International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp.1-10

https://doi.org/10.70917/ijcisim-2026-0088
Article

Application of Convolutional Neural Networks to Achieve Performance Enhancement of Robot Vision Inspection Systems for Industrial Production Lines

Jianjia Qi *

Heilongjiang Institute of Technology, Harbin 150050, Heilongjiang, China; hgcqjj@163.com

Abstract: In response to the issue of dynamic complexity in robot vision inspection technology on industrial production lines, this paper proposes a method for robot vision inspection that combines convolutional neural networks (CNN) with background difference methods and spatio-temporal context target tracking technology. Using the established image database containing five types of workpieces—bearings, screwdrivers, gears, pliers, wrenches, and other five categories of workpiece images, and utilizing the proposed lightweight residual attention and decoupled attention mechanisms to design the FCN8s model. Experimental validation revealed that the model achieved an average pixel accuracy, average accuracy, and average IoU accuracy of 98.30%, 82.46%, and 73.62%, respectively, on the test set, meeting the real-time requirements of industrial visual inspection. In terms of accuracy, the recognition rate increased by approximately 15.97% compared to traditional methods, and the speed increased by approximately 83.26% to 84.65%. This paper adopts a multi-technology fusion solution to provide a more reliable and practical means for target detection and recognition in complex industrial environments, thereby providing a robot vision system for intelligent manufacturing.

Keywords: convolutional neural network; machine vision; industrial robot; target detection; intelligent manufacturing

1. Introduction

The advancement of smart manufacturing has imposed increasingly stringent requirements on the real-time performance and accuracy of machine vision systems in smart manufacturing production lines. Traditional machine vision image inspection methods generally function normally in fixed inspection environments [1-2]. However, they often struggle to cope with the rapidly changing conditions of industrial sites, particularly in harsh environments involving spectral changes, target deformation, and motion blur, where both inspection accuracy and speed fall short of requirements [3-5]. With the implementation of the “Made in China 2025” strategy and the advanced development of smart manufacturing equipment, industrial robots have also been upgraded, transitioning from single-task operations to concurrent multi-task execution and autonomous decision-making [6-8]. During this period, the “eyes” of industrial robots—machine vision systems—also require systematic improvements in target recognition accuracy, detection response latency, and detection model generalization [9].

During this period, the development of deep learning, especially CNN, has brought about a qualitative change in the fields of image recognition and target detection. CNN possesses strong feature expression capabilities and robustness, enabling it to automatically learn multi-layer image semantic expressions, significantly enhancing the dynamic adaptability of models in industrial scenarios [10-13]. In the application of robot vision detection on production lines, Reference [14] proposed a CNN-based grasping detection method for object recognition and introduced a grasping algorithm to determine the configuration of the gripper, enabling industrial robots to retrieve items from various objects. Reference [15] proposes a robot path planning algorithm combining deep Q-learning with CNN, which can analyze

various complex production environment conditions and adapt to changing conditions to achieve flexible and efficient movement. Literature [16] utilizes a CNN model to improve the visual recognition system of industrial robots. Specifically, by employing the FastRCN algorithm and a VGG-16 classification network enhanced with a super-block scheme, it achieves an 82.34% recognition accuracy rate, effectively addressing issues such as positioning errors, slow recognition speeds, and low accuracy. Reference [17] proposes a CNN-based visual classification system that integrates cloud-edge computing technology, suitable for flexible manufacturing systems. By leveraging CNN, it achieves high classification accuracy and efficient processing, meeting the demand for accurate part classification in smart manufacturing. Literature [18] addresses the issue of low efficiency in product defect detection on high-speed production lines by proposing a three-stage convolutional neural network model. This model combines object detection with defect classification and demonstrates high precision and rapid quality correction effects on actual production lines. With the continuous development of convolutional architectures and attention mechanisms, the model design of industrial vision inspection systems no longer relies on manual feature design [19-21]. However, factors such as data collection costs in industrial settings, the diversity of training samples, and the complexity of deployment environments constrain the performance and application of CNN models.

