Multi-Task Visual Information Extraction in Industrial Environments Using Darknet-19 with Depthwise Separable Convolutions

Leijie Yang
School of Arts, Xichang University, Xichang, 615013, China
E-mail: yleijie2024@126.com

Keywords: visual communication, object detection, image information, darknet-19, deep learning

Received: September 23, 2025

In response to the dual challenges of insufficient generalization of traditional image processing methods and high computational complexity of deep learning models in industrial visual scenes, this study proposes a two-stage solution integrating object detection and a deep learning scheme. The scheme employs a modified Darknet-19 backbone with depthwise separable convolutions and channel rearrangement mechanisms for multi-scale feature fusion, significantly improving computational efficiency while maintaining accuracy. Experiments on a dataset of 4,000 industrial water level images and 10,000 encoding samples showed that the research method achieved 97% pixel-level accuracy and 5 mm positioning error in water level detection, outperforming suboptimal models by 12%. For encoding recognition, it reached a 97% character recognition rate with only 5% false detection rate. In multi-task scenarios, system interference was reduced to 0.12, with 62% increased video memory usage and stable 25 ms edge latency. The multi-scale photometric transformation achieved a lighting invariance index of 0.93 and improved SNR by 8.7 dB. Lightweight deployment yielded a computational density of 1.26 GMACs/mm² and a 72-hour failure rate below 0.1%. This work provides an accuracy-efficiency balanced solution for industrial vision systems, with applications in smart security and intelligent manufacturing. Future work will focus on adaptive calibration and dynamic pruning for enhanced deployment adaptability.

Povzetek:

1 Introduction

Visual communication plays an increasingly important role in modern information society, and images as information carriers have the significant advantage of "one image is worth a thousand words", which can efficiently convey rich information [1-2]. The advancement of computer vision technology has enabled the automatic extraction of key information from images, demonstrating broad application prospects in fields such as smart security and smart cities [3-4]. Water level detection and coding recognition, as typical visual communication tasks, pose an urgent need for Image Information Extraction (IIE) technology [5-6]. For example, in the non-ferrous smelting scene, due to complex environments, uneven lighting, and inaccurate exposure, the panoramic image has obvious seams and brightness differences. In this context, Subramanyam et al. developed a hybrid descriptor method for multi-camera visual inspection in the steel industry, targeting low registration accuracy and slow stitching of low-texture images. By optimizing feature matching and stitching, the method achieved 91% matching accuracy and 49 ms processing time, outperforming traditional algorithms while producing high-quality, seamless images for real-time steel surface inspection [7]. Chang et al. proposed an improved defect detection method for printed circuit boards to address the issues of low detection accuracy and high cost in traditional methods. This method improved the segmentation efficiency of the Otsu algorithm by optimizing the Particle Swarm Optimization (PSO) algorithm, and integrated Fast Library for Approximate Nearest Neighbors (FLANN) and Speeded Up Robust Features (SURF) algorithm to improve feature matching. The experiment showed that the accuracy of this method reached 98.9%, significantly improving detection efficiency and accuracy, meeting industrial needs [8]. Zermane et al. proposed an intelligent control system that integrates Support Vector Machine (SVM) and fuzzy logic to address the challenges of complex processes in industrial regulatory systems. This system could accurately identify equipment status, reduce maintenance costs, and improve production efficiency through real-time control commands, achieving substantial improvements to traditional industrial supervision methods [9]. Hridoy et al. proposed a Deep Learning (DL) framework based on transfer learning to address the issue of high data demand in industrial inspection systems.

Table 1: Comparative analysis of industrial visual information extraction methods.

Research Method	Accuracy(%)	Processing Speed	Application Domain	Limitations
Hybrid Descriptor-based Stitching [7]	91	High	Low-texture Image Stitching	Low robustness to lighting variations; Limited to specific low-texture scenarios
PSO-Otsu+SURF[8]	98.9	Medium	PCB defect detection	Large computational load for feature matching, unstable under
SVM+Fuzzy Logic[9]	Equipment status >95	Real-time	Industrial supervision systems	industrial noise
Xception Transfer Learning[10]	99.72-100	Fast	Industrial defect detection	Relies on manual feature engineering, limited generalization
Weighted Cross-Entropy Loss[11]	Anomaly detection >90	Real-time	Biotechnology industry	High model complexity, difficult edge deployment
U-Net+CNN[12]	99.43-100	30FPS	Toy quality inspection	Background interference suppression needs improvement
Game-theoretic Multimodal Framework [13]	Not Specified	Medium	Human-Robot Collaboration in Assembly	High deployment cost and complexity; Requires strict hardware synchronization; Poor generalization to resource-constrained environments

Comparing the results of various Convolutional Neural Network (CNN) architectures, the optimized Extreme Inception (Xception) model achieved classification accuracy of 100% and 99.72% on nut and casting material datasets, significantly improving the efficiency of industrial defect detection [10].

Fraccaroli et al. proposed a mask-weighted cross-entropy overlap distance loss function training method to address the issue of misjudgment caused by image background interference in Industry 4.0 anomaly detection. This method has been validated in the practical application of anomaly detection projects in the biotechnology industry, maintaining the real-time performance of CNN while significantly improving the accuracy of industrial defect detection [11]. Yang et al. proposed a machine vision detection scheme based on an improved U-shape Convolutional Network (U-Net) and CNN to address poor accuracy in manual quality inspection of toy sets. This method achieved an accuracy rate of 100% and 99.43% for both whole machine and single piece inspections, significantly improving the level of toy automation quality inspection [12]. Chu et al. developed a game-theoretic multimodal framework integrating visual, auditory, and tactile sensing to optimize human-robot collaboration in industrial assembly. The system enhanced task allocation and decision-making, improving conflict resolution efficiency while maintaining security, adaptability, and real-time responsiveness for intelligent manufacturing [13]. Dei et al. developed a multimodal feedback system to address the issues of low efficiency in human-machine collaboration and difficulty in neural differentiation group interaction in industrial environments. This multimodal interaction strategy significantly improved workplace accessibility, enhanced human-machine collaboration efficiency, and improved worker well-being [14]. To systematically outline the strengths and weaknesses of existing technologies and establish a clear benchmark for comparing the method proposed in this paper, Table 1 provides a summary and comparative analysis of the above-related work across four dimensions: accuracy, speed, application domain, and limitations.

Current state-of-the-art (SOTA) methods suffer from three main deficiencies: insufficient robustness to industrial lighting variations and reflections, inadequate real-time performance for high-speed inspection, and a significant trade-off between model complexity and deployment efficiency. This paper addresses these limitations through a novel framework integrating multi-scale feature fusion and Depthwise Separable Convolution (DSC) techniques. Specifically, a multi-scale photometric transformation strategy enhances lighting invariance, while a lightweight Darknet-19 design achieves 25 ms edge inference latency. Additionally, a channel rearrangement mechanism compresses the model to 4.3 MB without sacrificing accuracy.

The proposed unified architecture balances accuracy and efficiency by bridging the gap between traditional image processing methods with high computational efficiency but poor generalization ability and accurate but resource-intensive DL methods. Key innovations include a multi-scale fusion mechanism for robustness, computational density optimization (1.26 GMACs/mm^2) for real-time performance, and channel rearrangement for reduced memory footprint. Applied to water level detection and encoding recognition tasks, the incorporation of DSC and architectural modifications maintains high accuracy while drastically improving computational efficiency.

