Optimization of Robot Vision Object Detection and Recognition Algorithm Based on Convolutional Neural Network

Xin Zhong
[1]Sussex Artificial Intelligence Institute, Zhejiang Gongshang University, Hangzhou 310018, Zhejiang, China
Corresponding author: Xin Zhong (e-mail: 13758116770@163.com)

ABSTRACT Accurate object detection and recognition remain fundamental challenges in robot vision systems operating in complex environments. To improve detection accuracy, robustness, and computational efficiency, this study proposes a multi-stage collaborative optimization framework based on convolutional neural networks. A lightweight backbone architecture combining depthwise separable convolution and channel attention mechanisms is first designed to reduce computational complexity while preserving semantic representation capability. An adaptive feature pyramid network is then developed to enhance multi-scale feature fusion and improve small-target recognition performance. Furthermore, a GIoU-based localization optimization strategy and a joint knowledge-distillation–pruning framework are introduced to improve localization accuracy and model efficiency simultaneously. Experimental evaluations on public and self-constructed datasets demonstrate superior performance in terms of detection accuracy, real-time processing capability, and parameter reduction. The proposed framework provides an efficient solution for intelligent robotic perception and offers potential applications in machine vision, electromagnetic imaging interpretation, and autonomous sensing systems.

INDEX TERMS robot vision, object detection, convolutional neural networks, autonomous sensing, electromagnetic imaging

I. INTRODUCTION

As robots increasingly enter domestic and professional environments, endowing them with accurate environmental perception capabilities becomes the foundation for autonomous grasping and interaction with diverse materials and surfaces. This precise recognition of textures and patterns is essential for navigating spaces characterized by complex soft materials and ensuring the delicate handling of home materials [1-2]. Meanwhile, visual target detection and recognition as an important part of perception system directly affects the success rate and reliability of robot grasping [3]. However, indoor scenes are extremely complex, and sudden changes in lighting, mutual occlusion of objects, large variation of target scale and stacking often appear in scenes, which brings a huge challenge to detection algorithms. Although traditional detection models based on convolutional neural network can achieve good performance on general datasets, they also meet two bottlenecks when they are applied to robot embedded platform: first, the model parameter amount is large and computational complexity is high, which cannot meet real-time processing requirements; second, false negative rate of small-scale target and occluded target is relatively high, and positioning accuracy is insufficient. How to realize lightweighting and acceleration of the model while keeping the accuracy of the model is a key problem that needs to be solved urgently in robot vision.

In response to the above challenges, scholars at home and abroad have conducted extensive research. For example, Aboyomi and Daniel conducted a systematic comparative analysis of three representative target detection algorithms, providing a reference for selecting appropriate methods in applications such as autonomous driving, monitoring and image recognition [4]. To address the issues of low timeliness and high energy consumption in deep learning target detection in intelligent transportation, Hawlader F et al. proposed “Edge-Optimized Detection” [5], a target detection system based on edge cloud collaboration and reconstructed convolutional neural networks. Zhao and Wang proposed an improved lightweight network that further reduced computational overhead while maintaining high accuracy by optimizing the bottleneck structure and activation function [6]. Shan et al. studied the interaction between symbolic service remedies and robot social perception on consumer service evaluation in the context of service robot service failure, and further revealed its mechanism and boundary conditions [7]. Zou et al. systematically reviewed target detection algorithms over the past two decades, comprehensively outlining the evolution from traditional methods

to the deep learning era [8]. However, existing methods still have limitations: lightweight networks often sacrifice accuracy, feature pyramids do not fully integrate multi-scale information, and general model compression techniques lack targeted optimization for robot vision tasks, resulting in the need to improve detection robustness in complex scenarios.

To address these issues, Liu et al. proposed a multi-stage feature redistribution algorithm for small sample target detection, which significantly improves the target domain category detection performance while maintaining source domain category accuracy as much as possible [9]. To address the issues of insufficient information and limited feature capture in few-sample target detection, Tang et al. combined attention distillation and multi-scale attention modules within the meta-learning framework, making full use of query sets and support sets to enhance global feature representation [10]. These studies show that a single optimization method is difficult to simultaneously meet the requirements of accuracy, speed, and lightweight design. Therefore, this paper proposes a multi-stage collaborative optimization method for robot platforms, systematically solving the above problems through the organic combination of a lightweight backbone network, adaptive feature pyramid, improved loss function, and knowledge distillation pruning.