To address this, this paper introduces background subtraction methods, spatio-temporal context tracking algorithms, and various attention networks, and proposes a lightweight visual inspection system solution suitable for industrial settings, based on the premise of training a high-quality image library. By simplifying the model structure, improving inference speed, and enhancing robustness, this solution aims to assist industrial robot vision systems, developing them into smarter, more efficient, and more stable systems, thereby laying a solid technical foundation for future intelligent manufacturing. This paper establishes a high-precision database of industrial objects containing five typical industrial parts, various lighting conditions, and dynamic backgrounds. Sample preprocessing and augmentation are performed to ensure the diversity and generalization of training objects. A spatio-temporal integrated tracking method based on background difference and context is proposed to accurately initialize and track the target, thereby constructing the front-end unit of the detection system. We propose using residual attention and decoupled attention mechanisms to enhance the spatial sensitivity of object deep representation and shallow recognition, and employ comb-shaped convolutions and overparameterized convolutions to improve the model's expressive capability and computational efficiency to meet industrial site requirements. Additionally, we use model compression and marginalization techniques to enhance the model's rapid response and real-time processing capabilities. Finally, the system's performance is comprehensively validated across various scenarios in terms of accuracy, robustness, and speed using real-world industrial test datasets.

2. Overview of Performance Improvements in Industrial Production Line Robot Vision Inspection Systems

2.1. Theoretical Basis of the Study

Convolutional neural networks (CNNs) are one of the key technologies in deep learning. In the field of computer vision, they have become an essential core technology due to their powerful feature extraction capabilities. The advantages of CNNs are particularly evident in industrial vision inspection [22-25]. The key difference between convolutional neural networks and traditional neural networks lies in the introduction of convolution operations. By utilizing weight sharing and local connection mechanisms, CNNs optimize the number of parameters while retaining spatial information. The mathematical expression of the convolution layer is as follows:

$$y_{i,j} = \sum_{m,n} x_{i+m,j+n} \cdot w_{m,n} + b \tag{1}$$

In the equation, $y_{i,j}$ represents the value of the output feature map at position (i, j), $x_{i+m,j+n}$ represents the value of the input feature map at position $(i+m, j+n)$, $w_{m,n}$ is the weight of the convolution kernel at position (m, n), and b is the bias term.

It uses a convolutional kernel to slide across the input image and perform dot product operations at local positions, enabling the extraction of feature information such as image edges, textures, and shapes. Convolution differs from fully connected networks in that it uses the same set of parameters to process the entire input image, allowing for parameter sharing during computation. This not only reduces the number of parameters in the model and improves computational efficiency but also helps reduce overfitting. The pooling layer in the network is used for downsampling, reducing the spatial dimensions

of the feature map while extracting key feature information. This also improves computational efficiency and enhances the model's robustness to translation transformations. The fully connected layer in the network is used to flatten the feature map and connect all neurons, integrating high-level features to perform classification or regression tasks.

Object detection is commonly found in industrial vision systems where targets need to be detected in dynamic background images. In such scenarios, the background subtraction method has its unique advantages. It extracts foreground targets by calculating the difference between the current image and a pre-constructed background image, as expressed below:

$$D\_t = \cdot | I\_t - B\_t | \tag{2}$$

In the equation, D_t represents the difference image, I_t denotes the current frame image, and B_t denotes the background image. When the pixel value in the difference result D_t exceeds the set threshold τ, the pixel is classified as a foreground object; otherwise, it is classified as background. This method has significant advantages in terms of effectiveness and speed, but it struggles to adapt to complex background environments and lighting changes. To overcome these limitations, some researchers have employed a Gaussian background model, treating each pixel as a mixture of Gaussian distributions to model the background, thereby addressing background changes caused by complex environments. The Vibe method uses the neighborhood of each pixel as a sample for background modeling, offering greater universality. The hybrid background difference method validated via inter-frame difference analysis is more effective in overcoming noise interference and achieving more precise target localization.

Additionally, spatio-temporal context-based tracking methods have advantages in target tracking under motion conditions. Spatio-temporal context-based tracking is a method that utilizes temporal and spatial constraint information to establish a target motion model, and then applies this model to search for the current frame based on the previous frame, thereby achieving more precise target detection, reducing false tracking, and enhancing tracking robustness [26]. A typical tracking algorithm based on spatio-temporal context is the tracking algorithm under the Bayesian framework, which establishes a mapping relationship between target state and observation data, transforming the tracking problem into a probability estimation problem. This method uses the target's position and motion state from the previous frame to predict the target's location in the current frame, thereby reducing the search space. Within the predicted location, it utilizes spatial context information to establish a confidence mapping function to determine the target's precise position. The fast Fourier transform converts cumbersome frequency-domain convolution operations into convolution operations in the spatial domain, effectively improving the computational efficiency of the algorithm and making it suitable for real-time industrial applications.

