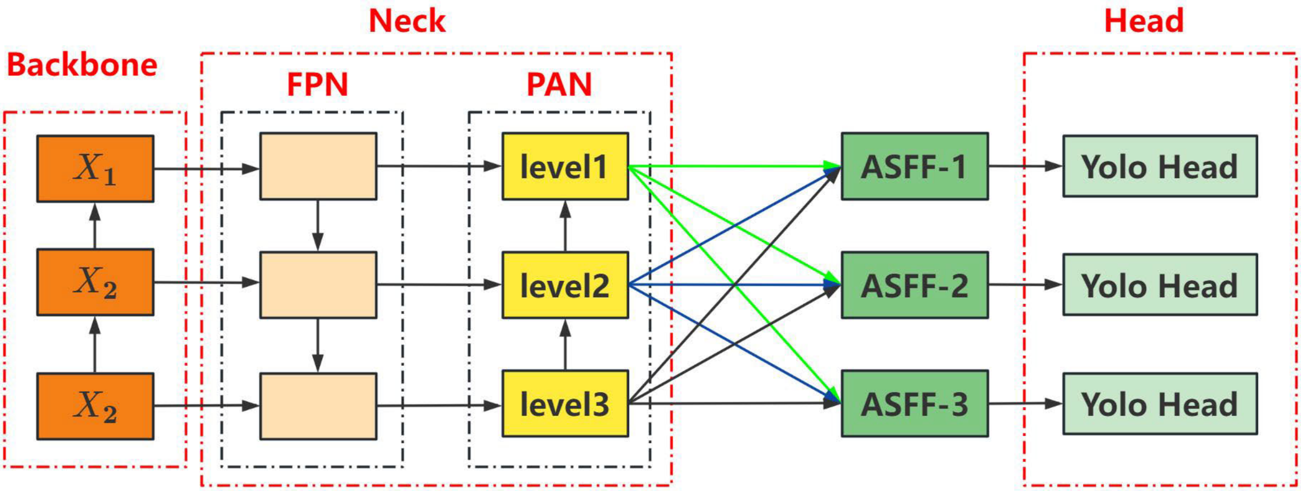

FIGURE 11. Adaptive spatial feature fusion structure diagram.

previous convolutional blocks, and its own feature maps are passed to all subsequent layers. Also, an L-layer network has L(L+1)/2 connections in an L-layer network instead of only L connections, as in the traditional structure.

In DenseBlock, layer l receives the feature maps of all previous layers, and x_l is defined as shown in Equation (4):

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{4}$$

where $[x_0, x_1, \ldots, x_{l-1}]$ are $0, 1, 2, \ldots, l-1$ layers of the cascade, and $H_l(\cdot)$ is a composite function of three consecutive operations: batch normalization (BN) [43], followed by a ReLU activation function [44] and a 3 × 3 convolution (Conv).

The transition layer primarily consists of convolution and pooling, which connects different DenseBlocks modules by adjusting the width and height of the preceding DenseBlocks and combines the characteristics of DenseBlocks. The DenseBlock modules are stacked so that the functions will stack continuously, which also makes the connection between the layers tighter.

E. ADAPTIVE SPATIAL FEATURE FUSION(ASFF)

During defect detection with LGPs, it is easy to lose small-size targets because the size of defect targets is uncertain, and small-size defects dilute their semantic information more quickly as the number of layers in the network increases. Although generic feature fusion methods can enrich the overall semantic feature information, there are often prediction conflicts between different dimensions. ASFF can resolve prediction conflicts in different dimensions and can compensate for small targets that are difficult to extract in high dimensions from low dimensions, solving the problem of detecting small targets that are easily lost. Thus, we use adaptive spatial feature fusion (ASFF) in the feature fusion structure of the model, which is shown in FIGURE 11.

In the neck structure of the YOLOv7 algorithm, FPN transmits the stronger semantic feature features from the deep layer to the shallow layer to enhance the entire pyramid, thus enhancing the semantic representation on multiple scales. PAN transmits the stronger location information from the shallow layer to the deep layer, enhancing the localization on multiple scales. The primary purpose of adding the adaptive spatial feature fusion module at the end of the PAN layer in front of the head layer is to ensure that the model can take full advantage of feature information at different scales. By adjusting the feature fusion and weight parameters of the PAN layer, the weight parameters are derived from the output of the convolutional feature layer, and the weight parameters become learnable after gradient back-propagation to be adaptive when performing weighted fusion, which effectively improves the feature extraction capability of the network and fully realizes the multiscale feature fusion of the model.

X_1, X_2 and X_3 are feature maps extracted from the YOLOv7 backbone network. Level 1, level 2 and level 3 are feature maps that can be obtained from the PAN structure. ASFF-1, ASFF-2 and ASFF-3 are the final fusion results obtained using the ASFF algorithm. The adaptive spatial feature fusion process with ASFF-3 as an example is as follows.

(1) For the level 1 feature map, we let level 1 obtain the same number of channels as the level 3 feature map by convolution. We then upsample the convolved level 1 feature map to keep the same size as level 3, and the result is $x^{1\rightarrow 3}$.

(2) For the level 2 feature map, we let level 2 obtain the same number of channels as the level 3 feature map by convolution. We then upsample the convolved level 2 feature map to keep the same size as level 3, and the result is $x^{2\rightarrow 3}$.