This paper focuses on designing a lightweight and high-precision target detection and recognition algorithm for robot vision to identify defects and material irregularities in high-speed manufacturing environments. By optimizing the detection of material breakages and production flaws, the proposed method ensures the structural integrity of advanced functional materials and technical materials during automated inspection. Firstly, we construct a lightweight backbone network with depthwise separable convolution and channel attention to improve semantic feature expression and lighten computational cost. Secondly, we design an adaptive feature pyramid fusion structure to further improve multi-scale target detection ability by cascading in cross layer and path enhancement. Thirdly, we use GIOU loss function to optimize the accuracy of target localization. Finally, we compress and speed up the model by knowledge distillation and channel pruning method.

II. LIGHTWEIGHT BACKBONE FEATURE EXTRACTION NETWORK

While traditional backbone networks such as VGG and ResNet perform excellently in image classification tasks, their large number of parameters and computational overhead make it difficult to meet the real-time requirements of robotic embedded platforms. To address this issue, this paper constructs a lightweight feature extraction backbone network. The specific structural parameters are shown in Table 1.

TABLE 1. Lightweight Backbone Network Structural Parameters

Stage	Input Size	Operation Type	Number of Kernels	Stride	Repetitions	Output Size
1	320×320	Conv2d	32	2	1	160×160
2	160×160	Bottleneck	16	1	2	160×160
3	160×160	Bottleneck	24	2	3	80×80
4	80×80	Bottleneck	32	2	4	40×40
5	40×40	Bottleneck	64	2	3	20×20
6	20×20	Bottleneck	96	1	2	20×20
7	20×20	Bottleneck	160	2	2	10×10

Specifically, depthwise separable convolution is used instead of standard convolution as the basic operator. Depthwise separable convolution decomposes standard convolution into two stages: depthwise convolution and pointwise convolution. Depthwise convolution performs convolution operations independently on each input channel, capturing features in the spatial dimension; pointwise convolution, on the other hand, linearly combines the outputs of the depthwise convolution through 1×1 convolutions, achieving cross-channel information fusion. Assuming the input feature map size is $D_F \times D_F \times M$, the convolution kernel size is $D_K \times D_K$, and the number of output channels is N, the computational cost of standard convolution is $D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F$, while the computational cost of depthwise separable convolution is $D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F$. Taking a 3×3 convolution as an example, depthwise separable convolution can compress the computational cost to approximately 1/8 to 1/9 of that of standard convolution, significantly reducing model complexity.

To compensate for the reduced representational capacity inherent in lightweight architectures further, this paper introduces channel attention into the depthwise separable convolution module. Taking SENet (SENet: Squeeze-and-Excitation Networks) for example, we select global average pooling to squeeze spatial dimension and generate statistical vectors of each channel. Then we use two fully connected layers to learn the importance weights of each channel from the statistical vector. Finally, the weights learned are multiplied with original feature map of each channel by element. That is, the importance weights of each channel are multiplied with each channel of the feature map by channel-wise. Therefore, the network can also extract saliency features of target region and suppress the influence of background noise in the case of limited computational resources, even if the channels of feature map are large.

As for the whole network structure, this paper extends MobileNetV3. The whole network contains initial standard convolution layer and lightweight bottleneck block as well as global average pooling layer. The bottleneck block includes depthwise separable convolution and SE attention mechanism, which can improve the reuse efficiency of feature while enhancing the gradient propagation. by adjusting the channel width and layer stacking depth, we can control the trade-off between accuracy and speed. Finally, the feature map extracted by this backbone network is input into adaptive feature pyramid module for multi-scale

fusion. The feature representation extracted by this network contains both rich semantic information and enough spatial information, which can be used for robot vision task.