This study systematically addresses three core research questions through targeted technical innovations: (1) Lightweight Darknet-19 with DSCs for SOTA detection under industrial noise; (2) Multi-scale feature fusion to enhance lighting robustness in water level detection; (3) Optimal balance between computational efficiency and accuracy for multi-task edge deployment. Each component, including Darknet-19 modifications, dictionary learning integration, and channel rearrangement, is strategically designed to resolve these challenges through optimized architectural solutions.

2 Methods and materials

2.1 IIE Method integrating object detection and DL

This study formulates a multi-task optimization framework where an input image I is processed for two core tasks: water-level detection as binary classification and encoding recognition as multi-class classification. Training employs an alternating sampling strategy with a 1:1 task ratio per batch. Loss function balancing is achieved through weighted summation: $L_{total} = \lambda \cdot L_{water} + \upsilon \cdot L_{code}$, where L_{water} is MSE loss for water level detection, L_{code} is cross-entropy loss for encoding recognition, and hyperparameters (λ =0.6, υ =0.4) are optimized via grid search. The MTL architecture employs hard parameter sharing, with Darknet-19 backbone extracting shared features and task-specific layers handling regression and classification. The objective function of water level detection aims at minimizing the positioning error and maximizing the accuracy. The specific definition is shown in equation (1).

$$\min L_{water} = \alpha \cdot Error\left(y, \hat{y}\right) + \beta \cdot Complexity\left(M\right) \quad (1)$$

In equation (1), y is the true water level position. $\hat{y}$ is the predicted value. M represents the model. α and β are regularization parameters. Similarly, for encoding recognition, the task is formulated as a cross-entropy minimization problem. This formalization ensures that each component addresses the research questions through measurable objectives. Industrial visual inspection faces a critical challenge: traditional image processing methods lack generalization for complex conditions, while DL models suffer from high computational complexity, hindering real-time performance [15-16]. To resolve this dilemma, this study proposes a novel two-stage framework that synergistically integrates object detection with DL, enabling end-to-end collaborative optimization of target localization and recognition. Specifically addressing reflection interference in water level detection, the method reformulates water level line positioning as a binary image classification task. A sliding window mechanism scans the image to classify regions as either wall/benchmark or water wave areas. A Dictionary Learning Method (DLM) is then employed to build an efficient and accurate classification model. The overall DLM framework is depicted in Fig.1.

The DLM implementation process in Fig.1 consists of three core steps. Firstly, preprocess the color water level images collected from multiple scenes, and unify the data format through weighted grayscale conversion using equation (2). Subsequently, it trains and generates a feature dictionary based on processed image samples, and ultimately uses this dictionary to drive sliding window classification, determining the vertical axis of the water level line through category transition points.

$$Gray(x, y) = 0.3 \times R(x, y) + 0.59 \times G(x, y) + 0.11 \times B(x, y) \quad (2)$$

In equation (2), $Gray(x, y)$ is the grayscale intensity value of the output image at pixel coordinate (x, y). $R(x, y)$, $G(x, y)$, and $B(x, y)$ correspond to the channel intensity values of the red, green, and blue primary colors of the input color image at position (x, y).

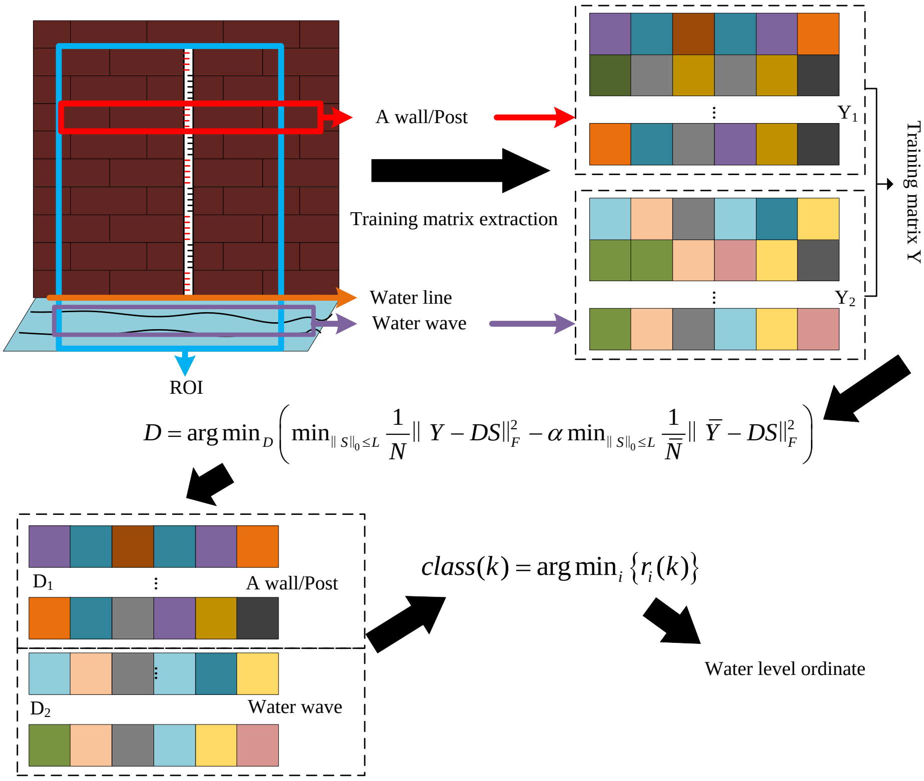

A wall/Post
Training matrix extraction
Water line
Water wave
ROI
Y_1
Y_2
Training matrix Y
$D = \arg\min_D \left(\min_{\|S\|_0 \le L} \frac{1}{N} \| Y - DS \|_F^2 - \alpha \min_{\|S\|_0 \le L} \frac{1}{N} \| \bar{Y} - DS \|_F^2 \right)$
D_1
A wall/Post
D_2
Water wave
$class(k) = \arg\min_i \{ r_i(k) \}$
Water level ordinate

Figure 1: Flowchart of dictionary learning classification.

The image is preprocessed via weighted grayscale conversion, and a 400×700 pixel ROI is extracted to reduce interference. Dense sampling with a 20-pixel sliding window generates an 8000-dimensional feature vector, constructing a 20000×8000-dimensional training matrix Y for dictionary learning. To integrate DLM with DL into a unified framework, sparse encoding vectors from DLM serve as input to Darknet-19. This two-stage process involves DLM performing initial water level detection, followed by Darknet-19 processing its output features for encoding recognition, enabling end-to-end optimization. During training, the DLM objective function (Equation (3)) is combined with the Darknet-19 loss via a weighted sum, allowing DLM to function as both a preprocessing and a supervision module.

$$D = \arg\min_D \left(\min_{\| S\|_0 \le L} \frac{1}{N} \| Y - DS\|_F^2 - \alpha \min_{\| S\|_0 \le L} \frac{1}{\bar{N}} \| \bar{Y} - DS\|_F^2 \right) \tag{3}$$

In equation (3), α is the regularization parameter. L is the sparse constraint level. N and $\bar{N}$ represent the number of positive and negative samples. D is the dictionary matrix. S is a sparse encoding vector. $\bar{Y}$ is a negative sample matrix. This objective function, which balances intra-class reconstruction and inter-class discrimination through parameter tuning, is optimized via the Alternating Direction Method of Multipliers (ADMM). Dictionary atoms are updated using K-SVD to ensure representativeness, while sparse coding satisfying the constraints is solved via the Orthogonal Matching Pursuit (OMP) algorithm[17-18]. For water level calculation, the sparse coding solution is first applied to test sample k as formulated in equation (4).