Attention mechanisms based on deep convolutional neural networks are an effective way to improve model performance. Inspired by the selective attention characteristics of the human eye, they automatically focus on important positions in the feature maps of input data while ignoring unimportant information. Attention mechanisms in convolutional neural networks can be divided into two types: spatial attention and channel attention. Spatial attention generates a two-dimensional weight map, enhancing the distinguishability of spatial position features in feature maps. Channel attention produces a one-dimensional vector, used to generate weight values indicating the importance of features across different channels. A representative achievement of channel attention is SENet, which first uses global average pooling to compress the spatial dimension, then uses fully connected layers to learn the dependencies between channels, thereby learning the weights of the channels and improving model performance. Combining channel attention and spatial attention results in CBAM, which ultimately outperforms either channel attention or spatial attention alone.

Based on the above analysis, when applied to industrial robot vision inspection systems, due to the complexity and dynamic changes of the detection targets, it is difficult to find a single method that can achieve the expected results. By integrating convolutional neural networks with background subtraction methods, spatio-temporal context tracking methods, and attention mechanisms, and combining the advantages of these various methods, a more effective and robust vision inspection system can be established. Through this paper, we observe that the background subtraction method provides the initial position of the target, offering a good starting point for the tracking algorithm. The spatio-temporal context tracking algorithm performs stable tracking of the target, while the convolutional neural network enables accurate target classification and recognition. The attention mechanism focuses on key regions. This method effectively addresses issues in the field of industrial visual inspection by fully leveraging the advantages of multiple methods, resulting in significant performance improvements.

2.2. Current Status of Industrial Production Line Robot Vision Inspection Systems

Currently, machine vision is rapidly developing in industrial production lines. It has broken through the initial application of machine vision, which was limited to the detection of simple items, and is now being applied in fields such as the automotive industry, electronics industry, and automated machinery. However, the difficulty of achieving both detection accuracy and real-time performance remains a key issue affecting the entire field.

3. Methods for Improving the Performance of Robot Vision Inspection Systems

3.1. Data Collection and Preprocessing

The object detection capability of industrial inspection vision robots is significantly constrained by the quantity and variety of training data available. To address this, a comprehensive image library targeting common mechanical tools and components has been developed. The designed image library encompasses five representative industrial application targets: bearings, screwdrivers, gears, pliers, and wrenches. These objects share similar shapes, are made of soft materials, and serve distinct functional purposes. Their images exhibit diversity and typicality, providing the model with ample training data. Image acquisition was performed using a high-speed industrial camera (resolution: 1920×1080 pixels), with the camera lens fixed above the conveyor belt at a distance of 50 cm, ensuring stable and pixel-clear target object images. To simulate real-world industrial conditions, image acquisition was designed under various scenarios: single-object image acquisition on a static background, multi-object image acquisition on a static background, and object image acquisition under simulated production conditions (conveyor speed ranging from 5 to 20 cm/s). Lighting conditions were set as follows: normal lighting (500–600 lux), low lighting (200–300 lux), and high lighting (800–1000 lux).

Through the aforementioned data collection methods, a database comprising 3,157 high-quality images was constructed, including 2,248 single-object single-class images and 909 multi-object multi-class images, with the specific composition detailed in Table 1. The raw collected data contained significant noise and interference, and using unprocessed images directly for training would impact detection accuracy. To address this, we designed a systematic preprocessing method, which can be divided into preprocessing and enhanced preprocessing. Preprocessing addresses quality issues by converting RGB three-channel images into single-channel images through grayscale processing, thereby reducing image computational complexity while preserving structural information. A combination of Gaussian filtering and median filtering is used to eliminate Gaussian noise and salt-and-pepper noise. The filter kernel size varies depending on the target size, typically ranging from 3x3 to 5x5. Histogram equalization is applied to enhance image contrast and address uneven lighting conditions. An adaptive threshold segmentation method is used for image binarization to obtain the target contour lines, which serve as the initial positioning for the subsequent background subtraction method.

Table 1. Mechanical tools and parts image database.

