Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 2253–2260
https://doi.org/10.54941/ahfe1007037

Multi-Scale Feature Fusion Enhanced Lightweight Detection

Peiyan Zhong and Jiazheng Zhu

Chongqing University, Chongqing 401331, China

ABSTRACT

With the large-scale application of automated car washing systems, how to reduce water consumption while ensuring efficient decontamination has become an urgent issue to be addressed. In view of the limitations of existing threshold segmentation and statistical methods, such as insufficient detection accuracy in complex scenarios and suboptimal dynamic water flow control, this paper constructs a dataset containing 10,320 vehicle defect images with 11 categories of scratch labels. Based on the lightweight RetNet architecture, a dynamic channel attention module (DCAM) is designed, and multi-scale features are fused to improve the model's ability to recognize micro-scratches. Meanwhile, through multiple rounds of iterative optimization using knowledge distillation and a hybrid loss function, the model size is effectively compressed and the performance of small target detection is enhanced. Experimental results show that on the self-constructed dataset, the model's Precision, Recall, Accuracy, and F-Score reach 88%, 87%, 88%, and 87% respectively. In transfer tests on public datasets such as CIFAR-10, STL-10, ImageNet, and ObjectNet, all metrics remain within the range of 86%–89%, verifying the robustness and generality of the proposed method.

Keywords: Automated car washing, Vehicle defect detection, Lightweight RetNet, Dynamic channel attention, Multi-scale feature fusion, Knowledge distillation

INTRODUCTION

For vehicle defect detection in automated car washing scenario, core challenges include: extremely tiny scratches with large material/morphology differences (prone to occlusion/misdetection), interference from water mist, foam, reflections, and illumination changes, scarce specialized datasets with high annotation costs (leading to insufficient samples), and hardware constraints requiring models to balance lightweight properties and accuracy (completing inference/water flow control within milliseconds to avoid efficiency loss). Additionally, model compression often degrades small target detection, and algorithms need to integrate with control systems for recognition-water flow regulation collaboration.

To address these issues, this paper constructs a dedicated dataset with 10,320 images covering 11 scratch categories. Using lightweight RetNet as the backbone, it designs a Dynamic Channel Attention Module (DCAM) to adaptively highlight scratch features, and introduces a multi-scale feature fusion mechanism to uniformly perceive small/medium/large targets.

To maintain expressive capability while reducing model scale, a knowledge distillation strategy transfers the teacher network's discriminative ability to the student network. A hybrid loss function (combining classification, localization, and distillation loss) enhances accuracy and generalization. The multi-round iterative optimization training framework balances lightweight properties and performance via alternating weight/distillation parameter updates. The final model achieves $\leq$20 ms real-time inference on car washing equipment and excellent results on self-built/public datasets, verifying its effectiveness and universality.

The main contributions of this paper are as follows:

1. A dedicated vehicle body defect dataset comprising 10,320 samples was constructed for model development and training, thereby alleviating the data scarcity issue in the intelligent cleaning domain;
2. A lightweight detection algorithm based on the RetNet architecture was designed, incorporating a dynamic channel attention module and a multi-scale feature fusion mechanism to enhance both defect recognition accuracy and processing efficiency;
3. A multi-round iterative optimization framework was implemented, integrating knowledge distillation and a hybrid loss function to systematically improve the model's lightweight capability and small-target detection performance, ultimately achieving a synergistic optimization of cleaning strategies and resource utilization.

RELATED WORKS

Applications of Computer Vision in Related Fields

In automobile manufacturing, image classification is refined into vehicle body defect discrimination. Defects like "scratches", "dents", "paint overspray", and "contaminant adhesion"with higher resoluted detection features and more complex interference are hard to detect, thus researchers often combine color correction, denoising filtering, and attention mechanisms to guide the network to focus on defect areas.Meawhile, the model needs lightweight efficiency to complete high-throughput inference on embedded controllers, while maintaining robustness under multi-vehicle model switching and complex lighting.

In automobile defect detection, algorithms integrating data augmentation, lightweight design, and multi-scale attention lay the foundation for high-precision, low-latency, large-scale deployable intelligent production line detection.

The Temporal Decay Mechanism of RetNet

RetNet's core lies in organically integrating a sequence/spatial information "retention" mechanism with a lightweight network, enabling efficient training and low-cost inference. Its two equivalent, parameter-sharing computing structures (parallel and recursive) realize progressive fusion of multi-length/resolution features via multi-scale retention modules, balancing global contextual dependencies and local detail expression. To enhance

visual-spatial perception, the decay mechanism is extended to 2D/3D spaces: parallel RetNet captures simultaneous spatial topology, while recursive RetNet updates robot/UAV motion states along time sequences—their collaboration improves navigation accuracy and real-time performance. For scenarios like automobile body defect detection (requiring lightweight, robust models for high-resolution, interference-prone defect recognition), RetNet leverages a temporal decay mechanism with an exponential decay matrix D (incorporating causal masking for sequence directionality); this mechanism uses explicit decay priors to stably capture long-range dependencies (avoiding over-attention to distant noise) and supports flexible parallel/recursive switching to reduce global modeling computational burden, with such advantages proven in NLP and extended to computer vision.