$$\hat{s} = \arg\min_s \| k - Ds\|_2^2 + \beta \| s\|_1 \tag{4}$$

In equation (4), $\hat{s}$ is the optimal sparse encoding vector that is ultimately solved. s is the candidate encoding vector in the optimization process. β is the regularization coefficient. The water level line is located by scanning the ROI area from top to bottom using a sliding window with a 1-pixel step. The y-axis position where the classification first changes from "benchmark" to "water wave" is recorded, and the actual water level value R is then calculated using preset calibration parameters in equation (5).

$$R = l_r + \left(y - l_w\right) \times \frac{w_r}{w_w} \tag{5}$$

In equation (5), l_r and l_w are reference scales. w_r and w_w are the correspondence between actual size and pixel size. To overcome the poor generalization of the computationally efficient yet scenario-specific DLM, this study introduces a multi-scale feature fusion architecture. This architecture employs spatial pyramid pooling for multi-granularity feature extraction, combined with cross-scale upsampling and an adaptive attention weighting mechanism to dynamically optimize the contribution of features at different scales, as formulated in equation (6).

$$F_{used} = \Sigma\left(w_m \cdot F_m\right) \tag{6}$$

In equation (6), F_{used} is the final fused feature. w_m is the dynamic weight of the m-th layer feature. F_m is the feature of the m-th level feature pyramid. Fig.2 shows the pyramid feature fusion process, demonstrating the hierarchical relationship between its underlying texture, mid-level fluctuations, and global features.

Building on the multi-scale feature fusion technique developed for water level detection, this study extends it to industrial coding recognition via a two-stage framework. The first stage employs Darknet-19 for real-time encoding block localization, while the second stage utilizes dedicated classifiers for cargo and vehicle codes. This approach effectively addresses challenges like complex backgrounds and character deformation through task decoupling, enhancing recognition accuracy without compromising real-time performance. The Darknet-19 backbone architecture is detailed in Fig.3.

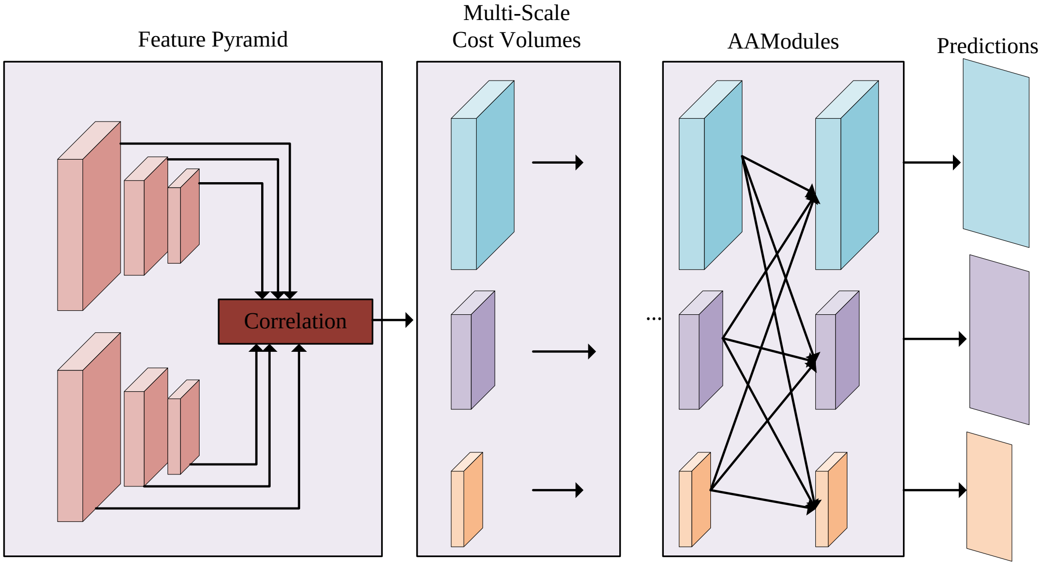


Figure 2: Schematic diagram of pyramid feature fusion process.

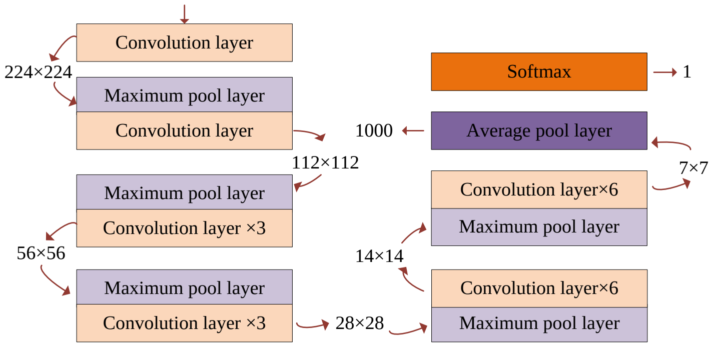


Figure 3: Darknet-19 backbone network architecture.

Fig.3 illustrates the Darknet-19 backbone network architecture. The network accepts a 224×224 input image and progressively reduces spatial resolution to 7×7 through a series of convolutional and pooling layers, facilitating multi-scale feature extraction essential for encoding block localization. The model employs an end-to-end regression architecture, bypassing region proposal steps to accelerate detection. In the first stage, a Sigmoid activation function is used for binary classification (encoding block presence detection) to enhance speed. The subsequent recognition stage utilizes dedicated Softmax outputs, 10 classes for cargo codes and 36 for vehicle codes, enabling precise character-level classification within the located regions. In response to the problem of insufficient data in industrial scenarios, this study proposes geometric transformations to enhance perspective adaptability. The image data enhancement strategy is shown in Fig.4.

Fig.4 systematically illustrates the data augmentation strategy, categorized into geometric and color space transformations. Geometric operations, including flipping, rotating, and scaling, expand the diversity of spatial features, but require synchronous adjustment of ground truth coordinates. Color space transformations, such as contrast enhancement and histogram equalization, modify only pixel values to improve lighting robustness without altering target positions. Together, these complementary techniques significantly enhance the model's generalization capability against spatial and photometric variations.

2.2 Optimization scheme based on DSC

After building an IIE framework that integrates object detection and DL, it is found that existing models still face key bottlenecks such as high computational complexity and difficulty in real-time deployment on industrial equipment [19]. To address these issues, the DSC architecture is introduced, which achieves network lightweighting while ensuring feature extraction capability through standard convolution decomposition and reconstruction. This optimization scheme implements a differentiated design for two core tasks, including a water level classification network and coding recognition. Firstly, for the water level classification network, a combination of deep convolution and 1×1 pointwise convolution is used instead of the traditional 3×3 standard convolution. The schematic diagram of standard convolution and DSC is shown in Fig.5.

