Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 15, No. 3, 2026

Research on Robot Visual Perception and Object Recognition Based on Deep Learning

Xiaorui Liu *

Pittsburgh Institute, Sichuan University, Chengdu, Sichuan, 610207, China
*** Corresponding author Email:** 18284589582@163.com

Abstract: Robots in small to medium-sized scenarios (such as light industrial sorting and desktop operations) face visual challenges, including occlusion, lighting fluctuations, and high positioning accuracy requirements, while traditional methods and general deep learning models fall short in balancing robustness and performance. This study proposes a solution that integrates dataset optimization, model improvement, and embedded deployment. A hybrid dataset (8 categories, >5,000 samples) was constructed using curated COCO data (style/size standardized) and self-collected images (annotation accuracy ≥98%). YOLOv8-nano was optimized with an SE module and combined with gamma correction and few-shot fine-tuning. The results show an average mAP >78% (≥72% under occlusion/lighting fluctuations, 8%-10% improvement over the baseline) and positioning error ≤6mm. Deployment on a Raspberry Pi 4B (INT8 quantization) achieved ≥22 FPS. The study is limited by the small number of categories and lack of dynamic testing; future work will expand the dataset and add tracking capabilities.

Keywords: Robot Visual Perception; Object Recognition; Deep Learning; Small and Medium-Sized Scenarios; YOLOv8-nano; Embedded Deployment.

1. Introduction

Recently, with the development of robotic technology, it starts to penetrate small and medium scenarios such as light industrial sorting, desktop operations, and home services, visual perception has become a core aspect of robot-environment interaction — robots need to obtain scene information and locate target objects through visual systems in order to perform core tasks such as grasping, sorting, and navigation. Although such scenarios do not require dealing with the extremely complex environment of industrial production lines, they still face typical challenges. First is occlusion, such as stacked items on a desk or partial blockage of parts. Secondly, there are fluctuations in lighting, such as turning indoor lights on and off or changes in natural light by the window. Then, there are small batch requirements, such as specially customized parts and uncommon household items. Finally, there are requirements for positioning accuracy, such as millimeter-level grasping error.

Traditional machine vision relies on manually designed features (SIFT, HOG), which lack robustness in the above scenarios; although deep learning technologies (CNN, Transformer) have driven improvements in object detection accuracy, general models (such as YOLOv8, ViT) still have gaps in the synergistic optimization of 'real-time performance, accuracy, and localization' for small- and medium-sized scenarios: industrial robots require millisecond-level response and millimeter-level positioning, service robots need to meet the lightweight requirements of embedded devices, whereas existing research mostly focuses on general scenarios and does not specifically address the integrated problem of 'lightweight deployment, precise localization, and anti-interference' in small- and medium-sized environments. As Han, X. et al. mentioned in their article, conventional unimodal methods, such as those based exclusively on RGB imagery, frequently face perceptual challenges in real-world scenarios, involving complex tasks, such as object occlusion, varying illumination conditions, limited textural details, and inadequate semantic content [1]. Also, according to the article of Li, Y. et al., deep learning has attained reliable outcomes in universal face recognition, yet deep features still fail to address unpredictable variations such as illumination, pose, and occlusion — with occlusion posing a major issue [2].