Type	Single target, single category	Multiple goals and multiple categories	Background type	Total
Bearing	465	203	Static/Motion	668
Screwdriver	427	178	Static/Motion	605
Gear	510	196	Static/Motion	706
Pliers	398	152	Static/Motion	550
Wrench	448	180	Static/Motion	628
Total	2248	909	-	3157

Data augmentation is primarily aimed at addressing data heterogeneity, and we have implemented various data augmentation techniques. Random rotation (-15° to +15°) simulates changes in the orientation of objects on a conveyor belt, random scaling (0.8 to 1.2 times) simulates size changes due to distance, random changes in brightness (-20% to +20%) and contrast (-15% to +15%) simulate the effects of brightness changes, and random horizontal flipping, vertical flipping, and random cropping simulate changes in object states. These methods can expand the original dataset by approximately three times, effectively enhancing the model's generalization and robustness against unknown information. For images with motion backgrounds, we designed a specialized preprocessing process. We employed adaptive background modeling technology to establish a dynamic background model using a Gaussian mixture model, detected moving objects using frame difference methods combined with temporal filters, and repaired object contours using morphological operations (opening and closing operations) to remove fragments and holes during object detection.

We performed semi-automatic annotation on the entire dataset, using a pre-trained YOLOv3 model for automatic annotation and initial inspection, followed by manual refinement by workers. For single-object, single-class images, the annotation information includes the category and bounding box coordinates; for multi-object, multi-class images, the annotation information is provided for the corresponding targets. We also annotated pixel-level mask information for some images to provide more detailed training supervision for semantic segmentation using fully convolutional neural networks. After annotation, the dataset is divided into training, validation, and test sets in a 7:2:1 ratio, ensuring balanced distribution across categories and types. Through these systematic data collection and preprocessing steps, a high-quality, content-rich image database for industrial machine parts and tools has been established, laying a solid foundation for subsequent convolutional neural network model training.

3.2. Model Training and Optimization

Based on the database and preprocessing workflow established in the preceding section, this section delves into model training and optimization strategies, focusing on the integration of background difference methods with spatio-temporal context-aware object tracking techniques, as well as the process of enhancing the performance of convolutional neural network models. We utilize the background difference method to automatically obtain the initial position of the target, establishing a precise foundation for object tracking. The specific implementation employs an improved adaptive hybrid Gaussian background model to handle complex dynamic backgrounds, with its mathematical expression defined as:

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} \cdot \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{3}$$

In the equation, X_t represents the pixel value at time t, K is the number of Gaussian distributions (in this study, $K = 3$), $w_{i,t}$ is the weight of the i th Gaussian distribution, η represents the Gaussian distribution function, and $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and covariance matrix of the i th Gaussian distribution, respectively. This method can accurately locate the target in the initial frame image with an error controlled within 2 pixels, thereby introducing a spatio-temporal context-based target tracking method. By utilizing the spatio-temporal relationships of the target to construct a representation model, it overcomes the deficiency of traditional tracking algorithms that tend to lose the target in dynamic backgrounds. This method is based on a Bayesian framework and employs a confidence function, i.e.:

$$c(x) = be^{-|x-x^*|^\alpha/\sigma^2} \tag{4}$$

Calculate the target position.

In the formula, $c(x)$ is the confidence at position x, x^* is the center position of the target, and b, α and σ are control parameters. v is To improve computational efficiency, the fast Fourier transform is used to convert the convolution operation into a frequency domain multiplication, that is:

$$\mathcal{F}(c) = \mathcal{F}(h^{sc}) \odot \mathcal{F}(I) \tag{5}$$

In the equation, $\mathcal{F}$ is the Fourier transform, h^{sc} is the spatio-temporal context filter, I is the input image, and $\odot$ is element-wise multiplication.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{Dice} \tag{6}$$

In the formula, the weight coefficients λ_1 and λ_2 are set to 0.7 and 0.3, respectively.

To further enhance the model's performance, we introduce multiple attention mechanisms and deep learning optimization techniques. In particular, we designed a residual attention module, which compresses spatial information through global average pooling and then uses a two-layer fully connected network to learn the dependencies between channels, expressed as:

$$F_{out} = F_{in} \cdot \sigma(W_2 \delta(W_1 GAP(F_{in}))) \tag{7}$$

In the equation, F_{in} and F_{out} represent the input and output feature maps, respectively, GAP denotes the global average pooling operation, W_1 and W_2 are the weights of the fully connected layers, δ is the ReLU activation function, and σ is the Sigmoid function. This module effectively suppresses

the activation of irrelevant features in the deep layers of the network.

At the same time, we developed a decoupled attention module that separates the spatial and channel dimensions for processing, mathematically expressed as:

$$F_{out} = F_{in} \cdot M_c \cdot M_s \tag{8}$$

In the equation, M_c and M_s represent the channel and spatial attention masks, respectively. Placing this module at the shallow jump connection significantly enhances feature expression capabilities. Considering the importance of inference speed, we draw inspiration from comb convolution and use a binary mask tensor to reduce spatial connections by approximately 50%, represented as:

$$Y = X \odot M * W \tag{9}$$

In the equation, X is the input feature map, M is the binary mask, W is the convolution kernel weight, and * denotes the convolution operation.