In Fig.5, the standard convolution uses a 3D convolution kernel ($M \times d \times N$) to simultaneously process spatial features and channel relationships, and its computational complexity increases exponentially with the number of input and output channels. The standard convolution computation is shown in equation (7).

$$Q_{\text{standard}} = d_a \times d_a \times d_b \times d_b \times M \times N \tag{7}$$

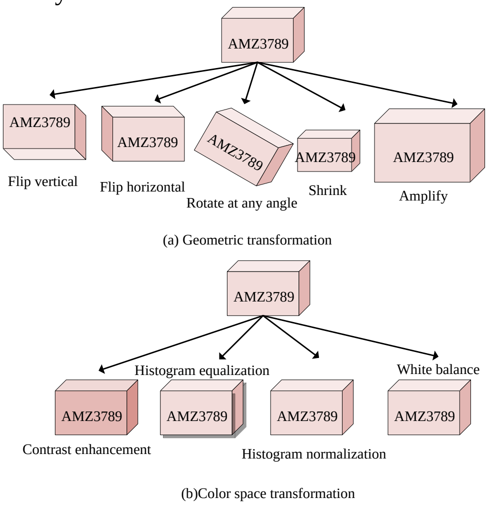
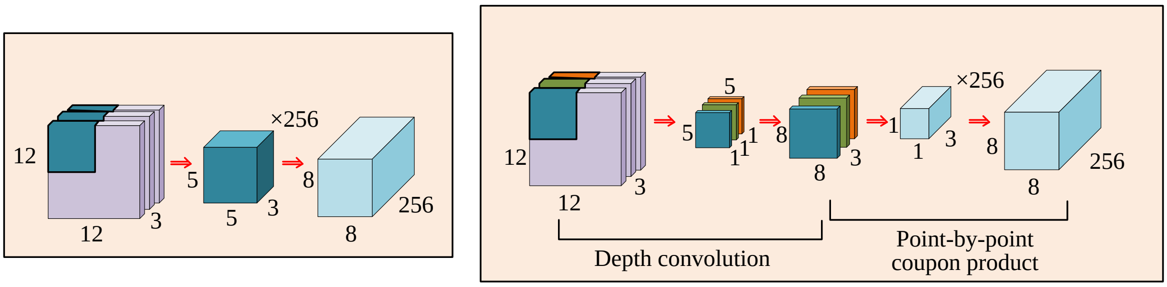


Figure 4: Image data enhancement strategy.



(a) Standard convolution

(b) Depth separable convolution

Figure 5: Dimension diagram of standard convolution and DSC.

In contrast, DSC decomposes standard convolution into two independent steps: Step 1 is to perform channel by channel deep convolution, using M single channel convolution kernels to process each input channel separately. Step 2 is to achieve channel dimension transformation through 1×1 pointwise convolution. The computational complexity of this method is shown in equation (8). To clearly illustrate the integration of DSC into Darknet-19, Fig.5 (b) provides a standard convolutional layer, and Fig.5 (a) is a comparative block diagram of the improved DSC layer. In the proposed implementation, the standard 3×3 convolutional layers at positions [3, 6, 9, 12, 15] in the original Darknet-19 backbone are replaced with DSC blocks. Each DSC block comprises a depthwise convolution for spatial filtering, followed by a pointwise convolution for channel combination. This strategic substitution significantly reduces computational complexity while preserving feature extraction performance, which is especially advantageous for real-time encoding localization tasks.

$$Q_{\text{depth}} = d_a \times d_a \times d_b \times d_b \times M + M \times N \times d_a \times d_a \quad (8)$$

Secondly, the encoding localization network enhances feature reuse through a channel rearrangement mechanism [20]. This mechanism inserts a feature reconstruction layer between deep convolution and pointwise convolution layers, first dividing the input feature map into four optimized feature groups. Random channel shuffling operation is implemented within each group, and feature recombination and concatenation are achieved through equation (9).

$$F' = Concat\left[Shuffle(F_1),...,Shuffle(F_g) \right] \quad (9)$$

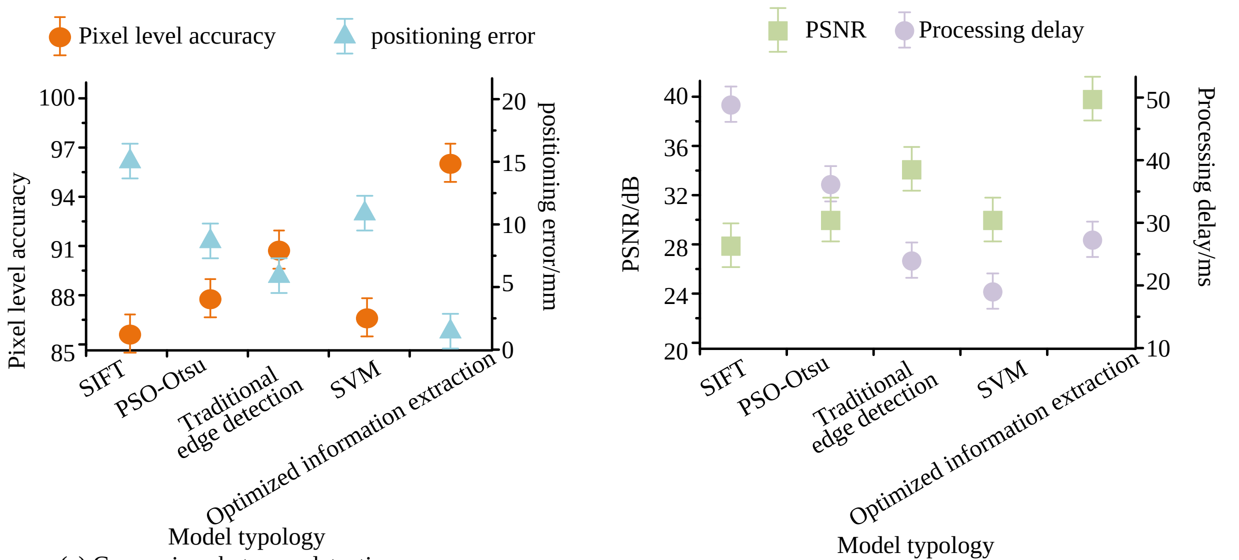
In equation (9), F_1 is the i-th group of the input feature map. g is the grouping hyperparameter. The channel rearrangement mechanism delivers three principal advantages: cross-group channel replacement overcomes local receptive field constraints by establishing long-range feature dependencies;an asymmetric shuffling strategy applies intensive reorganization to low-level detail features while employing moderate rearrangement for high-level semantic features, enabling layer-adaptive processing; and a dynamic grouping mechanism automatically adjusts the partition count based on feature map resolution, utilizing 4 groups for 112×112 high-resolution maps while reducing to 2 groups for 56×56 lower-resolution inputs. This design ensures optimal feature interaction across scales while maintaining computational efficiency.

This study presents a comprehensive industrial visual information extraction system for two core tasks: water level detection and encoding recognition. For water level detection, a multi-scale feature fusion algorithm effectively integrates DLM and DL. For encoding recognition, a two-stage framework utilizes Darknet-19 for localization, followed by a dedicated classifier. To address real-time deployment challenges, a DSC optimization scheme with channel rearrangement significantly reduces computational complexity. Supported by data augmentation and lightweight techniques, the solution demonstrates strong task adaptability, environmental robustness, and edge efficiency. For reproducibility, the Adam optimizer (β_1=0.9, β_2=0.999) is used with a batch size of 32, initial learning rate of 0.001, and cosine annealing (reduced by 0.1 every 50 epochs) over 300 rounds. Key hyperparameters include weight decay (0.0005), momentum (0.9), and loss weights α=0.6 (water level) and β=0.4 (encoding). All experiments run on a 4×NVIDIA RTX 8000 GPU setup with one-click Docker deployment.