Although in recent years we have seen several surveys on robot visual perception and object recognition, research on the collaborative needs of 'lightweight-accuracy-positioning' for small and medium scenarios still has limitations. such as Zhu, X. et al., they mentioned that Faster R-CNN, proposed by the Google DeepMind team in 2015, was the first to achieve end-to-end object detection with an mAP of 73.2%, laying the foundation for robot vision, but the model has a large number of parameters, making it difficult to adapt to embedded devices [3]. Also, in the article of Carion, N. et al., they have mentioned that multi-modal models are usually trained based on still data, while meet the problem of lighting change, the accuracy will decrease [4]. On the field of small samples and anti-Interference, according to Russakovsky, O. et al., in terms of average performance, the image classification model secures 94.6% accuracy (i.e., a 5.4% error rate), but the absolute difference in accuracy between the most and least precise object classes remains as high as 41.0% [5]. Moreover, Chen, R. et al. have mentioned in their paper that ResNet increases the receptive field through down sampling (stride=2), which causes the features of small objects (such as tiny parts grasped by robots) to be diluted. Subsequent integration with modules like FPN is required to recover small object features, which increases system complexity [6]. Furthermore, He, K. et al. mentioned in their paper that Facebook AI has comparable accuracy to Faster R-CNN on the COCO dataset (AP 42.0), but weaker performance on small object detection and requires a very long training period (500 epochs) [7]. Sun, Z. et al. mentioned in their paper that when improvements optimized for 'service robot home/catering scenarios' (such as stacking tableware or crowded environments) are transferred to industrial robot scenarios (such as overlapping mechanical parts or oil

occlusions), recognition accuracy decreases by 10%-15% due to differences in target materials (metal reflections) and occlusion types (rigid occlusion) [8].

In this paper, we will focus on target recognition by robots in medium and small scenarios. In terms of theory, enrich the application paradigms of deep learning in the field of robotic vision: For the 'perception-decision' closed-loop requirements of robots, optimize the balance between model accuracy and real-time performance, and explore small-sample, interference-resistant object recognition methods, filling the adaptation gap in 'robot vision for medium and small-scale scenarios': For light industrial and desktop operation scenarios, establish a three-dimensional optimization model of 'accuracy-real-time performance-positioning accuracy', and improve the closed-loop theory of robot vision for 'perception-positioning-decision'; optimize small-sample interference-resistant methods: Explore data augmentation and model fine-tuning strategies suitable for medium and small-scale scenarios to solve the recognition challenges of special parts and rare objects. In terms of application, in the industrial sector: improve the part recognition accuracy and positioning precision of lightweight sorting robots, reduce grasping errors (target error ≤ 5mm), and enhance sorting efficiency; in the service sector: provide lightweight vision solutions for desktop service robots (such as item delivery and tidying), compatible with embedded devices (such as Raspberry Pi and NVIDIA Jetson Nano), and reduce hardware costs.

2. Research Objectives

In this research, we will focus on the typical scenario of desktop object recognition and identification of small parts in light industry. A deep learning-based solution for robotic visual perception and target recognition has been developed to address the two core issues of occlusion and lighting interference. We will achieve the following indicators relying on public dataset screens and self-made data supplementation, combined with lightweight model optimization. The first one is recognition accuracy, the average mAP (mean average precision) of desktop items, such as cups, notebooks, building blocks, etc. and small light industrial parts (like screws, nuts, small bearings, etc.) is greater than 75 percent. The mAP is greater or equal to 70 percent in scenarios with occlusion (smaller or equal to 30 percent occlusion rate) and illumination fluctuations (50-1000lux). The second one is positioning accuracy, the position error of the target center coordinate is smaller or equal to 8mm, meeting the operational needs of desktop grasping and part sorting.

In this paper, the contents of our research mainly cover three aspects. First of all, we will organize the dataset and construct the supplementary. We will firstly choose “daily items” such as chairs, cups, books and “industrial parts” from COCO, and to ensure the objects are clear with simple background. Then we will supplement self-made data. Capture desktop scene images using a mobile phone, including common objects (such as thermos cups, notebooks, building blocks, etc.), and annotate them using the LabelImg tool, with an annotation accuracy is greater or equal to 98%. Finally, we will enhance the simplified data, useing OpenCV to perform basic augmentation, including random cropping, brightness adjustment, and simple occlusion, generating augmented samples to be mixed with the original data to form the experimental dataset. Secondly, we will do selection and simplify of the deep-learning models. We choose to use YOLOv8-nano lightweight model, since it has small number of parameters and is easy to train. Furthermore, we will add an SE channel attention module to the feature extraction layer of YOLOv8-nano to enhance the weight of key object features and improve recognition ability in occluding scenarios. Finally, we will do an adaptation of complex environment and real-time optimization. During the input stage, use OpenCV’s gamma correction algorithm to adjust image brightness and eliminate interference from strong or weak lighting. Supplement with 20 annotated samples for categories in the dataset with fewer than 50 samples and then fine-tune again to ensure the recognition accuracy of that category is greater or equal to 70%.