(3) The level 3 feature map is not adjusted but is renamed $x^{3\rightarrow 3}$.

(4) After processing the three feature maps using the softmax function, the weight coefficients α_{ij}^3, β_{ij}^3 and γ_{ij}^3 of $x^{1\rightarrow 3}$,

$x^{2\to 3}$ and $x^{3\to 3}$ are obtained, respectively. Then, y_{ij}^3 $(ASFF-3)$ is calculated using a weighted summation according to Equation (5) (i.e., y_{ij}^3 $(ASFF-3)$ is calculated by the Adaptive Spatial Feature Fusion (ASFF) algorithm to obtain the new feature map):

$$y_{ij}^3 = \alpha_{ij}^3 * x_{ij}^{1\to 3} + \beta_{ij}^3 * x_{ij}^{2\to 3} + \gamma_{ij}^3 * x_{ij}^{3\to 3} \tag{5}$$

The adaptive spatial feature fusion (ASFF) module is calculated as shown in Equation (6):

$$y_{ij}^l = \alpha_{ij}^l * x_{ij}^{1\to l} + \beta_{ij}^l * x_{ij}^{2\to l} + \gamma_{ij}^l * x_{ij}^{3\to l} \tag{6}$$

where y_{ij}^l $(ASFF-l)$ is the new feature map obtained using the ASFF module and is the valid feature for the target prediction. α_{ij}^l, β_{ij}^l and γ_{ij}^l are the weight coefficients of the three feature maps that are defined by the softmax function as shown in Equation (7). α_{ij}^l, β_{ij}^l and γ_{ij}^l are processed by the softmax [45] function to satisfy $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$, and $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0, 1]$.

The parameters $\lambda_{\alpha_{ij}}^l$, $\lambda_{\beta_{ij}}^l$ and $\lambda_{\gamma_{ij}}^l$ are the control parameters for Equation (7):

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \tag{7}$$

F. LOSS FUNCTION

The loss function is used to measure how far the predictions made by the model deviate from the true value and to quantify the deviation to guide the next training step in the right direction. The loss function of YOLOv7 can be divided into three parts: localization loss, confidence loss and classification loss. The total loss is the weighted sum of the three losses, as shown in Equation (8):

$$LOSS = W_1 \times L_{box} + W_2 \times L_{cls} + W_3 \times L_{obj} \tag{8}$$

The localization loss is used to measure the error between the predicted box and the calibrated box. Confidence loss is used to measure the probability of the presence of the target in the prediction box. The larger the confidence loss is, the smaller the probability of the presence of the target. The classification loss is used to measure the probability that the target in the prediction box belongs to a certain classification. The larger the classification loss is, the lower the probability that the target belongs to a certain classification.

1) LOCALIZATION LOSS

The localization loss is based on the CIOU loss, as shown in Equation (9):

$$CIOU\ loss = 1 - CIOU \tag{9}$$

The formula for CIOU is shown in Equation (10). The three terms of CIOU correspond exactly to the calculation of IOU, center point distance and aspect ratio:

$$CIOU = IOU - \left(\frac{\rho^2(b, b^{gt})}{c^2} + \alpha v\right) \tag{10}$$

where b is the parameter for the center coordinates of the prediction box. b^{gt} is the parameter for the center coordinates of the real box. ρ is the Euclidean distance between the prediction and real boxes. c is the diagonal length of the smallest outer rectangle that completely encloses the prediction and real boxes.

The full name of IOU is Intersection over Union (IOU), which is used to calculate the ratio between the intersection of the prediction box and the real box and the union, as shown in Equation (11),

$$IOU = \frac{A \cap B}{A \cup B} \tag{11}$$

v is used to measure the consistency of the scale between the prediction box and the real box, as shown in Equation (12). And, α is the parameter used to balance the scale, as shown in Equation (13):

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{12}$$

$$\alpha = \frac{v}{(1 - IOU) + v} \tag{13}$$

where w and h are the width and height of the predicted box. w^{gt} and h^{gt} are the width and height of the real box.

2) THE CONFIDENCE LOSS AND CLASSIFICATION LOSS

The loss of confidence and the loss of classification are shown in Equation (14) using the binary cross-entropy loss function:

$$L_{cls} = L_{obj} = -\frac{1}{n}\sum(y_n \times \ln x_n + (1 - y_n) \times \ln(1 - x_n)) \tag{14}$$

where n is the number of samples of the input. y_n is the target value and x_n is the predicted value of the network.

V. EXPERIMENTAL VERIFICATION

A. EXPERIMENTAL DATASET OF HOT-PRESSED LGPS

The dataset of the hot-pressed light guide was acquired on the factory production line by a high-precision line-scan camera, and the defective images were manually selected. Because the resolution of the original image is large and cannot be directly applied to train in the deep learning model, the defective part of the image is manually intercepted with a window of 416 × 416, and the dataset is divided into white point defects, bright line defects, dark line defects, and area defects. In this study, we make labels for these defective images and extend the dataset by other means such as panning and flipping. The size of the original dataset is 1500 and the size of the expanded dataset is 4127, and then the dataset is divided into training, validation and test sets in the ratio of 6:2:2, and the results are shown in TABLE 2:

B. EXPERIMENTAL SETUP

The hardware environment and software versions for the experiments are shown in TABLE 3.