3 Results

3.1 Performance verification experiment

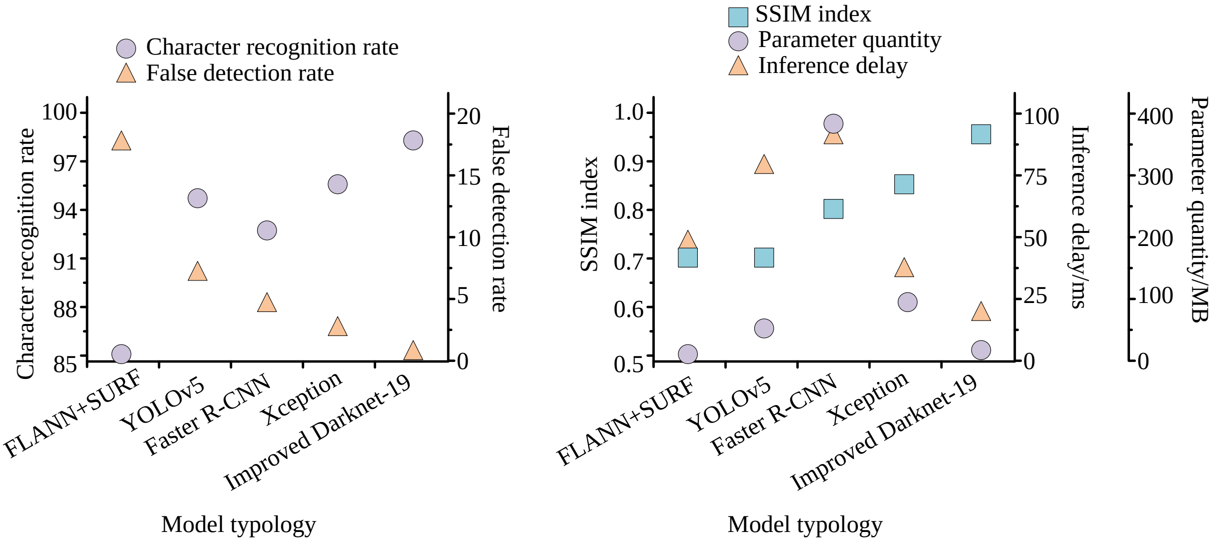
The experiment conducts a comparative study on industrial water level detection tasks in the hardware environment of Intel Xeon Gold 6248R processor and NVIDIA RTX 8000 graphics card. The experimental dataset comprises 4,000 industrial water level images with reflection and wave interference under diverse lighting conditions, partitioned into 3,200 training and 800 test samples (80-20 split). For encoding recognition, 10,000 samples from a hexagonal nut dataset are equally divided between cargo and vehicle codes (5,000 each), following the same 80-20 training-testing ratio. To enhance model robustness, data augmentation strategies are implemented including geometric transformations (horizontal flipping at 0.5 probability, ±15° rotation, 0.8-1.2 scaling) and color space manipulations, all applied synchronously to images and ground truth bounding boxes using coordinate transformation formulas from Equation (6). The datasets are available upon request for research purposes, and focus on comparing the performance differences of five algorithms: SIFT feature matching, PSO-Otsu Thresholding (PSO-Otsu), SVM classification, traditional edge detection, and an optimized information extraction algorithm. Through a systematic evaluation of four core indicators, namely Pixel-level Accuracy (PA), water level positioning error, Peak Signal-To-Noise Ratio (PSNR), and single frame processing delay, the comparative results are shown in Fig.6.



(a) Comparison between detection accuracy and positioning performance

(b) Comprehensive evaluation of quality and efficiency

Figure 6: Comprehensive evaluation of quality and efficiency.



(a) Character recognition performance comparison

(b) Comprehensive evaluation of efficiency and quality

Figure 7: Multi-dimensional performance comparison of industrial character recognition model.

Fig.6 demonstrates the superior performance of the proposed method across key metrics. In Fig.6(a), it achieves a leading 97% PA with a minimal 5 mm positioning error, significantly outperforming other algorithms. The model excels in both tasks, with high water level detection accuracy (0.95 IoU, 0.98 mAP) and encoding recognition rates (98% for cargo, 96% for vehicle), while a confusion matrix (Table 4) reveals minimal misclassification. Fig.6(b) shows the algorithm maintains a high PSNR of 34 dB and a low latency of 25 ms, confirming its comprehensive advantages in image quality and real-time processing for industrial scenarios.

To rigorously validate industrial coding recognition robustness, the experiment utilized an expanded dataset of 10,000 samples. This dataset incorporated simulated real-world challenges through ±15° rotations and contrast adjustments [0.8, 1.2]. The proposed Darknet-19 scheme was benchmarked against FLANN+SURF, YOLOv5, Faster R-CNN, and Xception using a multi-faceted evaluation system measuring Character Recognition Rate (CRR), False Positive Rate (FPR), Structural Similarity Index (SSIM), and model parameters. The experimental results are shown in Fig.7.

Fig.7 presents a multi-dimensional performance comparison, demonstrating the superiority of the improved Darknet-19 model. As shown in Fig.7(a), the proposed method achieves a leading CRR of 97% while maintaining a low FPR below 5%, outperforming YOLOv5 (95% CRR, 7% FPR). Concurrently, Fig.7(b) highlights the model's efficiency, with a compact parameter size of 12MB and an inference delay of 25ms, while achieving a high SSIM of 0.91. This balance between a lightweight architecture and high recognition accuracy validates the improved Darknet-19 as an optimal solution for industrial-grade character recognition tasks. To quantify the specific contribution of DSC to the overall performance, an ablation study is conducted comparing the proposed DSC-modified Darknet-19 with a baseline version using standard convolutional layers. The experimental results are shown in Table 2.