III. ADAPTIVE MULTI-SCALE FEATURE FUSION MODULE

In robot vision tasks, we design an adaptive multi-scale feature fusion module, which takes three feature maps from lightweight backbone network as input. They are denoted as C3, C4 and C5, and their downsampling ratios are 8, 16 and 32, respectively. These three feature maps have different receptive fields and resolutions. C3 has high resolution but low semantic level, and it is used for locating small targets; C5 has rich semantic information but low resolution, and it is used for recognizing large targets. In traditional feature pyramid networks, high-level semantic information is passed layer by layer from top to bottom through a top-down path, and then connected laterally with bottom-up multi-scale features. However, these lateral connections simply add lateral skip connections between every two adjacent layers, and they treat all levels of features equally, i.e., they do not consider the different importance of features at different levels in different scenes.

Our adaptive feature pyramid makes two improvements compared with traditional FPN structure. First, we construct a top-down feature enhancement path. After 1×1 convolutional dimensionality reduction, C5 is upsampled and fused with C4 to generate P4; then, P4 is upsampled and fused with C3 to generate P3. This enables rich detail localization information of lower-level features to be supplemented by high-level semantic information. Secondly, we introduce a channel attention mechanism at lateral connections to learn to adaptively weight features from different levels. Specifically, when generating P4, the upsampled C5 feature is concatenated with C4 feature along the channel dimension. Global average pooling and two fully connected layers are used to learn the fusion weights of each channel, and finally a weighted sum is obtained to get the enhanced P4. In this way, the network can select the contribution of each layer's feature adaptively according to the content of input image. When the input scene is seriously occluded or contains dense targets, it will give more attention to detail information. When the scene is complex, it will enhance the semantic guidance.

To enhance the expressive power of multi-scale features, we add a bottom-up enhancement branch after the pyramid structure. P3 is downsampled through convolution with stride=2 and fused with P4 to generate N4. Then, N4 is downsampled and fused with P5 to generate N5. This reverse path can pass the low-level detail localization information to higher levels, and they form an bidirectional information flow. Finally, the module outputs three kinds of feature maps. P3 is used for the detection of small targets; N4 is used for the detection of medium targets; N5 is used for the detection of large targets.

After feature fusion, use deformable convolution to adjust adaptively the sampling points on output feature map of each layer. Different from the fixed sampling grid in standard convolution, deformable convolution learn the offset of each sampling point and allow the convolution kernel to adjust the receptive field adaptively according to the real shape and pose of target. This will bring large adjustment on stacked targets with different poses. By this design, adaptive multi-scale feature fusion module can alleviate the impact of scale change and occlusion adaptively and offer more accurate feature for following detection heads as input.

IV. IMPROVED BOUNDING BOX REGRESSION LOSS FUNCTION

In object detection tasks, the design of the bounding box regression loss function directly affects the model's localization accuracy. Traditional regression loss functions, such as Smooth L1, can measure the coordinate difference between the predicted and ground truth boxes, but their optimization objective is inconsistent with the final evaluation metric, IoU (Intersection over Union). More importantly, when the predicted and ground truth boxes have no overlapping area, the IoU value is zero and the gradient vanishes, causing the regression direction to lose guidance, making it difficult for the model to accurately locate densely stacked or mutually occluded targets in complex scenes.

To address these issues, this paper adopts Complete Intersection over Union (CIOU) as the loss function for bounding box regression. Compared to GIOU, CIOU not only addresses the gradient vanishing issue in non-overlapping cases but also accounts for the aspect ratio and center point distance between boxes. This is particularly critical for the densely stacked packages and multi-scale production defects mentioned in this study, where precise overlap and shape alignment are required for quality control. GIOU introduces the minimum bounding rectangle as a geometric constraint on top of the standard IoU, effectively reflecting the spatial relationship between the predicted and ground truth boxes, and still providing gradient information even when there is no overlap. The specific calculation process is as follows: First, calculate the IoU value between the predicted box B_p and the ground truth box B_g; second, find the smallest closed rectangle C that can simultaneously contain B_p and B_g; then calculate the proportion of the area in C that does not belong to $B_p \cup B_g$ to the total area of C; finally, obtain the expression for GIOU as (1):