At the same time, overparameterized convolution technology is introduced to enhance the model's expressive power without increasing the inference computation load. To address the overfitting issue caused by the limited number of industrial site images, regularization techniques such as batch normalization, weight decay (coefficient of 0.0005), and Dropout (rate of 0.5) were applied. After model training was completed, model compression processing was performed to adapt to edge computing device deployment requirements. The convolution layers were compressed using an energy transfer-based low-rank projection algorithm, and the floating-point weights were converted to 8-bit integers using quantization techniques.

4. Experiments and Analysis

4.1. Experimental Results

The convolutional neural network designed in this paper is applied to an industrial production line robot vision inspection system. Through systematic experimental analysis, it has been proven to have excellent performance. The hardware environment of the experimental platform uses an NVIDIA GeForce RTX 3080 GPU workstation, and the software environment uses the Ubuntu 20.04 operating system and the PyTorch 1.9.0 deep learning framework, ensuring that the experimental results are reliable and reproducible. For this experiment, we primarily focused on evaluating the target tracking performance of industrial production line robots and assessing the classification and recognition performance of fully convolutional neural network models. The target tracking performance evaluation criteria used the Euclidean distance criterion with a pixel threshold of 10. The experimental data statistics are shown in Table 2. From this, we can clearly see that there are significant differences in tracking accuracy among different types of mechanical tools and parts. Industrial parts such as bearings and gears, which have regular geometric structures with distinct features and clear boundaries, and pliers with prominent jaw shapes achieved 100% ideal tracking accuracy in the experiment. This is because they possess relatively stable geometric shapes and appearance features, enabling the establishment of accurate target models in spatio-temporal context tracking algorithms. In contrast, screwdrivers and wrenches achieved tracking accuracy rates of only 72.3% and 65.8%, respectively, in the experiment. Their relatively poor tracking performance is due to their slender shapes and relatively sparse shape features, which pose challenges for the tracking algorithm in terms of conveyor belt movement obstruction and angle conversion. Especially the open end of the wrench is easily confused with the background at certain angles, making feature extraction extremely difficult, severely affecting the stability and accuracy of tracking.

Table 2. Target tracking precision experiment results.

Type	Test sample	Successful tracking	Tracking accuracy (%)	Average tracking time (ms)
Bearing	150	150	100.0	12.3
Screwdriver	145	105	72.3	15.7
Gear	160	160	100.0	11.8
Pliers	135	135	100.0	14.2
Wrench	140	92	65.8	16.9
Average	146	128.4	87.6	14.2

The classification performance test results of the fully convolutional neural network FCN8S are based on 309 test images, which include different lighting conditions, different background complexities,

and different target pose differences. The comprehensive test results are shown in Table 3. The average pixel accuracy is 98.30%, indicating strong pixel-level classification capability. The network effectively distinguishes the classification of each pixel in the image, demonstrating that the designed network can correctly classify the pixel types in each image. Additionally, the average accuracy reaches 82.46%. Considering the precision and recall rate metrics of this method, it reflects that the convolutional neural network with the attention mechanism fusion performs well in industrial part recognition. The average IoU is 73.62%, which is lower than the average pixel accuracy and average precision metrics. However, due to the complexity of industrial targets with varying background environments and complex target boundaries, as well as unavoidable human influences during the annotation process, the actual target classification performance is better. This proves that the model has excellent target localization performance. Further analysis of category differentiation confirms a pattern similar to that observed in target tracking experiments: parts with regular geometric shapes, such as bearings and gears, exhibit higher classification accuracy, while parts with irregular geometric shapes, such as screwdrivers and wrenches, have relatively lower classification accuracy. This validates that target shape features are the primary factor influencing the performance of visual inspection systems.

Table 3. Experimental results of the FCN8S model.