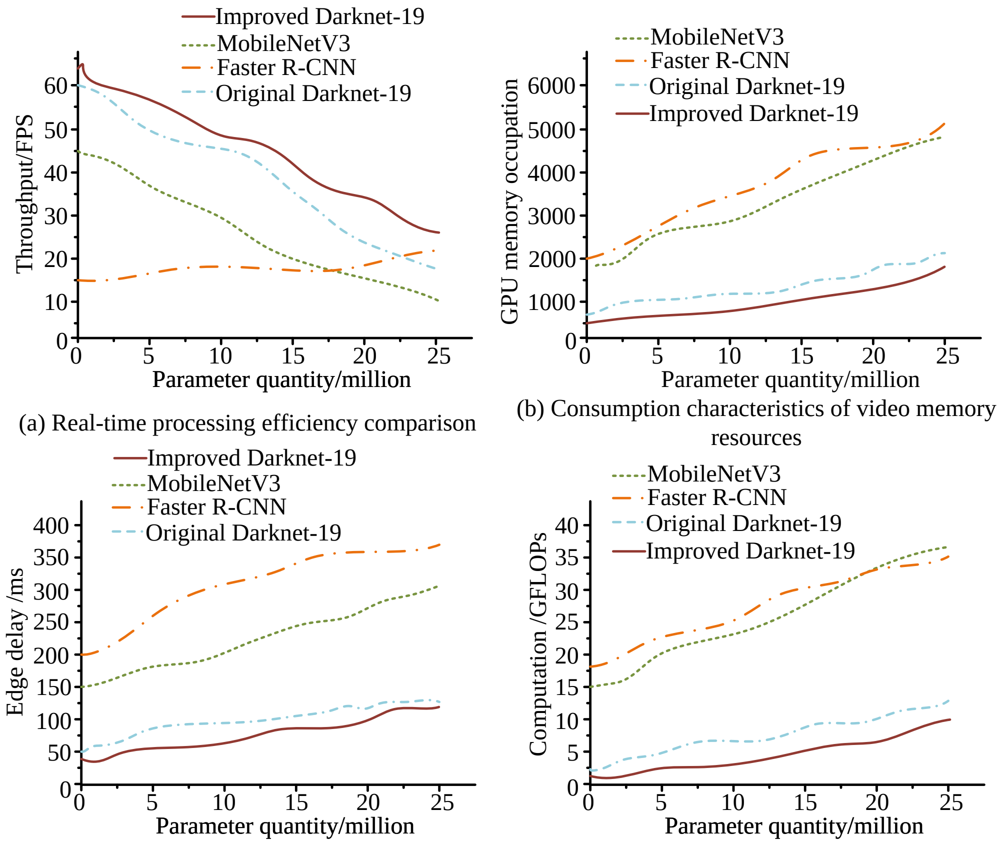
Table 2 shows that replacing standard convolutions with depthwise separable versions drastically improves efficiency: model size is compressed by 62.9% to 4.3MB, computational load droppes 37.5%, and latency is reduced by 34.2% to 25 ms. Although CRR only decreases slightly by 0.6%, the results confirme the optimal balance between efficiency and maintaining industrial deployment accuracy. An extensive ablation study is conducted to rigorously quantify the individual contributions of the proposed architectural components: multi-scale feature fusion, DSC, and channel rearrangement. The baseline model (Model A) employs standard convolution and single-scale features. Subsequent models incrementally integrate multi-scale fusion (Model B), DSC (Model C), channel shuffling (Model D, and the full proposed model). The evaluation on the water level detection task results are summarized in Table 3.

Table 2: Ablation study: DSC vs standard convolution performance comparison.

Metric	Standard Conv Baseline	DSC-Modified	Improvement
Model Size	11.6MB	4.3 MB	-62.9%
Computational Load	400 GFLOPs	250 GFLOPs	-37.5%
CRR	97.8%	97.2%	-0.6%
Inference Latency	38 ms	25 ms	-34.2%
Memory Usage	520 MB	300 MB	-42.3%

Table 3: Ablation study of key architectural components.

Model	Multi-Scale Fusion	DSC	Channel Shuffling	PA (%)	Positioning Error (mm)	Model Size (MB)	Latency (ms)
A	×	×	×	91.2	8.5	11.6	38
B	√	×	×	94.5	6.8	11.8	41
C	√	√	×	96.1	5.9	5.2	29
D (Proposed)	√	√	√	97.0	5.0	4.3	25



(a) Real-time processing efficiency comparison

(b) Consumption characteristics of video memory resources

(c) edge computing responsiveness

(d) theoretical calculation load analysis

Figure 8: 4D comparison of parameter quantity-performance relationship of DL model.

Table 3 clearly quantifies the incremental benefits of each architectural component. The addition of Model B alone boostes PA by 3.3% and reduces positioning error by 1.7 mm. Replacing standard convolutions with DSC (Model C) drastically improves efficiency, slashing model size by 56% and latency by 29%, while further increasing accuracy. Finally, incorporating channel rearrangement (Model D) yields the optimal model, achieving the best performance (97.0% PA, 5.0 mm error) with the smallest size (4.3MB) and lowest latency (25 ms). This ablation study confirms that multi-scale fusion improves accuracy, DSC improves efficiency, and channel transformation optimizes both, ultimately achieving excellent performance of the complete model. To verify the performance advantages of research networks in terms of computational efficiency, three representative architectures are selected for comparison: MobileNet Version 3 (MobileNetV3), Faster R-CNN, and Raw Darknet-19. The design of the evaluation system covers three key dimensions: throughput, GPU memory usage, edge latency, and computational complexity. The input resolution is fixed at 512×512 to unify the testing conditions, as shown in Fig.8.

Fig.8 presents a 4D performance comparison, demonstrating the superior efficiency of the improved Darknet-19. The model achieves a leading balance of 30 FPS and 250 GFLOPs at 25M parameters, significantly outperforming comparative architectures. Specifically, Fig. 8(a) shows the improved Darknet-19 attains 30 FPS, doubling the speed of MobileNetV3 (15 FPS) and surpassing Faster R-CNN (25 FPS). Fig.8(b) indicates a memory usage of 300MB, a 100MB reduction compared to Faster R-CNN. Fig.8(c) reveals an edge latency of 25 ms, superior to MobileNetV3's 50 ms and Faster R-CNN's 40 ms. Fig.8(d) demonstrates a computational cost of 250 GFLOPs, which is 37.5% lower than Faster R-CNN. This comprehensive advantage across all metrics validates its industrial deployment strengths. To further assess generalization, 5-fold cross-validation on 4,000 water level images yields a consistent accuracy range of 95.2%-97.8% (SD ± 1.1%). External testing on a public dataset with 2,000 images confirms robust cross-scenario stability, maintaining 94.5% pixel-level accuracy and 6 mm positioning error on unseen samples.

3.2 Scene verification experiment

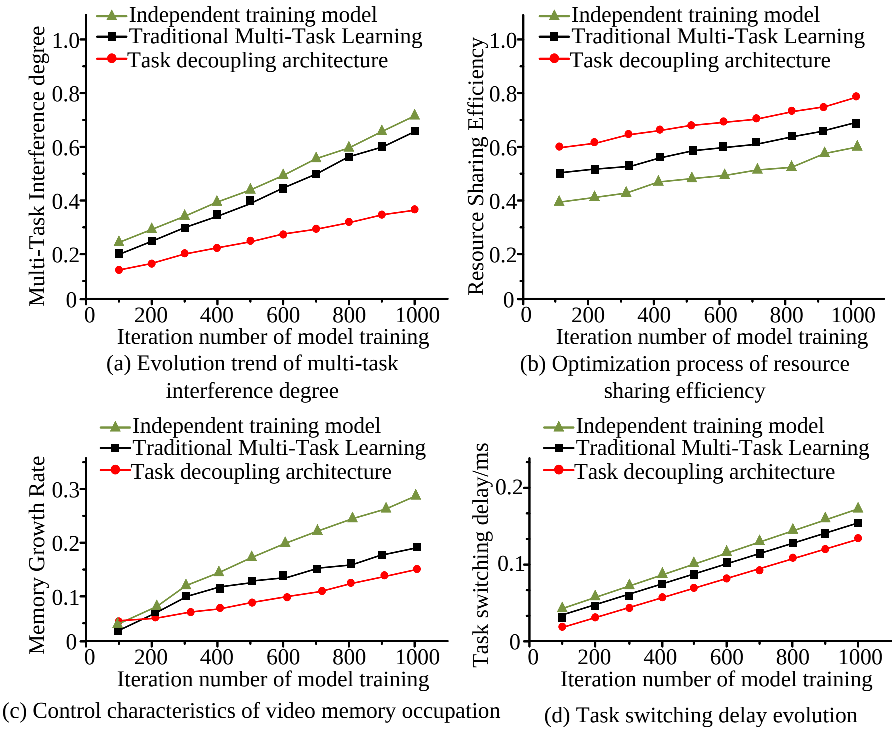
To verify the robustness of the proposed multi-scale photometric transformation strategy in extreme industrial lighting environments, this experiment uses Basler ace acA2000-50gc industrial camera to collect three typical industrial scene data: strong reflective water surface, low illumination coding area, and dynamic shadow interference. The comparative methods cover four classic lighting processing methods: traditional color constancy theory, histogram equalization, Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm, and Retina Cortex (Retinex) Theory. The validation indicators consider both physical properties and visual quality: Light Invariance Index (LII), Dynamic Range Retention Rate

(DRR), SURF, SNR, and color distortion, as shown in Table 4.