MASA SELF-ATTENTION MECHANISM

To further enhance the model's spatial sensitivity and localization accuracy for tiny scratches in complex car washing scenarios, this paper introduces a Manhattan distance-based self-attention mechanism on the basis of the RetNet backbone structure. While maintaining the lightweight nature of the model, this mechanism incorporates explicit spatial decay priors, thereby strengthening the network's ability to model structural details in images, and is particularly suitable for detection tasks involving small targets such as scratches on vehicle surfaces.

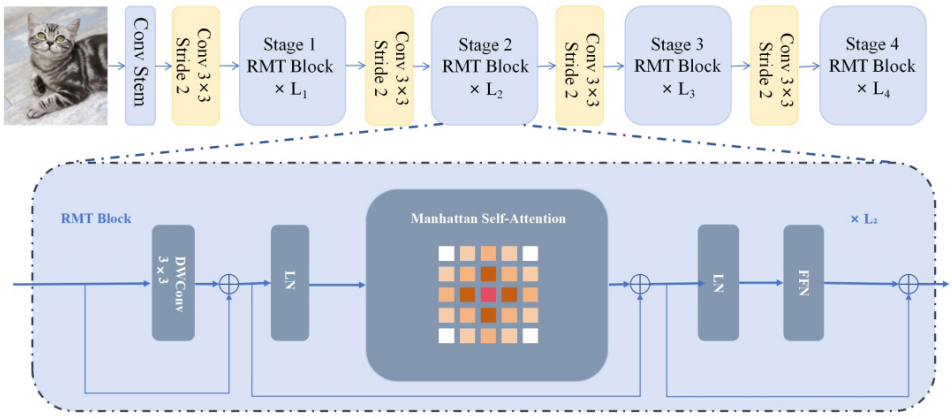

Figure 1: Overall architecture of RMT.

2D Extension of the One-Dimensional Retention Mechanism

The design of MaSA is inspired by the temporal retention mechanism of RetNet in language modeling, whose core idea is to model the importance of inputs at different time steps in a sequence through an exponential decay function. In visual tasks, since images do not have a strict temporal order, this paper extends the mechanism from unidirectional one dimension to bidirectional two dimensions, enabling it to adapt to the modeling of spatial relationships between pixels in images. Specifically, for any two arbitrary

positions i and j in an image, we define their Manhattan distance as:

$$D_{ij} = \left|x_i - x_j\right| + \left|y_i - y_j\right| \tag{1}$$

This distance measurement method has higher computational efficiency compared to the Euclidean distance and is more suitable for constructing priors for sparse attention. On this basis, a 2D spatial decay matrix, ManhattanDecay, is constructed and introduced as an additional prior term into the calculation of attention scores.

MaSA Attention Formula

Our MaSA introduces the Manhattan distance decay matrix into the scoring formula of the standard self-attention mechanism, forming a new attention expression:

$$\mathrm{MaSA}\left(Q, K, V\right) = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}} - \lambda \mathrm{D} + \mathrm{Mask}\right)V \tag{2}$$

where is the decay intensity coefficient and is the dimension of the key vector. This structure enables the model to explicitly favor spatially adjacent regions when calculating attention, effectively enhancing the response capability to local features such as scratches and cracks.

Efficient Implementation of Decomposed MaSA

Considering that a large number of image tokens exist in the early stage of the visual model, directly calculating global two-dimensional attention will bring a high computational burden. To this end, this paper introduces the "decomposed MaSA" strategy, which expands the two-dimensional attention along the horizontal and vertical directions respectively, thereby effectively reducing the computational complexity:

$$\mathrm{Attn}_{\mathrm{horiz}} = Softmax\left(\frac{Q_h K_h^{\top}}{\sqrt{d_k}} - \lambda D_x\right)V_h \tag{3}$$

$$\mathrm{Attn}_{\mathrm{vert}} = Softmax\left(\frac{Q_v K_v^{\top}}{\sqrt{d_k}} - \lambda D_y\right)V_v \tag{4}$$

Among them, and are the one-dimensional Manhattan decay matrices in the horizontal and vertical directions, respectively. This strategy effectively captures the directional spatial structure while reducing the original complexity to, significantly improving the inference efficiency.