Type	Pixel precision (%)	Precision (%)	IoU precision (%)	Reasoning time (ms)
Bearing	99.2	89.3	81.7	23.1
Screwdriver	96.8	76.2	65.2	24.7
Gear	99.5	91.1	83.4	22.8
Pliers	98.7	85.6	78.9	23.9
Wrench	97.3	70.1	58.9	25.2
Average	98.30	82.46	73.62	23.94

Table 4 presents a comparison of performance across different methods. Comparative experiments demonstrate that our research approach differs significantly from traditional visual recognition methods. Traditional methods rely on manually constructed feature extractors such as SIFT and SURF, combined with SVM or random forest algorithms for classification. On our test data, the best average recognition accuracy achieved by traditional methods was only 67.83%, which is far inferior to the performance of our designed FCN8S model. The actual runtime performance comparison is equally satisfactory. Traditional methods take approximately 143ms to 156ms to process an image, while the improved FCN8S model achieves an average inference runtime of just 23.94ms, representing an improvement of 83.26% to 84.65%, thereby meeting the real-time detection requirements of industrial production lines. The research results demonstrate that the CNN-based industrial assembly line robot vision detection system we designed can accurately identify targets in dynamic industrial assembly line environments and achieve high precision, providing strong support for industrial automation production. Of course, this is just the beginning. The accuracy of object detection varies significantly depending on the type of workpiece and object. Through comparison, it was found that the geometric features of the target significantly influence the accuracy of detection results. This provides critical guidance for future algorithm optimization and model improvement, facilitating further enhancements in the reliability and robustness of the method in actual complex industrial environments.

Table 4 Performance comparison results of different methods.

Method	Recognition accuracy (%)	Inference time (ms)
SIFT+SVM	65.71	143.18
SIIFT+RF	66.04	145.94
SURF+SVM	66.39	152.67
SURF+RF	67.83	155.36
FCN8S	82.46	23.94

4.2. Analysis of Results

This paper presents the results of testing and data analysis of the designed visual detection system. The target detection system, based on background difference method and spatio-temporal context-aware target tracking, demonstrates high tracking accuracy in complex and dynamic environments, with an average tracking accuracy of approximately 87.6%. Especially for industrial parts such as bearings and gears, the tracking accuracy can reach 100%. This is attributed to the high initial positioning accuracy of the background difference method (error within 2 pixels) and the spatio-temporal context algorithm,

which considers the environmental information around the target. Additionally, tracking performance is significantly influenced by target characteristics, with regular-shaped objects (bearings and gears) achieving better tracking results than irregular-shaped objects (screwdrivers and wrenches). This also provides relevant insights for future research directions, suggesting the selection of more appropriate tracking strategies for different types of workpieces. The average tracking duration of the overall detection system in this paper is only 14.2 ms, far below the standard threshold requirement of 50 ms, demonstrating significant application potential on high-speed production lines. The fully convolutional neural network achieves an excellent accuracy of approximately 98.30% during workpiece recognition, further validating the superior adaptability of deep learning methods in industrial applications. Analysis of experimental data also reveals that the accuracy of target recognition varies depending on the target's characteristics. Objects with clear edges and stable textures generally exhibit higher recognition accuracy, while objects like screwdrivers and wrenches, which have unclear edges and unstable shapes, exhibit lower recognition accuracy.

After introducing multiple attention mechanisms, the overall system performance was significantly improved. By incorporating residual attention blocks to suppress the activation of unrelated features, the network focuses solely on key features within the target region. Through decoupled attention blocks, attention is separated and processed in both spatial and channel directions, enhancing the richness of feature representations. The average runtime of this method is 23.94 ms, far meeting the real-time detection requirements of industrial environments. Compared to traditional methods, the visual detection method based on convolutional neural networks proposed in this paper achieves better results. Traditional methods based on SIFT or SURF features achieve an average recognition rate of only 66.49% on the test set used in this paper and are also slower. This is because, in real-world scenarios, these methods require prior manual design to identify key features, making them difficult to adapt to complex working environments, lighting conditions, and changes in perspective. In contrast, fully convolutional neural networks adaptively learn to express suitable and key features during training. Therefore, the method proposed in this paper can accurately detect different features in various complex images with fewer parameters and higher efficiency.

In summary, this visual detection system adopts a high-precision dynamic environment target detection and recognition method combining background difference, spatio-temporal context target tracking, and fully convolutional neural networks. It is suitable for industrial production lines with high requirements for detection accuracy and real-time performance. In particular, it has high recognition detection rates and fast real-time processing speeds for different industrial tool parts. By understanding the difficulty of detecting different objects, it points to future research directions. Further improvements in the system's applicability for industrial robot vision detection systems can be achieved by enhancing feature extraction capabilities for complex-shaped objects, integrating 3D point cloud and 2D image detection, and increasing system robustness under extremely low illumination and severe occlusion conditions.