Parameters for network training, as shown in TABLE 4.

TABLE 2. Defect dataset for hot-pressed LGPs.

	Training	Validation	Test	Total
white dot	628	211	207	1046
bright line	777	259	259	1295
dark line	580	193	194	967
area	491	163	165	819

TABLE 3. Hardware environment and software version.

Hardware and Software	Configuration
Hardware	Operating system：Ubuntu 18.04
	CPU：Intel(R) Xeon(R) Platinum8358P @2.6 GHz
	GPU：RTX A5000
Software	Python3.8+PyTorch1.8.1+Cuda11.1

TABLE 4. Network training parameters.

Training parameters	Value
Batch Size	64
epoch	300
Dynamic Parameters	0.937
Initial learning rate	0.01
Recurrent learning rate	0.1
Image size	416×416

C. MOSAIC DATA AUGMENTATION

Using Mosaic data augmentation can enrich the detection dataset using panning, scaling, rotation and changing the hue values. In particular, random scaling adds many small targets to make the network more robust and improve the discriminative power of the model on the test data.

The steps for implementing mosaic data augmentation include the following: (a). Read four random images at a time. (b). Flip, scale and change the color gamut of the four images separately, and place these four images in the top left, bottom left, top right and bottom right corners, respectively. (c). Stitch the design regions of the four images together using a matrix to create a new image. This new image also contains the bounding box information of the target, which has been processed during the stitching process.

D. PERFORMANCE INDICATORS

This study uses six primary metrics to test the performance of the model. Precision (P) describes the probability that the positive class classified by the classifier is indeed a positive class and is calculated as shown in Equation (15). Recall (R) describes the ability of the classifier to find all positive classes and is calculated as shown in Equation (16). Average precision (AP) is each class consists of precision (P) and recall (R), which takes the area of the P-R curve under different thresholds. The larger the value is, the better the class recognition accuracy is. The formula is shown in Equation (17). The mean average accuracy (mAP) is the average AP of all categories, and the relevant formula is Equation (18). The larger the value is, the better the model recognizes the higher the accuracy of the target. Frames per second (FPS) is the number of frames per second processed by the model and describes the speed of model inference. The larger the value is, the faster the model inference, and the better the model performance. Billions of floating point operations per second (GFLOPS) is the number of computations required by a model and is used to measure the complexity of the model:

$$P(Precision) = \frac{TP}{TP+FP} \tag{15}$$

$$R(Recall) = \frac{TP}{TP+FN} \tag{16}$$

$$AP = \int_0^1 P(R)dR \tag{17}$$

$$mAP = \frac{\sum_{n=0}^{c} AP(C)}{C} \tag{18}$$

where TP is a positive class judged as positive. FP is a negative class judged as positive. FN is a positive class judged as negative, and TN is a negative class judged as negative.

E. ABLATION EXPERIMENTS

In this paper, three improvements were made to YOLOv7. To verify the effectiveness of each improvement and the effectiveness of the combination of the three improvements, ablation experiments were conducted, and experimental results are shown in TABLE 5.

TABLE 5 shows that the mAP of YOLOv7 is 96.4%. mAP is improved by 2% by adding only the GCBlock module to the backbone of the YOLOv7 network because GCBlock allows the network to focus on different regions and detect defective targets more effectively by constructing long-range dependencies between all feature pixels for the feature map itself.

The YOLOv7 network only adds the DenseNet module in Neck, and the mAP is improved by 2.1% because each layer of DenseNet accesses the gradient directly from the loss function and the original input signal, ensuring the maximum information flow of the network while enhancing the feature extraction ability, alleviating the gradient disappearance problem and helping to train a deeper network architecture.

The YOLOv7 network only uses the ASFF module in the feature fusion structure, and the mAP is improved by 2.1%.

TABLE 5. Results of ablation experiments on the hot-pressed LGP dataset.

Method	GCBlock	DenseNet	ASFF	GFLOPS	Params	mAP	FPS
YOLOv7				105.2	37.213M	96.4%	220
YOLOv7+GCBlock	✓			108.8	38.858M	98.4%	161
YOLOv7+DenseNet		✓		109.9	37.420M	98.5%	152
YOLOv7+ASFF			✓	138.6	58.872M	98.5%	130
GDA-YOLOv7	✓	✓	✓	147.0	60.067M	99.1%	127