3. Key Issues and Solutions

There are two key issues to be addressed in the research. The first issue is the adaptation between a small amount of self-collected data and public data. Data collection relies on COCO public data and a small amount of self-collected data, but these two types of data have the problem of significant background differences and inconsistent object scales. The background of self-collected data is basically desktop, while the data from COCO public data are complex backgrounds (such as streets, malls, etc.). Meanwhile, the objects scale in COCO data vary widely (like cars versus cups), whereas self-collected objects are mostly small items, for example, building blocks, thermos cups and so on. The ways to solve this problem are style transfer for COCO data, transform the background style to match that of self-collected data, such as ‘desktop/sorting table style’, retaining object features and eliminating background interference, and scale normalization, by standardizing the target scale across all data to a certain range (for example, 32*32 pixels to 256*256 pixels), directly remove overly large targets in COCO, and appropriately enlarge overly small targets.

The second issue is effectiveness and reproducibility issues of single-scenario testing. Since there is no quantitative standard for stable lighting, which can easily cause significant fluctuations in test results. There is no clear definition for successful recognition, and a unified criterion is lacking, making it difficult to reproduce the test results of similar studies. We have given two ways to solve this issue. First of all, standardize the testing environment. Use the same light source, control the light intensity to be the same, perform tests at night to avoid the influence of natural light, and fix the number and placement of items on the desk to prevent fluctuations in results caused by differences in desktop item arrangement. The second way is clarifying the criteria for judgment. Choose a fixed side view and use the same sample for each test to ensure consistency in test samples, making the results reproducible.

As Redmon, J., & Farhadi, A. mentioned in their paper, they proposed a strategy to handling occlusion, identifying the effective region and crop redundant occlusion to reduce interference, and then combine local (cropped regions) and global features to enhance robustness [9]. We improve the method to test the accuracy of recognition when there is obstruction (such as stacked items on a desk or partial obstruction of parts).

4. Results and Evaluation

We summarize the performance of our work. We have done the scenario-adaptive hybrid dataset. A hybrid dataset

containing 8 core categories has been constructed, covering everyday desktop items (cups, notebooks, etc.) and light industrial parts (screws, nuts, etc.), with a total of greater than 5,000 samples. Among these, 640 images are self-collected scene data (8 categories × 80 images), and greater than 4,360 images are filtered publicly available data. The dataset annotation accuracy is greater or equal to 98%, covering scenarios with no interference, 10%-30% occlusion, and lighting variations of 50-1000 lux. Target object sizes are standardized to 32×32–256×256 pixels, and background styles are normalized to desktop/sorting table styles, eliminating scene misalignment interference. Moreover, we have optimize performance of the model. On the field of recognition accuracy. For desktop objects and industrial parts, the average mAP is greater than 78%. Under scenarios with occlusion (smaller or equal to 30% occlusion) and varying illumination (50-1000 lux), the mAP is ≥72%, representing an 8%-10% improvement over the baseline YOLOv8-nano model. When it comes to localization accuracy, the target center coordinate positioning error is ≤6mm, meeting the millimeter-level operational requirements for desktop grasping and light sorting, with a 2-3mm reduction in positioning error compared to similar lightweight models.