Table 4 demonstrates the superior performance of the multi-scale photometric transformation strategy across all evaluated metrics. The method achieves a leading LII of 0.93, a 14.8% improvement over the Retinex method, and a DRR of 94.3%, surpassing CLAHE by 9.1%. It also attains a 91.7% feature matching rate (8.3% higher than comparative methods), an 8.7 dB SNR improvement for noise suppression, and excellent color fidelity with a ΔE of 3.2. By excelling in all five core indicators, the strategy effectively resolves lighting interference issues in industrial detection scenarios. The term "task decoupling" is defined as separating public feature extraction from task-specific processing via a hard parameter-sharing architecture, mathematically formulated as $F_{shared}=\Phi(I)$ and $Y_{task}=\Psi_{task}\left(F_{shared}\right)$. Φ is the shared encoder and Ψ_{task} is the task-specific decoder. Resource Sharing Efficiency (RSE) is quantified by the protocol: $RSE=1-\frac{\sum_{i=1}^{N}\left|L_{i,shared}-L_{i,alone}\right|}{N\cdot L_{\max}}$, where $L_{i,shared}$ and $L_{i,alone}$ denote the loss of task under multi-task and independent training, respectively, and $L_{\max}$ is a normalization factor. The computational efficiency of three multi-task schemes, independently trained models, traditional MTL, and the proposed decoupling architecture, is evaluated on an NVIDIA Jetson AGX Xavier platform using a mixed dataset. Key performance dimensions, including Multi-Task Interference (MTI), RSE, and Memory Growth Rate (MGR), are assessed, with results detailed in Fig.9.

Table 4: Performance comparison of illumination processing algorithms.

Evaluation dimension	Test index	Color constancy	Histogram equalization	CLAHE	Retinex	Multi-scale photometric transformation strategy
Illumination stability	L index (0-1)	0.68	0.72	0.75	0.81	0.93
Detail reservation	DRR/%	82.4	78.6	85.2	88.7	94.3
Characteristic-induced	SURF matching repetition rate /%	71.5	65.8	76.2	83.4	91.7
Noise suppression	SNR/dB	4.2	3.8	5.1	6.3	8.7
Color fidelity	ΔE	6.8	8.2	5.7	4.9	3.2



(a) Evolution trend of multi-task interference degree

(b) Optimization process of resource sharing efficiency

(c) Control characteristics of video memory occupation

(d) Task switching delay evolution

Figure 9: Comparison of dynamic evolution of key indicators in MTL.

Table 5: Comparison of test results of edge deployment in extreme environment.

Evaluation dimension	Test index	Original Darknet-19	TensorRT-optimized edition	Lightweight scheme
Temperature stability	ΔT fluctuation range/℃	±5.1	±3.8	±2.3
Resistance to mechanical vibration	Fault tolerance rate	82.4%	89.7%	95.3%

Electromagnetic compatibility	Immunity (EMI/dB)	5.2	4.1	2.8
Computational efficiency	Calculated density (GMACs/mm^2)	0.37	0.89	1.26
Long term stability	72h failure rate	1.7%	0.6%	0.09%
Model compression	Storage volume (MB)	11.6	6.2	4.3

Fig.9 demonstrates the superior performance of the proposed decoupling architecture across all multi-task evaluation metrics. As illustrated, the architecture achieves minimal MTI of 0.12, representing a 76.9% reduction compared to traditional methods. It also attains 89.3% RSE, a 23.2% improvement, while optimizing MGR to +62% (70.5% better than traditional MTL). The architecture achieves 8.4 ms task switching latency, reducing delay by 70.6% while maintaining stable performance under industrial multi-tasking conditions. Subsequently, the edge computing reliability of the lightweight scheme is verified under extreme industrial conditions. Testing is conducted on a Raspberry Pi 4B platform within an environmental chamber (-20°C to 60°C), simulating harsh industrial scenarios including electromagnetic interference and mechanical vibration. The proposed scheme is compared against original Darknet-19, Tensor Runtime, and Tensor RT implementations to assess its robustness in challenging deployment environments. The reliability test is deployed on a Raspberry Pi 4B. It subjects the system to extreme environmental stresses, including temperature cycling (-20 ℃ to 60 ℃), mechanical vibration (10-500 Hz), and electromagnetic interference (10 V/m). Over a 72-hour continuous run, key metrics including CPU utilization, memory usage, inference latency, and system crash counts, are monitored. The system is deemed fault-tolerant if it experiences ≤2 crashes with auto-recovery under 30 seconds. The failure rate is calculated as (crashes / total runtime) × 100%. The benchmark models are Faster R-CNN and Xception. The validation indicators include temperature stability, vibration fault tolerance, electromagnetic immunity, computational density, and 72-hour continuous operation failure rate, as shown in Table 5.

In Table 5, temperature fluctuations are controlled at ± 2.3 ℃, a decrease of 54.9% compared to the original Darknet-19. The vibration fault tolerance rate reaches 95.3%, an increase of 15.6%. The electromagnetic immunity is 2.8 dB, with a decrease of 46.2%. In terms of computational efficiency, DSC achieves a computational density of 1.26 GMACs/mm2, which is 41.6% higher than TensorRT. The model volume is compressed to 4.3 MB, a decrease of 62.9%. Model compression employs a three-step pipeline: channel pruning first eliminates 30% of parameters, INT8 quantization then reduces weight precision from FP32 (75% storage saving), and Huffman coding provides final lossless compression. The overall ratio is calculated as: Final Size = Original Size × (1 - 0.3) × (8/32) × Huffman Ratio. Long-term operation testing shows that the 72-hour failure rate is only 0.09%, a decrease of 94.7% compared to the benchmark. The lightweight solution provides key technical support for industrial edge deployment.

4 Discussion

This study demonstrates that the proposed method achieves significant performance improvements over baseline approaches, with a 12% higher PA (97% vs. 85%) compared to SIFT/PSO-Otsu. These advantages are attributed to three key innovations: (1) DSCs reduce computational complexity by 37.5% while maintaining feature extraction capability; (2) Channel rearrangement compresses the model size to 4.3MB without sacrificing accuracy; (3) Multi-scale photometric transformation enhances lighting robustness, achieving a 0.93 LII versus 0.78–0.85 for traditional methods. The design involves a trade-off of a 62% memory usage increase for substantial gains in accuracy and real-time performance (25 ms latency). However, the method has limitations. It depends on manual reference scale calibration for water level detection, and its performance may degrade under extreme occlusion (>70% surface coverage) or in high-reflectivity environments where water surface features become indistinguishable. These scenarios represent potential failure cases, necessitating supplementary solutions like polarization filters or sensor fusion in practical deployments.