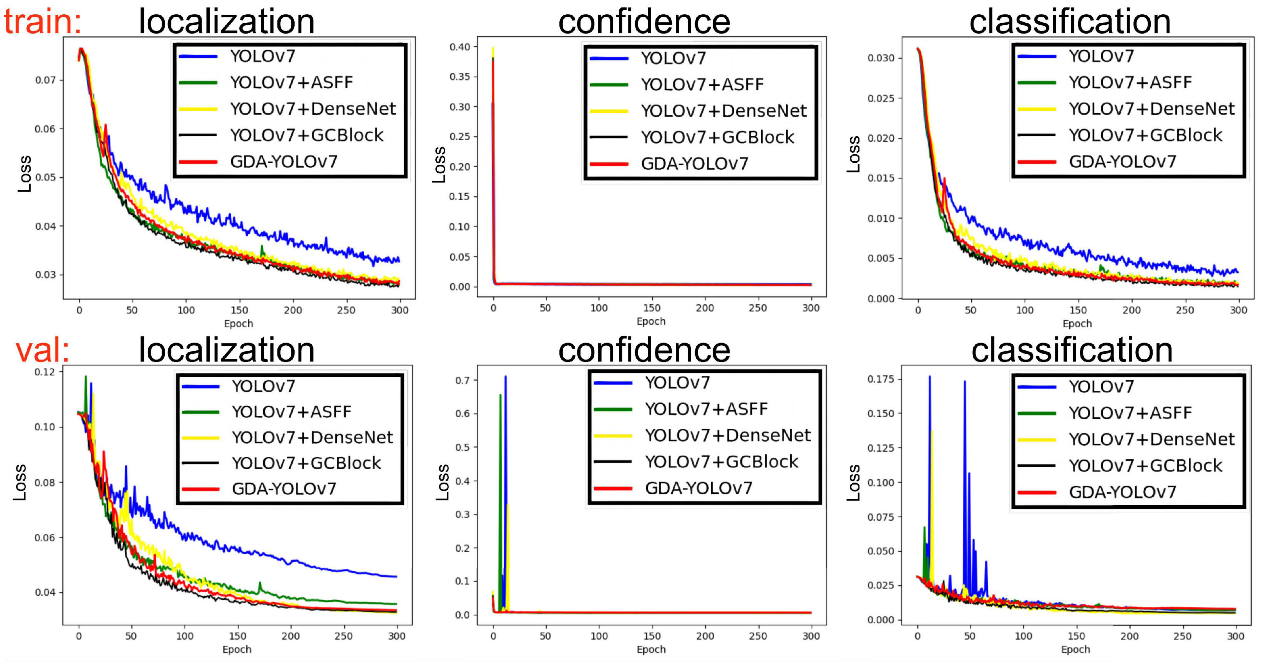

FIGURE 12. Training loss and validation loss. (a) Localization loss; (b) confidence loss; (c) classification loss.

Because ASFF improves the feature extraction ability of the network by adjusting the feature fusion and weight parameters of the PAN layer, the weight parameters are derived from the output of the convolutional feature layer, and the weight parameters become learnable after gradient back-propagation to be adaptive when performing weighted fusion, which effectively improves the feature extraction ability of the network and fully realizes the model multiscale feature fusion. This process can still compensate for small targets that are difficult to extract in high dimensions from low dimensions, solving the problem of detecting small targets that are easily lost.

As can be seen from FIGURE 12, the localization loss and classification loss functions of the GDA-YOLOv7 network converge quickly within the first 100 training times during training and validation, and converge when the number of training times reaches 300. The confidence loss function of the GDA-YOLOv7 network converges in the first few epochs, indicating that the improved network will hardly miss any detection during training and validation.

The results of the loss functions of the YOLOv7 network with various improvement modules are significantly better than those of the original YOLOv7 network. During validation, the confidence loss function and classification loss function of YOLOv7, YOLOv7+DenseNet, and YOLOv7+ASFF had several significant mutations in the first 80 epochs, that is, some data in the validation set were validated incorrectly, but YOLOv7+GCBlock did not have these problems. Obviously, our GDA-YOLOv7 network has absorbed the advantages of the YOLOv7+GCBlock network, making the results of the loss function very reasonable, without mutations during validation, and the network converges quickly.