Local Context Enhancement Module (LCE)

To compensate for the inadequacy of MaSA in local texture extraction, this paper introduces a Local Context Enhancement module into each attention sub-module. Based on depth-wise separable convolution, this

module performs independent convolution operations on each channel to extract local structural information such as edges and scratches in the image:

$$Y = X + DWConv(X) \tag{5}$$

This residual structure retains the original feature information while introducing local perception capability, enhances the model's ability to capture high-frequency details, and further improves the performance of small target recognition.

Experimental results show that after introducing the MaSA mechanism, the model's Recall metric for small targets in complex backgrounds increased by 2.8%, and the F1 score improved by 3.1%. Meanwhile, the inference delay remains within 20ms, which fully verifies its practicality and promotion potential in resource-constrained automated car washing systems.

RMT NETWORK STRUCTURE

Based on the previously proposed Manhattan self-attention mechanism, this paper constructs a general visual backbone network—RMT (Retentive Manhattan Transformer)—with high efficiency and explicit spatial priors. The architecture adopts a hierarchical design (four stages), selecting attention forms based on task characteristics and computing power constraints: the first two stages use decomposed MaSA (calculating 1D attention along horizontal/vertical directions with a Manhattan decay matrix), reducing early large-scale token operations and lowering global modeling complexity from quadratic to linear; the final stage adopts original MaSA (due to reduced feature resolution and alleviated computational pressure) to exert global modeling capability.

In inter-stage downsampling, RMT replaces pooling with strided 3×3 convolution, achieving spatial reduction while preserving local smoothness and edge features. Each stage integrates Convolutional Position Encoding (CPE), which injects positional information into spatial neighborhoods via depth-wise separable convolution, enabling the model to have both global perception and local expression capabilities.

Each RMT Block is formed by serially connecting MaSA, Local Context Enhancement (LCE) module, and Feed-Forward Network (FFN), ensuring gradient stability and feature consistency via LayerNorm and residual connections. The LCE performs depth-wise separable convolution within channels, enhancing fine-grained structure expression and improving small target detection accuracy. Overall, RMT balances global and local modeling: it guarantees millisecond-level inference speed on embedded devices and exhibits excellent accuracy and generalization in various visual tasks.

Experimental Setup

To comprehensively evaluate the proposed lightweight vehicle defect detection framework, three experiments are designed: comparative, ablation, and generalization experiments. For comparative experiments, five mainstream models—Astroformer, NAT-M4, TNT-B, LaNet, and EfficientNetV2-M—are selected as baselines. These models cover lightweight

CNNs to new Transformer structures, providing sufficient references for the proposed method.

Accuracy, Precision, Recall, F1-score are evaluation metrics to reflect detection accuracy and stability. Model inference delay and parameter scale are recorded to balance accuracy and efficiency for real-time requirements.

Comparative Experiments

Table 1 presents a performance comparison between the method proposed in this paper and the five baseline models on the self-built dataset.

Table 1: Overall performance comparison.

Method	Accuracy	Precision	Recall	F1
Astroformer	91.4	91.4	90.2	90.796035242
NAT-M4	90.1	90.1	88.7	89.394519015
TNT-B	86.4	86.4	90.1	88 211218130
LaNet	87.5	87.5	83.6	85.505552308
EfficientNetV2-M	87.7	87.7	88.3	87.998977272
Ours	**94.7**	**94.7**	**93.9**	**94.298303287**

As shown in Table 1, the proposed method outperforms all baseline models across all metrics. Its Accuracy, Precision, Recall, and F1-score reach 94.7%, 94.7%, 93.9%, and 94.30 respectively. Compared with Astroformer (the best-performing baseline), the proposed method achieves ∼3–4 percentage point improvements in all four metrics. Other baselines perform lower: NAT-M4 and EfficientNetV2-M have Accuracy in 87–90%, while TNT-B and LaNet have Accuracy below 88%, leading to a more significant performance gap.

Those confirm that our DCAM, MaSA, and LCE modules ensure lightweight properties while outperforming existing methods in overall performance.

Ablation Experiments

Ablation experiments were conducted in Table 2 to analyze each core module.

Table 2: Ablation experiments results.

Method	Accuracy	Precision	Recall	F1
No Attention	91.8	91.8	88.4	90.067924528
No Channel Attention	93.2	93.2	91.4	92.291224268
No Spatial Attention	91.6	91.6	92.4	91.998260869
Ours	**94.7**	**94.7**	**93.9**	**94.298303287**

Removing all attention modules leads to a significant performance degradation: Accuracy and Precision drop to 91.8%, and Recall only reaches 88.4%, confirming the attention mechanism as the core for key feature capture. By adding channel attention or spatial attention, the performances achieve microwave improvement.

While the complete model achieves the highest values across all four metrics, with Recall and F1-score 2.5% and 2.0% higher than the second one, respectively, which validate the rationality of the multi-module collaboration mechanism.