5. Conclusion and Outlook

5.1. Conclusion

This paper primarily focuses on improving the visual inspection performance of industrial assembly line robots. By combining convolutional neural networks with background subtraction methods and spatio-temporal context-aware object tracking techniques, the study achieved stable tracking of both regular components (e.g., bearings and gears) and complex components (e.g., screwdrivers and wrenches) in complex assembly line scenarios. The accurate tracking rate for regular components reached 100%, while those for complex components were 72.3% and 65.8%, respectively. An improved FCN8s network incorporating residual and decoupled attention mechanisms was proposed, enhancing detection accuracy and robustness. The average pixel accuracy, average recognition accuracy, and IoU accuracy reached 98.30%, 82.46%, and 73.62%, respectively, with an average inference time of only 23.9ms. This fully meets the real-time requirements of industrial assembly lines. The dataset consists of five types of mechanical parts, with more diverse data augmentation techniques and improved model generalization performance.

5.2. Research Limitations and Future Prospects

Although this paper has achieved certain research results in improving the performance of robotic vision inspection systems on industrial production lines, it still has some shortcomings and requires further research to provide reference and guidance for subsequent research work. For example, this paper primarily focuses on image detection, specifically the detection and recognition of two-dimensional images. While the use of background subtraction methods and spatio-temporal context-based object

tracking techniques has effectively improved the accuracy of object detection, two-dimensional images inherently have certain limitations and cannot directly capture the three-dimensional structural characteristics of objects. In industrial applications, mechanical parts often have complex geometric shapes with multiple faces, and single two-dimensional image data alone is insufficient to comprehensively perceive the spatial distribution and geometric features of objects. For tools like wrenches and screwdrivers, which are difficult to detect, their pose transformations in three-dimensional space often cannot be intuitively represented and identified using two-dimensional images, leading to low detection accuracy for these tools. There is a phenomenon where two-dimensional image detection methods have not fully overcome the limitations of three-dimensional information. Future research that utilizes more diverse and abundant three-dimensional point cloud data is expected to significantly enhance the performance and application scope of detection systems.

The geometric features of the target extracted by detection methods based on three-dimensional point cloud data are related to spatial information and provide a more comprehensive description of the target's geometric information, making them highly valuable for future research. Three-dimensional point cloud data generated by laser radars, structured light cameras, and other devices have shown broad application prospects in many areas and have achieved certain results in industrial applications. Combining advanced technologies based on 3D point cloud data, such as the 3D point cloud detection network VoteNet, with existing 2D image detection system methods to construct a multi-modal fusion detection system is an effective means of improving the level of industrial detection systems.

In addition to the aforementioned shortcomings, regarding the non-structural issues of point clouds, future improvements could be made by increasing the processing modes of seed point clouds and introducing symmetric function training modules to enhance the detection accuracy of complex industrial parts. For feature extraction on objects with complex shapes or simple surface structures, the aforementioned models still have significant room for improvement. Capturing temporal features remains one of the bottlenecks hindering the model's ability to handle future temporal structural objects, so future research should focus more on developing deep network architectures adapted to complex industrial environments. For future research, improving filter selection can also address adaptability issues in complex industrial environments. In addition, due to deployment requirements in practical applications, future research should also focus on model lightweighting and model deployment to edge computing. At the same time, while maintaining a certain model capacity, the computational complexity of the model can be significantly reduced, enabling the deployment of more powerful industrial vision inspection systems on resource-constrained edge devices. This lays the foundation for the future application of industrial vision inspection systems in more scenarios. As mentioned above, future research will further explore multimodal data fusion, novel network architectures, enhanced environmental adaptability, and model lightweighting and miniaturization, enabling industrial visual inspection systems to provide more flexible, intelligent, and efficient services in fields such as industrial automation and smart manufacturing.

References

1. Lou, P., Li, J., Zeng, Y., Chen, B., & Zhang, X. (2022). Real-time monitoring for manual operations with machine vision in smart manufacturing. Journal of Manufacturing Systems, 65, 709-719.
2. Jain, T. (2022). Industrial objects recognition in intelligent manufacturing for computer vision. International Journal of Intelligent Unmanned Systems, 10(4), 401-415.
3. Benbarrad, T., Salhaoui, M., Kenitar, S. B., & Arioua, M. (2021). Intelligent machine vision model for defective product inspection based on machine learning. Journal of Sensor and Actuator Networks, 10(1), 7.
4. Li, Y., Zhou, Y., & Liu, H. (2024). Research on the influence of image motion blur on the effectiveness of machine vision-based metal scraps separation system. Journal of Material Cycles and Waste Management, 26(4), 2509-2517.

a) Mei, W., Zheng, Y., & Gu, Y. (2023). Design of intelligent 3d collaborative manufacturing platform for non-holonomic mobile industrial robots based on improved binocular vision. International journal of intelligent robotics and applications, 7(4), 740-751.