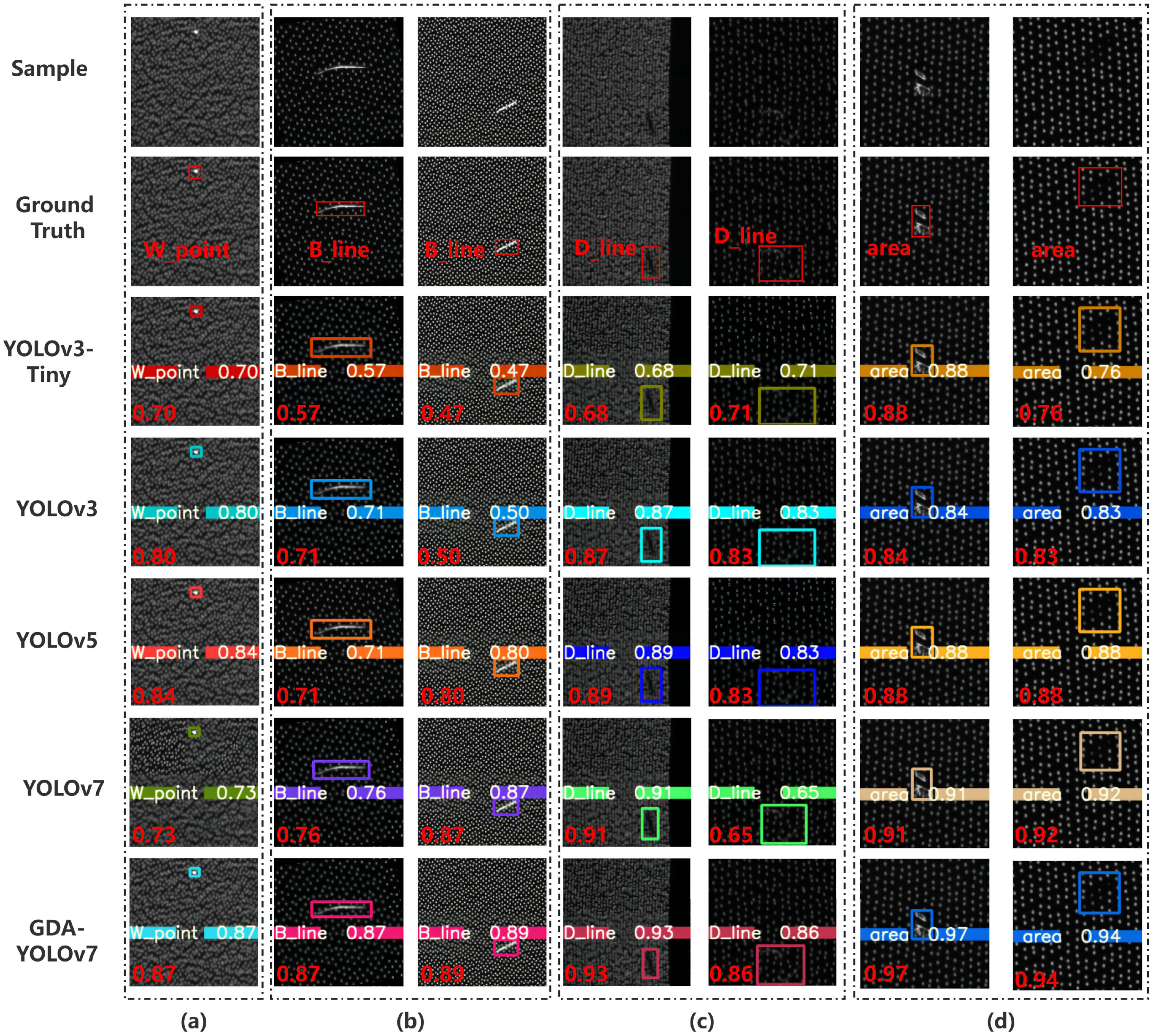

FIGURE 13. Detection results of the hot-pressed LGP dataset in different networks. (a) White dot defects; (b) Bright line defects; (c) Dark line defects; (d) Area defects.

F. HOT-PRESSED LGP COMPARISON EXPERIMENT

The improved YOLOv7 model was compared with the YOLOv3-Tiny, YOLOv3, YOLOv5 and YOLOv7 networks to verify its accuracy and validity. These models were all built based on the same parameters as the improved YOLOv7 model and used the original dataset of hot-pressed LGPs in the model. The comparison of mAP and FPS of different networks is shown in TABLE 6. The detection results of the hot-pressed LGPs dataset in different networks are shown in FIGURE 13.

In this study, seven images were randomly selected for testing on each model, and the results are shown in FIGURE 13. The detection accuracy of the improved YOLOv7 network is much higher than that of the other networks and meets the requirement of high-accuracy detection.

As shown in TABLE 6, Comparison of the hot-pressed LGPs dataset on the correlation network shows that the total mAP of the improved YOLOv7 model is 10.4%, 7.7%, 1.9% and 2.7% higher than the YOLOv3-Tiny, YOLOv3, YOLOv5 and YOLOv7 models, respectively. The improved YOLOv7 model shows a 2.1% improvement in mAP for white point defects, 6.2% improvement in mAP for dark line defects, 1.3% improvement in mAP for bright line defects and 1.1% improvement in mAP for area defects compared to YOLOv7. The FPS of the improved YOLOv7 model is 127. Thus, the

TABLE 6. Comparison results of hot-pressed LGPs dataset on related networks.

Method	AP				GFLOPS	Params	mAP	FPS
	White dot	Dark line	Bright line	Area				
YOLOv3-Tiny	86.4%	82%	91.8%	94.4%	13.0	8.677M	88.7%	314
YOLOv3	95.9%	87.2%	88.7%	94%	155.3	61.540M	91.4%	263
YOLOv5	96.1%	95.5%	97.8%	99.4%	48.3	20.883M	97.2%	303
YOLOv7	97.3%	91.9%	98.1%	98.4%	105.2	37.213M	96.4%	220
GDA-YOLOv7	99.4%	98.1%	99.4%	99.5%	147	60.067M	99.1%	127

TABLE 7. Results of comparison of NEU-DET dataset on related networks.

Methods	AP						mAP	FPS
	crazing	inclusion	patches	pitted	rolled	scratches		
YOLOv3-Tiny	40.9%	78.3%	88.2%	75.2%	64.8%	84.1%	71.9%	212
YOLOv3	45.9%	81.2%	86.1%	79.7%	65.1%	87.6%	74.3%	158
YOLOv5	47.3%	85.2%	88.3%	82.1%	68.6%	90.1%	76.9%	203
YOLOv7	45.2%	88.3%	87.4%	83.2%	66.5%	92.3%	77.2%	170
GDA-YOLOv7	49.8%	90.5%	89.2%	83.6%	70.4%	93.2%	79.5%	152

improved YOLOv7 model can meet the demand for real-time high accuracy.