$$GIOU = IoU - \frac{|C \setminus (B_p \cup B_g)|}{|C|} \tag{1}$$

The value range of GIOU is $[-1, 1]$. When the two boxes completely overlap, GIOU=1. When the two boxes are infinitely far apart, GIOU approaches -1. The corresponding regression loss function is defined as (2):

$$L_{reg} = 1 - GIOU \tag{2}$$

Compared with IoU loss, GIOU loss not only focuses on optimizing the overlapping area, but also prompts the predicted box to move towards the ground truth box to reduce the area of the smallest closed rectangle, thereby continuously providing effective regression gradients during training. This characteristic is particularly beneficial for overlapping targets in robot grasping scenarios. When multiple objects are stacked, there is often partial overlap or proximity between the detection boxes. GIOU loss can guide the predicted boxes to converge quickly to the correct position and avoid getting trapped in local optima. In the algorithm of this paper, the detection head outputs the class probability and bounding box coordinates at the same time. The overall loss function is composed of the weighted classification loss and regression loss, as shown in formula (3):

$$L = L_{cls} + \lambda L_{reg} \tag{3}$$

The classification loss adopts the focal loss to alleviate the problem of imbalance between positive and negative samples. The regression loss is the above-mentioned GIOU loss. The weight coefficient λ is set to 1.5 to balance the contribution of the two types of tasks [11].

V. MODEL COMPRESSION PRUNING BASED ON KNOWLEDGE DISTILLATION

The essence of knowledge distillation is to use a pre-trained complex teacher network to assist in training a lightweight student network. Different from using hard labels (i.e., true category annotations) to supervise the training of networks, the soft label output by the teacher network can provide richer inter-category similarity information and help to train the student network. In this paper, we choose the complete model without compression pruning as the teacher network. This model has high accuracy in detecting the objects on our own dataset, but it has many parameters. And the student network is the backbone network to lighten the parameters. During the training process, the teacher network outputs the category probability distribution and the bounding box regression results for the input image, and the student network outputs the category probability and bounding box regression results at the same time, and then the two objectives should be optimized: cross-entropy loss and regression loss with true labels, and distillation loss with the soft labels output by the teacher network.

The generation of soft labels requires the introduction of a temperature parameter T to smooth the softmax function, as shown in the following formula (4):

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{4}$$

where z_i is the logits output of the teacher network. The higher the temperature T, the smoother the generated category distribution, and the more implicit relationships between categories can be revealed. In this experiment, T is set to 3 to balance the information content of soft labels with training stability. The distillation loss uses KL divergence to measure the difference between the student and teacher output distributions. The final total loss function is the weighted sum of the hard label loss and the distillation loss. The weight coefficients are determined to be 0.7 and 0.3 through cross-validation. This training method enables the student network to inherit the superior discrimination ability of a high-parameter teacher network (e.g., a non-pruned ResNet-101 based detector) for complex scenes. Through knowledge distillation, the student network captures complex inter-class relationships and refined feature representations that exceed the capacity of standard supervised training. Consequently, the final optimized model achieves an mAP of 43.5%, significantly outperforming the unpruned baseline (36.2%) and the initial lightweight version (35.8%) by leveraging the 'dark knowledge' transferred from the teacher [12]. After completing initial training of the backbone, this paper further uses channel pruning technology to remove redundant convolutional kernels, followed by knowledge distillation to recover and enhance the performance of the pruned student network. By performing distillation after pruning, the student network effectively inherits the complex feature-mapping capabilities of the teacher network, mitigating the accuracy degradation typically caused by structural sparsification. The pruning process evaluates the importance based on the scaling factor of the BN layer (Batch Normalization). In the batch normalization (BN) layer following each convolutional layer, the scaling factor γ reflects the contribution of that channel to the output. If the value of γ is close to zero, the corresponding convolutional kernel can be considered redundant. This paper uses the L1 norm as a sparsity regularization term to guide γ to converge to zero during the fine-tuning training after distillation. After training, a global threshold is set, and all channels with γ below the threshold are pruned. To reduce accuracy loss, pruning is done iteratively: approximately 20% of the channels are pruned in each round, followed by short-cycle fine-tuning to restore accuracy. This process is repeated until the target compression ratio is achieved.