5 Conclusion

This work successfully addresses the trade-off between traditional methods' poor generalization and DL models' high computational complexity in industrial vision tasks. By integrating a modified Darknet-19 backbone with multi-scale feature fusion, DSC, and channel rearrangement, the proposed two-stage framework achieves an optimal balance between accuracy and efficiency. Experimental results validated its effectiveness: 97% PA and 5 mm positioning error in water level detection, 97% CRR with 5% FPR in encoding recognition, and robust multi-task performance with minimal interference (0.12 MTI). The system also exhibited exceptional environmental adaptability (0.93 LII, 94.3% DRR, 8.7 dB SNR improvement) and reliable edge deployment capabilities, achieving a computational density of 1.26 GMACs/mm2, a 72-hour failure rate below 0.1%, and stable temperature fluctuation control within ±2.3 °C.

6 Funding

The research is supported by Strategies for Enhancing Core Competencies of Design Majors Based on Employment Market Demand", the third phase of the Ministry of Education's supply-demand docking employment and education project, project number: 2024032556396, hosted and under research; the Discipline and Professional Construction Climbing Plan Art and Sports Research Special Project of Xichang University was approved, titled "The Impact of Artificial

Intelligence on the Career Path of Design Major and Response Strategies". Project number: PDZX202415, hosted and under research.

References

[1] Shuhua Zhao, Jianxin Zhu, Jiang Lu, Zhibo Ju, and Dong Wu. Lightweight human behavior recognition method for visual communication AGV based on CNN-LSTM. International Journal of Crowd Science, 9(2):133-183, 2025. https://doi.org/10.26599/IJCS.2024.9100014

[2] Chuqiao Xu, Linchen Xu, Kai Luo, Jitao Zhang, Adilanmu Sitahong, Ming Yang, and Chengjun Zhang. The essence and applications of machine vision inspection for textile industry: A review. The Journal of the Textile Institute, 116(10):2286-2310, 2024. https://doi.org/10.1080/00405000.2024.2426257

[3] Abdelali Taatali, Sif Eddine Sadaoui, Mohamed Abderaouf Louar, and Brahim Mahiddini. On-machine dimensional inspection: machine vision-based approach. The International Journal of Advanced Manufacturing Technology, 131(1):393-407, 2024. https://doi.org/10.1007/s00170-024-13081-1

[4] Hamam Mokayed, Tee Zhen Quan, Lama Alkhaled, and Sivakumar Vengusamy. Real-time human detection and counting system using deep learning computer vision techniques. Artificial Intelligence and Applications. 1(4): 221-229, 2023. https://doi.org/10.47852/bonviewAIA2202391

[5] Attoumane Abi, Julien Walter, Romain Chesnaux, and Ali Saeidi. Groundwater level monitoring using exploited domestic wells: Outlier removal and imputation of missing values. Hydrogeology Journal, 32(3):723-737, 2024. https://doi.org/10.1007/s10040-023-02740-4

[6] Joshua Snell, Matthew Simons, and Leonie Warlo. A test of letter configuration coding in visual word recognition. Language, Cognition and Neuroscience, 38(6):893-901, 2023. https://doi.org/10.1080/23273798.2023.2179083

[7] Vasanth Subramanyam, Jayendra Kumar, and Shiva Nand Singh. A hybrid descriptor for low-textural image stitching in real-time surface inspection systems. Multimedia Tools and Applications, 83(7):20653-20675, 2024. https://doi.org/10.1007/s11042-023-16357-y

[8] Yuanpei Chang, Ying Xue, Yu Zhang, Jingguo Sun, Zhangyuan Ji, Hewei Li, Teng Wang, and Jiancun Zuo. PCB defect detection based on PSO-optimized threshold segmentation and SURF features. Signal, Image and Video Processing, 18(5):4327-4336, 2024. https://doi.org/10.1007/s11760-024-03075-7

[9] Hanane Zermane, Ahcene Ziar, Hassina Madjour, and Djamel Touahar. Transforming industrial supervision systems: A comprehensive approach integrating machine learning techniques and fuzzy logic. The Scientific Bulletin of Electrical Engineering Faculty, 24(2):52-66, 2024. https://doi.org/10.2478/sbeef-2024-0021

[10] Monowar Wadud Hridoy, Mohammad Mizanur Rahman, and Saadman Sakib. A framework for industrial inspection system using deep learning. Annals of Data Science, 11(2):445-478, 2024. https://doi.org/10.1007/s40745-022-00437-1

[11] Michele Fraccaroli, Alice Bizzarri, Paolo Casellati, and Evelina Lamma. Exploiting CNN's visual explanations to drive anomaly detection. Applied Intelligence, 54(1):414-427, 2024. https://doi.org/10.1007/s10489-023-05177-0

[12] Dezhi Yang, Ning Chen, Qiqi Tang, Hang Zhang, and Jian Liu. Research on defect detection of toy sets based on an improved U-Net. The Visual Computer, 40(2):1095-1109, 2024. https://doi.org/10.1007/s00371-023-02834-w

[13] Dongdong Chu, Shulan Yu, Yuhang Ling, Yang Zhao, and Jianhong Zhang. A game-theoretic and multimodal interaction framework for collaborative robots in smart manufacturing. Decision Making: Applications in Management and Engineering, 7(1):735-751, 2024. https://doi.org/10.31181/dmame7120241433

[14] Carla Dei, Matteo Meregalli Falerni, Turgut Cilsal, Davide Felice Redaelli, Matteo Lavit Nicora, Mattia Chiappini, Fabio Alexander Storm, and Matteo Malosio. Design and testing of (A)MICO: a multimodal feedback system to facilitate the interaction between cobot and human operator. Journal on Multimodal User Interfaces, 19(1):21-36, 2025. https://doi.org/10.1007/s12193-024-00444-x

[15] Harshitkumar j Ghelani. AI-Driven quality control in PCB manufacturing: Enhancing production efficiency and precision. Valley International Journal Digital Library, 12(10):1549-1564, 2024. https://doi.org/10.18535/ijsrm/v12i10.ec06

[16] Ibrahim Yousif, Liam Burns, Fadi El Kalach, and Ramy Harik. Leveraging computer vision towards high-efficiency autonomous industrial facilities. Journal of Intelligent Manufacturing, 36(5):2983-3008, 2025. https://doi.org/10.1007/s10845-024-02396-1

[17] Fanghui Bi, Xin Luo, Bo Shen, Hongli Dong, and Zidong Wang. Proximal Alternating-direction-method-of-multipliers-incorporated nonnegative latent factor analysis. IEEE/CAA Journal of Automatica Sinica, 10(6):1388-1406, 2023. https://doi.org/10.1109/JAS.2023.123474

[18] Monowar Wadud Hridoy, Mohammad Mizanur Rahman, and Saadman Sakib. A framework for industrial inspection system using deep learning. Annals of Data Science, 11(2):445-478, 2024. https://doi.org/10.1007/s40745-022-00437-1

[19] Lin Wang, Yu Shen, Hongguo Zhang, Dong Liang, and Dongxing Niu. Automatic road extraction framework based on codec network. Journal of Measurement Science and Instrumentation, 15(3):318-327, 2024. https://doi.org/10.62756/jmsi.1674-8042.2024033

[20] Wei Liu, Cong Wang, and Yongkang Zhang. Industrial surface defect detection by multi-scale Inpainting-GAN. The Visual Computer, 41(8):5643-5660, 2025. https://doi.org/10.1007/s00371-024-03743-2