5. Wen, H., & Zhao, Z. (2021). How does China's industrial policy affect firms' R&D investment? Evidence from 'Made in China 2025'. Applied Economics, 53(55), 6333-6347.
6. Lv, H., Shi, B., Li, N., & Kang, R. (2022). Intelligent manufacturing and carbon emissions reduction: evidence from the use of industrial robots in China. International Journal of Environmental Research and Public Health, 19(23), 15538.
7. Yu, L., Wang, Y., Wei, X., & Zeng, C. (2023). Towards low-carbon development: The role of industrial robots in decarbonization in Chinese cities. Journal of environmental management, 330, 117216.
8. Javaid, M., Haleem, A., Singh, R. P., Rab, S., & Suman, R. (2022). Exploring impact and features of machine vision for progressive industry 4.0 culture. Sensors international, 3, 100132.

9. Sunkara, R., & Luo, T. (2022, September). No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. In Joint European conference on machine learning and knowledge discovery in databases (pp. 443-459). Cham: Springer Nature Switzerland.
10. Jena, B., Nayak, G. K., & Saxena, S. (2022). Convolutional neural network and its pretrained models for image classification and object detection: A survey. Concurrency and Computation: Practice and Experience, 34(6), e6767.
11. Sharma, S., & Guleria, K. (2022, April). Deep learning models for image classification: comparison and applications. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 1733-1738). IEEE.
12. Ye, T., Qin, W., Zhao, Z., Gao, X., Deng, X., & Ouyang, Y. (2023). Real-time object detection network in UAV-vision based on CNN and transformer. IEEE Transactions on Instrumentation and Measurement, 72, 1-13.
13. Ogas, E., Avila, L., Larregay, G., & Moran, D. (2019). A grasp detection method for industrial robots using a Convolutional Neural Network. IEEE Latin America Transactions, 17(09), 1509-1516.
14. Bae, H., Kim, G., Kim, J., Qian, D., & Lee, S. (2019). Multi-robot path planning method using reinforcement learning. Applied sciences, 9(15), 3057.
15. Jin, Z., Liu, L., Gong, D., & Li, L. (2021). Target recognition of industrial robots using machine vision in 5G environment. Frontiers in Neurorobotics, 15, 624466.
16. Wang, Y., Hong, K., Zou, J., Peng, T., & Yang, H. (2019). A CNN-based visual sorting system with cloud-edge computing for flexible manufacturing systems. IEEE transactions on industrial informatics, 16(7), 4726-4735.
17. Wang, K. J., & Lee, Y. X. (2023). Measuring defects in high-speed production lines—a three-phase convolutional neural network model. Measurement Science and Technology, 34(10), 105903.
18. Chen, H., Du, Y., Fu, Y., Zhu, J., & Zeng, H. (2023). DCAM-Net: A rapid detection network for strip steel surface defects based on deformable convolution and attention mechanism. IEEE Transactions on Instrumentation and Measurement, 72, 1-12.
19. Lv, H., Chen, J., Pan, T., Zhang, T., Feng, Y., & Liu, S. (2022). Attention mechanism in intelligent fault diagnosis of machinery: A review of technique and application. Measurement, 199, 111594.
20. Li, G., Shi, J., Luo, H., & Tang, M. (2013). A computational model of vision attention for inspection of surface quality in production line. Machine vision and applications, 24, 835-844.
21. Weimer, D., Scholz-Reiter, B., & Shpitalni, M. (2016). Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP annals*, *65*(1), 417-420.
22. Singh, S. A., & Desai, K. A. (2023). Automated surface defect detection framework using machine vision and convolutional neural networks. *Journal of Intelligent Manufacturing*, *34*(4), 1995-2011.
23. Staar, B., Lütjen, M., & Freitag, M. (2019). Anomaly detection with convolutional neural networks for industrial surface inspection. *Procedia CIRP*, *79*, 484-489.
24. Rajan, A. J., Jayakrishna, K., Vignesh, T., Chandradass, J., & Kannan, T. T. M. (2021). Development of computer vision for inspection of bolt using convolutional neural network. *Materials Today: Proceedings*, *45*, 6931-6935.
25. Xue, W., Xu, C., & Feng, Z. (2017). Robust visual tracking via multi-scale spatio-temporal context learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 2849-2860.