VI. EVALUATION OF ALGORITHM DETECTION PERFORMANCE

A. EXPERIMENTAL SETUP

1) Datasets

Two datasets were used for evaluation. The publicly available dataset, COCO 2017 validation set, contained approximately 50,000 images and 80 object categories, used for general performance comparison of the algorithm. The self-built dataset is a specialized defect dataset for industrial robotic inspection, containing 8,000 images of various technical materials and functional materials in manufacturing environments. It covers 15 specific categories of industrial defects (such as material breakages, microscopic filament knots, texture irregularities, and oil stains), with approximately 45,000 labeled targets. It highlighted complex scenes

such as lighting changes, mutual occlusion, and small-scale targets, used to verify the algorithm's adaptability in real robot operating environments.

2) Experimental Platform

The hardware platform used an NVIDIA Jetson Orin Nano embedded module (simulating a modern robot's onboard computing environment), with an 6-core Arm Cortex-A78AE v8.2 64-bit CPU, an Ampere architecture GPU with 1024 CUDA cores, and 8GB of memory. The Orin platform was selected over legacy modules like the TX2 to provide a contemporary benchmark for high-speed inspection in 2026. The software environment consisted of Ubuntu 18.04 operating system, CUDA 10.2, CUDNN 8.0, and the PyTorch 1.8 deep learning framework.

3) Training Parameters

The input image size was uniformly adjusted to 320×320 pixels. The SGD optimizer was used with an initial learning rate of 0.01, momentum of 0.9, weight decay of 0.0005, and a batch size of 16. Training was conducted for 120 epochs, with the learning rate decaying by a factor of 10 at the 60th and 90th epochs. Data augmentation strategies included random flipping, color dithering, and mosaic enhancement.

B. COMPARISON EXPERIMENTS WITH MAINSTREAM ALGORITHMS

To validate the effectiveness of the proposed algorithm in this paper, we select the COCO 2017 validation set as the test set and compare the proposed algorithm with four mainstream object detection algorithms, YOLOv4, Faster R-CNN, SSD, and EfficientDet. The evaluation indicators are mean accuracy (mAP@0.5:0.95), detection speed (FPS), number of parameters (params), and floating-point operations (FLOPs). All speed tests are conducted on the NVIDIA Jetson TX2 embedded platform.

As shown in Figure 1, the optimized algorithm presented in this paper achieved a 43.5% mAP@0.5:0.95 on the COCO 2017 validation set, with a detection speed of 45 FPS, both of which are superior to the comparison algorithms. Compared to EfficientDet, mAP was improved by 0.7% while FPS increased by 40.6%; compared to YOLOv4, mAP was 2.3% higher and speed was improved by 60.7%. The number of parameters is compressed to 5.2M, only 1/8 of that of Faster R-CNN, and FLOPs are reduced to 6.8G. Experimental results show that the proposed algorithm significantly reduces computational overhead while maintaining high detection accuracy, better meeting the requirements of robot vision systems for real-time performance and lightweight design.

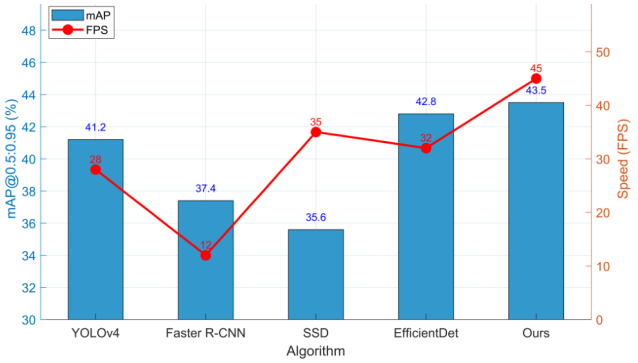

FIGURE 1. Comparison Evaluation of Mainstream Algorithms

C. ABLATION EXPERIMENTS

To analyze which optimization module makes what contribution to the performance of the detection, the ablation experiments were conducted. Taking the base detector without the optimizations as the initial stage, the lightweight backbone network, adaptive feature pyramid, GIOU loss function, and knowledge distillation pruning module were introduced one by one, and the mAP, FPS, parameter amount, and computational cost of the five combinations were compared on the self-built robot grasping dataset. The experimental results are shown in Table 2.