G. EXPERIMENTAL RESULTS OF NEU-DET DATASET

To further substantiate the efficacy of the proposed model in this paper, comparative experiments were conducted using the NEU-DET dataset. The methodology for these comparative experiments mirrors that of the experimental approach applied to the dataset of surface defects in hot-pressed light guide plates. The NEU-DET dataset consists of images of defective hot-rolled steel strips. The sample images in the dataset have a resolution of 200 × 200, totaling 1800 images with defects. The defect types include Crazing, Inclusion, Patches, Pitted, Rolled, and Scratches. In this study, labels were created for these 1800 defective images, and the dataset was partitioned into training, validation, and test sets in a 6:2:2 ratio. TABLE 7 presents the comparative results of detecting surface defects in hot-rolled steel strips using the GDA-YOLOv7 network, YOLOv3-tiny, YOLOv3, YOLOv5, and YOLOv7. Six images were randomly selected for testing each model, and the detection results are illustrated in FIGURE 14. The results in FIGURE 14 indicate that the detection accuracy of the GDA-YOLOv7 method surpasses that of the other models. As shown in TABLE 7, GDA-YOLOv7 not only exhibits superior detection accuracy but also operates at a speed of 152 FPS, meeting the requirements for industrial site detection.

VI. DISCUSSION

For this study, GDA-YOLOv7 has the following advantages: The introduction of the GCBlock module increases the network's awareness of contextual semantic information, which improves the accuracy of detecting defects. The introduction of the DenseNet module enhances the feature extraction capability and training of deeper models while ensuring maximum

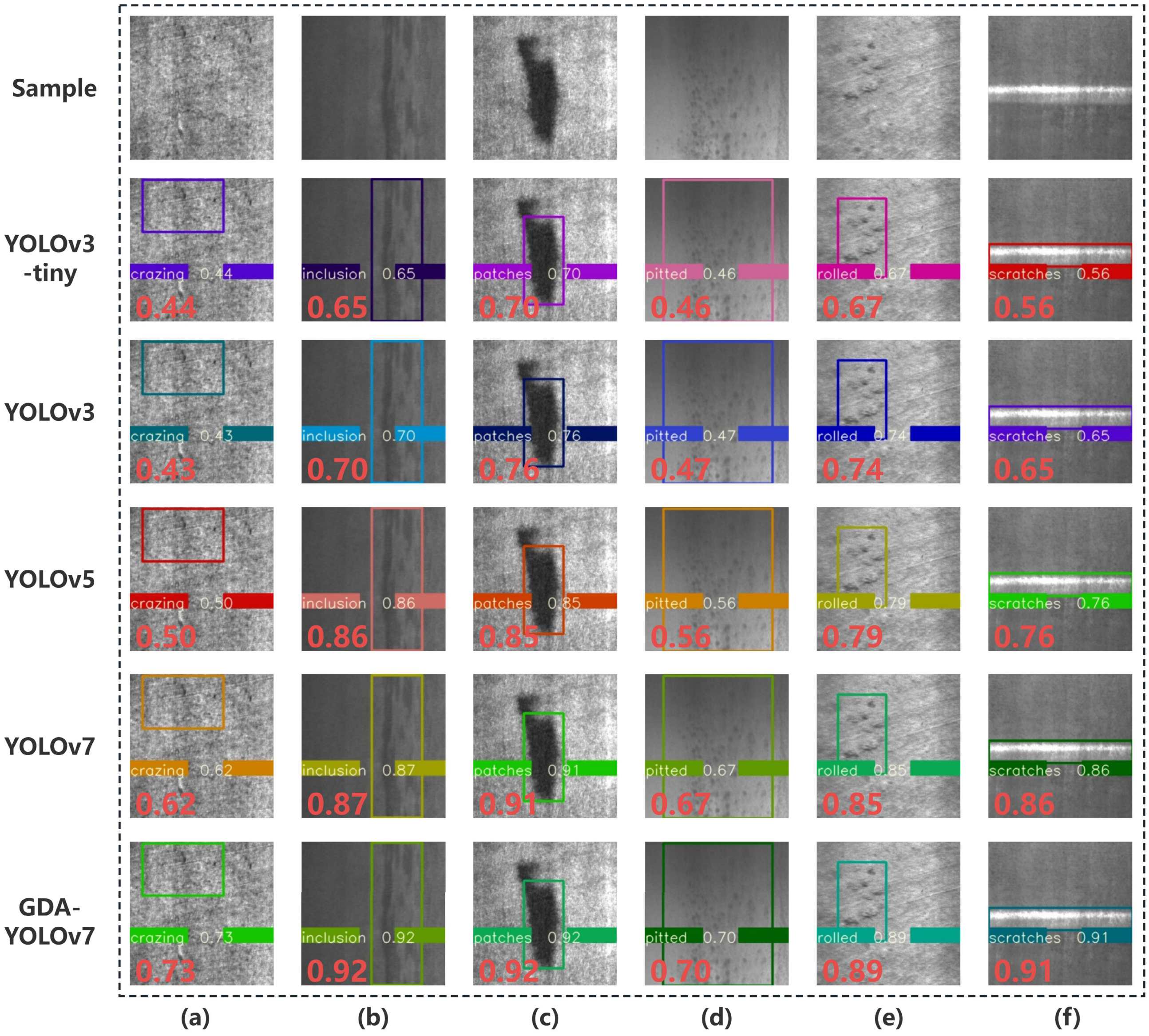

FIGURE 14. Detection results of NEU-DET dataset in different networks. (a) crazing defects; (b) inclusion defects; (c) patches defects; (d) pitted defects; (e) rolled defects; (f) scratches defects.

information flow. The application of ASFF module realizes multi-scale feature fusion, which makes the model more reliable in recognizing small targets.