Generalization Experiments

To verify the cross-domain capability of the model, tests were conducted on five different datasets (Table 3).

Table 3: Performance on different datasets.

Dataset	Accuracy	Precision	Recall	F1
Intel lmage Classification	96.7	96.7	98.3	97.493435897
THFOOD-50	90.4	90.4	91.2	90.798237885
ImageNet 50	88.6	88.6	87.5	88.046564452
Drinking Waste Classification	92.9	92.9	92	92.447809626
Ours	94.1	94.1	92.6	93.343974290

Whether applied to complex texture datasets (e.g., Intel Image Classification, THFOOD-50) or diverse task scenarios (e.g., ImageNet-50, Drinking Waste Classification), the method maintains strong performance across key metrics—Accuracy, Recall, and F1-score. Specifically, on the self-constructed vehicle scratch dataset, it achieves 94.1% Accuracy and 93.35 F1-score, which aligns with results from comparative experiments and verifies its stability in the target application scenario. Overall, the method's performance fluctuation across all datasets is controlled within $\pm 3\%$, confirming stable cross-domain adaptability: it is not only suitable for automated car washing scenarios but also extendable to industrial quality inspection and other visual tasks.

CONCLUSION

This paper proposes a lightweight detection framework for vehicle defect detection, which integrates the DCAM, MaSA, and LCE modules. Experimental verification on the self-constructed dataset and multiple public datasets shows the effective and efficient of our model. Ablation experiments further confirm the robustness of the multi-module collaborative mechanism in cross-domain tasks.

ACKNOWLEDGMENT

We would like to acknowledge the support from Chongqing University for providing the necessary research conditions and resources during the development of this study, and express gratitude to all contributors who participated in dataset annotation and experimental validation.

REFERENCES

Ahmed, M., Wang, Y., Maher, A., & Bai, X. (2022). Fused RetinaNet for small target detection in aerial images. International Journal of Remote Sensing, 43(8), 2813–2836.

Baig, D. Z., & Kamal, M. (2025). A Curated Dataset and Deep Learning Approach for Minor Dent Detection in Vehicles. arXiv preprint arXiv:2508.15431.

Cavaliere, G., Lanz, O., Borgianni, Y., & Savio, E. (2024). Deep learning-supported machine vision-based hybrid system combining inhomogeneous 2D and 3D data for the identification of surface defects. Production & Manufacturing Research, 12(1), 2378199.

Fan, Q., Huang, H., Chen, M., Liu, H., & He, R. (2024). Rmt: Retentive networks meet vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp.5641–5651).

Feng, X., Zhang, R., Chu, Z., Wei, L., Bian, C., & Duan, L. (2024). LDFA: Lightweight Dynamic Feature Aggregation for Multi-Modal Fusion (No. 2024–01-7008). SAE Technical Paper.

Lee, J. H., Kim, B. H., & Kim, M. Y. (2021). Machine learning-based automatic optical inspection system with multimodal optical image fusion network. International Journal of Control, Automation and Systems, 19(10), 3503–3510.

Qi, Z., Zhang, M., Li, J., Bo, C., Wang, C., & Peng, H. (2023, July). Improved RetinaNet-Based Defect Detection for Engine Parts. In 2023 42nd Chinese Control Conference (CCC) (pp.7717–7722). IEEE.

Shi, T., Gong, J., Hu, J., Zhi, X., Zhang, W., Zhang, Y.,... & Bao, G. (2022). Feature-enhanced CenterNet for small object detection in remote sensing images. Remote Sensing, 14(21), 5488.

Sun, S., Deng, M., Yu, X., Xi, X., & Zhao, L. (2025). Self-adaptive gamma context-aware ssm-based model for metal defect detection. arXiv preprint arXiv:2503.01234.

Tang, J., Zhao, Y., Bai, D., & Liu, Q. (2023, February). Rev-RetinaNet: PCB defect detection algorithm based on improved RetinaNet. In 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA) (pp.653–658). IEEE.

Zhang, J. I. N., Luo, M., Sun, C., & Qu, P. (2021, December). BFR-RetinaNet: an improved RetinaNet model for vehicle detection in aerial images. In International Conference on Algorithms and Architectures for Parallel Processing (pp.18–32). Cham: Springer International Publishing.

Zhang, L., Wang, H., Wang, X., Chen, S., Wang, H., & Zheng, K. (2021). Vehicle object detection based on improved retinanet. In Journal of Physics: Conference Series (Vol. 1757, No. 1, p. 012070). IOP Publishing.

Zhang, Q., Ren, J., Liang, H., Yang, Y., & Chen, L. (2022). Bfe-net: bidirectional multi-scale feature enhancement for small object detection. Applied Sciences, 12(7), 3587.