TABLE 2. Ablation Study Results of Self-Made Robot Grasping Dataset with Different Module Combinations

Lightweight Backbone	Adaptive Feature Pyramid	GIOU Loss	Knowledge Distillation & Pruning	mAP (%)	FPS	Params (M)	FLOPs (G)
				36.2	18	42	52
✓				35.8	32	11.5	12.3
✓	✓			39.1	28	13.2	14.5
✓	✓	✓		41.3	27	13.2	14.5
✓	✓	✓	✓	43.5	45	5.2	6.8

As shown in Table 2, the lightweight backbone reduces the number of parameters from 42.0M to 11.5M and increases FPS to 32, but mAP decreases by 0.4%. Introducing the adaptive feature pyramid recovers mAP to 39.1%, a 2.9% improvement over the baseline. GIOU loss further increases mAP to 41.3%. Finally, through knowledge distillation and pruning, the number of parameters is compressed to 5.2M, FLOPs decrease to 6.8G, FPS reaches 45, and mAP increases to 43.5%. Ablation experiments demonstrate that each module positively contributes to performance optimization, and the proposed algorithm achieves both lightweight model and real-time performance requirements while maintaining high accuracy.

D. ADAPTABILITY EXPERIMENTS IN COMPLEX SCENES

To evaluate the robustness of the proposed algorithm in complex environments, adaptation experiments in complex scenes were designed. Six typical challenging subsets were selected from a self-built robot grasping dataset: normal lighting, sudden lighting changes, severe occlusion, dense

small targets, multi-scale mixing, and dynamic blur. Compared with an unoptimized baseline detector, the mean accuracy (mAP) and false negative rate were statistically analyzed for each scene to examine the algorithm's adaptability to interference factors such as lighting changes, occlusion levels, and target scale.

Figures 2(a) and (b) show the comparison of detection accuracy (mAP) and false negative rate between the baseline algorithm and the proposed algorithm in six complex scenes, respectively. In the adaptation experiments in complex scenes, the proposed algorithm showed significant advantages in all challenging environments. Compared to the baseline algorithm, under normal lighting conditions, the mAP improved from 36.2% to 43.5%, and the false negative rate decreased from 20.0% to 12.0%. In scenarios with sudden changes in lighting, the mAP reached 40.2%, an improvement of 11.7 percentage points over the baseline. In severely occluded scenarios, the mAP improved by 14.4% to 38.7%, and the false negative rate decreased by 25%. In scenarios with dense small targets, the mAP improved by 15.4% to 37.5%, and the false negative rate decreased by 28%. In multi-scale mixed and dynamic blur scenarios, the mAP reached 41.0% and 39.3%, respectively, with the false negative rate controlled below 20%. Experimental results show that the proposed optimized algorithm has good robustness to complex factors such as lighting changes, occlusion, and scale differences, and can meet the detection requirements of actual robot operating environments.