The experiments show that compared with YOLOv3-tiny, the overall mAP of the GDA-YOLOv7 model has increased by 10.4%, and the mAP of white dot defects and dark line defects has increased the most significantly, at 13% and 16.1%, respectively. The mAP of bright line defects increased by 7.6%, and the mAP of regional defects increased by 5.1%. Compared with YOLOv3, the overall mAP increased by 7.7%, and the mAP of dark line defects and bright line defects increased the most significantly, at 10.9% and 10.7%, respectively. The mAP of white dot defects increased by 3.5%, and the mAP of regional defects increased by 5.5%. Compared with YOLOv5, the overall mAP increased by 1.9%, and the mAP of white dot defects and dark line defects increased the most significantly, at 3.3% and 2.6%, respectively. The mAP of bright line defects increased by 1.6%, and the mAP of regional defects increased by 0.1%. Compared with YOLOv7, the overall mAP increased by 2.7%, and the mAP of white dot defects and dark line defects increased the most significantly, at 2.1% and 6.2%, respectively. The mAP of bright line defects increased by 1.3%, and the mAP of regional defects increased by 1.1%. Therefore, the GDA-YOLOv7 model has shown significant improvements compared to other networks and is very suitable for high-precision detection tasks.

VII. CONCLUSION

In this study, an improved YOLOv7 hot-press guide plate defect detection method is proposed to accurately identify defects in hot-pressed guide plates in complex backgrounds. Incorporating the GCBlock module into the backbone network of YOLOv7 facilitates the transfer of additional contextual semantic information to the Neck layer. This enables the network to focus on different regions, enhancing the recognition of targets in complex backgrounds and improving the perceptual capability for small target defects. The DenseNet module is introduced in the neck to ensure maximum information flow of the network while enhancing feature extraction, alleviating gradient disappearance, and helping to train a deeper network architecture. Utilizing the ASFF module in the feature fusion architecture enables comprehensive multi-scale feature integration, addressing the challenge of potential small target loss. This enhances the recognition capability for light guide plate defects with intricate and variable appearances. Experimental results show that the GDA-YOLOv7 model improves the mAP by 2.7% compared with YOLOv7 and the detection speed of the model is 127 fps, which can meet the demand of real-time high accuracy.

The methodology of this study still has some limitations. Our main focus is on the accuracy and efficiency of defect detection to meet the demand for high efficiency in industrial production. However, future work can further improve the real-time and reliability of the algorithm to ensure real-time application in real industrial production environments. In addition, attention should also be paid to improving the robustness and security of the algorithms so that they can cope with various anomalies and potential attacks.

DATA AVAILABILITY

The dataset and code used in our research have been shared at the following link: https://www.kaggle.com/datasets/zhenyuli123/code-and-dataset

REFERENCES

[1] S. L. Lou, J. C. Ren, Y. L. Han, X. H. Yuan, and X. D. Zhou, "The preprocessing for infrared sea-surface target image," *Adv. Mater. Res.*, vol. 433, pp. 4512–4515, 2012.

[2] W. Wang, Z. Qin, S. Rong, and R. Y. Song, "A kind of method for selection of optimum threshold for segmentation of digital color plane image," in *Proc. 9th Int. Conf. Comput.-Aided Ind. Design Conceptual Design*, Nov. 2008, pp. 959–961.

[3] M. Zhang, J. Ma, M. Gong, H. Li, and J. Liu, "Memetic algorithm based feature selection for hyperspectral images classification," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jun. 2017, pp. 495–502.

[4] W. Lu, Z. Zhou, X. Ruan, Z. Yan, and G. Cui, "Insulator detection method based on improved faster R-CNN with aerial images," in *Proc. 2nd Int. Symp. Comput. Eng. Intell. Commun. (ISCEIC)*, Aug. 2021, pp. 417–420.

[5] C. Liu, Y. Wu, J. Liu, and J. Han, "MTI-YOLO: A light-weight and real-time deep neural network for insulator detection in complex aerial images," *Energies*, vol. 14, no. 5, p. 1426, Mar. 2021.

[6] X. Li, H. Su, and G. Liu, "Insulator defect recognition based on global detection and local segmentation," *IEEE Access*, vol. 8, pp. 59934–59946, 2020.

[7] H. S. Gill, G. Murugesan, B. S. Khehra, G. S. Sajja, G. Gupta, and A. Bhatt, "Fruit recognition from images using deep learning applications," *Multimedia Tools Appl.*, vol. 81, no. 23, pp. 33269–33290, Sep. 2022.

[8] Y. Yu, K. Zhang, L. Yang, and D. Zhang, "Fruit detection for strawberry harvesting robot in non-structural environment based on mask-RCNN," *Comput. Electron. Agricult.*, vol. 163, Aug. 2019, Art. no. 104846.

[9] Y. Tian, G. Yang, Z. Wang, E. Li, and Z. Liang, "Instance segmentation of apple flowers using the improved mask R–CNN model," *Biosyst. Eng.*, vol. 193, pp. 264–278, May 2020.

[10] X. Li, S. Lai, and X. Qian, "DBCFace: Towards pure convolutional neural network face detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1792–1804, Apr. 2022.

[11] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, "Dual-path deep fusion network for face image hallucination," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 1, pp. 378–391, Jan. 2022.