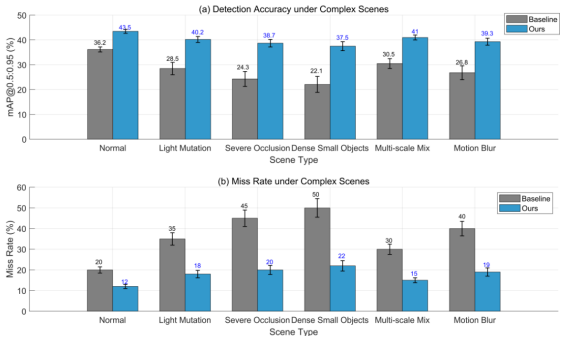

FIGURE 2. Comparison of Adaptability in Complex Scenes

VII. CONCLUSION

This paper proposes a multi-stage optimization algorithm for robot visual target detection in manufacturing, integrating backbone networks, feature fusion, loss function refinement, and model compression to identify production defects and material irregularities. By streamlining the detection of material breakages and texture inconsistencies, the algorithm ensures high-precision quality control for advanced functional materials and technical materials in real-time robotic inspection. The findings indicate that the ability to achieve a balance between the detection accuracy and the computational efficiency is achieved with the help of the lightweight design and the implementation of an attention mechanism. The pyramid of adaptive features and the regression loss is much better at raising the robustness of the model to work under occlusion and multi-scale conditions. Thus, despite its workability on publicly available datasets, as well as self-constructed situations, the algorithm does not take into account the extreme cases like the motion blur on the target in dynamic backgrounds and the drastic light variations. Future development will integrate temporal information to enhance the detection continuity of dynamic fiber irregularities and utilize neural architecture search to automate the design of models for monitoring high-speed spinning and weaving processes. This advancement will facilitate the seamless deployment of the algorithm across diverse robotic systems for the real-time quality inspection of advanced functional textiles and technical fibers.

AUTHOR CONTRIBUTIONS

Xin Zhong designed, collected and analyzed the data, and drafted the manuscript. Xin Zhong conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Xin Zhong participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] X. Liu, C. Huang, H. Zhu, et al., "State-of-the-art elderly service robot: Environmental perception, compliance control, intention recognition, and research challenges," IEEE Systems, Man, and Cybernetics Magazine, vol. 10, no. 1, pp. 2–16, 2024, doi: 10.1109/MSMC.2023.3238855.

[2] T. Zhang, D. B. Kaber, B. Zhu, et al., "Service robot feature design effects on user perceptions and emotional responses," Intelligent service robotics, vol. 3, no. 2, pp. 73–88, 2010, doi: 10.1007/s11370-010-0060-9.

[3] Y. Li and C. Wang, "Effect of customer's perception on service robot acceptance," International Journal of Consumer Studies, vol. 46, no. 4, pp. 1241–1261, 2022, doi: 10.1111/ijcs.12755.

[4] D. D. Aboyomi and C. Daniel, "A comparative analysis of modern object detection algorithms: YOLO vs. SSD vs. faster R-CNN," ITEJ (Information Technology Engineering Journals), vol. 8, no. 2, pp. 96–106, 2023, doi: 10.24235/itej.v8i2.123.

[5] F. Hawlader, F. Robinet, and R. Frank, "Leveraging the edge and cloud for V2X-based real-time object detection in autonomous driving," Computer communications, vol. 213, pp. 372–381, Jan. 2024, doi: 10.1016/j.comcom.2023.11.025.

[6] L. Zhao and L. Wang, "A new lightweight network based on MobileNetV3," KSII Transactions on Internet and Information Systems (TIIS), vol. 16, no. 1, pp. 1–15, 2022, doi: 10.3837/tiis.2022.01.001.

[7] M. Shan, Z. Zhu, H. Chen, et al., "Service robot's responses in service recovery and service evaluation: the moderating role of robots' social perception," Journal of Hospitality Marketing & Management, vol. 33, no. 2, pp. 145–168, 2024, doi: 10.1080/19368623.2023.2246456.
[8] Z. Zou, K. Chen, Z. Shi, et al., "Object detection in 20 years: A survey," Proceedings of the IEEE, vol. 111, no. 3, pp. 257–276, 2023, doi: 10.1109/JPROC.2023.3238524.
[9] Liululu, Z. He, Z. Ma, et al., "Small sample target detection based on multi-stage feature redistribution algorithm," Journal of UESTC, vol. 52, no. 1, pp. 116–124, 2023, doi: 10.12178/1001-0548.2022016.
[10] Y. Tang, R. Zhang, R. Ding, et al., "MSA net: a small sample target detection method based on multi-stage attention mechanism," Optical instruments, vol. 45, no. 6, pp. 14–24, 2023, doi: 10.3969/j.issn.1005-5630.20230203011.
[11] S. Yao, R. Guan, X. Huang, et al., "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 2094–2128, 2023, doi: 10.1109/TIV.2023.3307157.
[12] G. Cheng, X. Yuan, X. Yao, et al., "Towards large-scale small object detection: Survey and benchmarks," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 11, pp. 13467–13488, 2023, doi: 10.1109/TPAMI.2023.3290594.