[12] S. Prasad, Y. Li, D. Lin, and D. Sheng, "maskedFaceNet: A progressive semi-supervised masked face detector," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3388–3397.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[15] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[17] Z. Chen, R. Wu, Y. Lin, C. Li, S. Chen, Z. Yuan, S. Chen, and X. Zou, "Plant disease recognition model based on improved YOLOv5," *Agronomy*, vol. 12, no. 2, p. 365, Jan. 2022.

[18] Y. Cai, T. Luan, H. Gao, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "YOLOv4-5D: An effective and efficient object detector for autonomous driving," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[19] K. Jiang, T. Xie, R. Yan, X. Wen, D. Li, H. Jiang, N. Jiang, L. Feng, X. Duan, and J. Wang, "An attention mechanism-improved YOLOv7 object detection algorithm for hemp duck count estimation," *Agriculture*, vol. 12, no. 10, p. 1659, Oct. 2022.

[20] Z. Su, K. Han, W. Song, and K. Ning, "Railway fastener defect detection based on improved YOLOv5 algorithm," in *Proc. IEEE 6th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Oct. 2022, pp. 1923–1927.

[21] Z. Yang, Z. Xu, and Y. Wang, "Bidirection-fusion-YOLOv3: An improved method for insulator defect detection using UAV image," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–8, 2022.

[22] Y. Lu, C. Sun, X. Li, and L. Cheng, "Defect detection of integrated circuit based on YOLOv5," in *Proc. IEEE 2nd Int. Conf. Comput. Commun. Artif. Intell. (CCAI)*, May 2022, pp. 165–170.

[23] Y. Chen, H. Xu, X. Zhang, P. Gao, Z. Xu, and X. Huang, "An object detection method for bayberry trees based on an improved YOLO algorithm," *Int. J. Digit. Earth*, vol. 16, no. 1, pp. 781–805, Oct. 2023.

[24] L. Xu, S. Dong, H. Wei, Q. Ren, J. Huang, and J. Liu, "Defect signal intelligent recognition of weld radiographs based on YOLO V5-IMPROVEMENT," *J. Manuf. Processes*, vol. 99, pp. 373–381, Aug. 2023.

[25] Y. Li, H. Ding, P. Hu, Z. Yang, and G. Wang, "Real-time detection algorithm for non-motorized vehicles based on D-YOLO model," *Multimedia Tools Appl.*, pp. 1–24, 2023.

[26] H. Wu, Y. Wang, P. Zhao, and M. Qian, "Small-target weed-detection model based on YOLO-V4 with improved backbone and neck structures," *Precis. Agricult.*, vol. 24, no. 6, pp. 2149–2170, Dec. 2023.

[27] W. Ming, F. Shen, H. Zhang, X. Li, J. Ma, J. Du, and Y. Lu, "Defect detection of LGP based on combined classifier with dynamic weights," *Measurement*, vol. 143, pp. 211–225, Sep. 2019.

[28] Y. Li and J. Li, "An end-to-end defect detection method for mobile phone light guide plate via multitask learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[29] Z. Li, J. Li, and W. Dai, "A two-stage multiscale residual attention network for light guide plate defect detection," *IEEE Access*, vol. 9, pp. 2780–2792, 2021.

[30] L. Hong, X. Wu, D. Zhou, and F. Liu, "Effective defect detection method based on bilinear texture features for LGPs," *IEEE Access*, vol. 9, pp. 147958–147966, 2021.

[31] J. Yao and J. Li, "AYOLOv3-tiny: An improved convolutional neural network architecture for real-time defect detection of PAD light guide plates," *Comput. Ind.*, vol. 136, Apr. 2022, Art. no. 103588.

[32] J. Li and Y. Yang, "HM-YOLOv5: A fast and accurate network for defect detection of hot-pressed light guide plates," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105529.

[33] J. Li and H. Wang, "Surface defect detection of vehicle light guide plates based on an improved RetinaNet," *Meas. Sci. Technol.*, vol. 33, no. 4, Apr. 2022, Art. no. 045401.

[34] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[35] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[36] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1971–1980.

[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[39] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.

[40] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Scene parsing with global context embedding," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2650–2658.

[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[42] C.-Y. Lee et al., "Deeply-supervised nets," in *Proc. Artif. Intell. Statist.*, 2015, pp. 562–570.

[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015.

[44] B. Hanin, "Universal function approximation by deep neural nets with bounded width and ReLU activations," *Mathematics*, vol. 7, no. 10, p. 992, Oct. 2019.

[45] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 1989.

ZHENYU LI was born in Inner Mongolia, China, in 1996. He is currently pursuing the M.S. degree in control engineering with Zhejiang Sci-Tech University. Since 2021, he has been learning with the School of Information Science and Engineering. His research interests include deep learning, computer vision, industrial defect detection, and image processing.

JUNFENG LI received the B.S. degree in electrical engineering and automation from Zhengzhou University, in 2002, the M.S. degree in mechanical design and theory from Zhejiang Sci-Tech University, and the Ph.D. degree in mechanical design and theory from Donghua University. He is currently an Associate Professor at the School of Information Science and Engineering, Zhejiang Sci-Tech University. His research interests include medical image fusion, image quality evaluation, human behavior recognition, pattern recognition, and industrial defect detection.