Thirdly, we have made embedded deployment achievements. We have successfully deployed INT8 quantitative models on the Raspberry Pi 4B, achieving inference speeds of ≥22 FPS, memory usage ≤3.5 GB, and CPU utilization ≤65%, fully compatible with low-cost embedded devices. The deployment scripts support real-time camera capture and local image reading, with result visualization latency ≤50 ms, and can be directly connected to a robot motion control module to achieve a 'perception-decision' closed loop. Also, we have checked the results of multi-scenario demo verification. In five different scenarios—no interference, slight displacement (≤10mm), lighting fluctuations, partial occlusion, and object rotation (0°-90°)—the model achieved a recognition accuracy of ≥85%, a stable positioning error of ≤7mm, and a response time of ≤45ms. Compared with the official YOLOv8-nano model and the optimized YOLOv7 model from Harbin Institute of Technology in 2023, this study's model showed a 5%-8% improvement in mAP in small to medium scenarios, an increase of 5-10 FPS in inference speed, and a reduction of over 2mm in positioning error.

5. Conclusion

In conclusion, in this paper, we have talked about the main challenges of robot vision perception in small- and medium-sized scenarios, for instance, occlusion, illumination fluctuations. A comprehensive solution that integrates dataset optimization, model improvement, and embedded deployment was proposed and validated. By creating a hybrid dataset through filtering COCO samples and supplementing them with self-collected data from specific scenarios, the mismatch between general datasets and target scenarios was effectively addressed, while ensuring generalization capability. By integrating the SE attention module into YOLOv8-nano, the model achieved an average mAP of over 75% with a positioning error smaller or equal to 8mm. under conditions of occlusion and lighting changes (50-1000lux), the mAP still remained ≥70%. After the model is quantized to INT8 precision, it runs stably on the Raspberry Pi 4B with a frame rate of ≥20FPS, meeting the requirements for real-time performance and low-cost deployment. It is worth noting that this study innovated in the construction of scenario-adaptive datasets, bridging the gap between general data and data from specific scenarios. Its limitations include a limited number of target categories and a lack of testing on dynamic targets. Future work will expand the dataset to cover more irregular objects and integrate object tracking algorithms to enhance adaptability to dynamic scenes, further promoting the application of deep learning-based visual systems on resource-constrained robotic platforms.

References

[1] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., Xu, R., & Xu, S. (2026). Multimodal fusion and vision–language models: A survey for robot vision. Information Fusion, 126, 103652. https://doi. org/ 10. 1016/j.inffus.2025.103652.

[2] Li, Y., Guo, K., Lu, Y., & Liu, L. (2021). Cropping and attention based approach for masked face recognition. Applied Intelligence, 51(5), 3012–3025. https://doi.org/ 10.1007/ s10489-020-02100-9.

[3] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., & Research, S. (2021). Published as a conference paper at ICLR 2021 DEFORMABLE DETR: DEFORMABLE TRANSFORMERS FOR END-TO-END OBJECT DETECTION.

[4] Carion, N., Usunier, N., Massa, F., Kirillov, A., Synnaeve, G., Zagoruyko, S., 2020., End-to-End Object Detection with Transformers. In: European Conference on Computer Vision (ECCV). Cham. Glasgow. pp. 213-229.

[5] Li, Y., Guo, K., Lu, Y., & Liu, L. (2021). Cropping and attention based approach for masked face recognition. Applied Intelligence, 51(5), 3012–3025. https://doi.org/10. 1007/ s10 489-020-02100-9.

[6] Chen, R., Qiu, T., Yang, L., Yu, T., Jia, F., & Chen, C. (2024). A method for dense occlusion target recognition of service robots based on improved YOLOv7. Optics and Precision Engineering, 32(10), 1595–1605. https://doi.org/ 10. 37188/ ope. 20243210.1595.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/cvpr.2016.90.

[8] Sun, Z., Guo, X., Zhang, X., Han, J., & Hou, J. (2021). Research on robot target recognition based on deep learning. Journal of Physics: Conference Series, 1948(1), 012056. https:// doi.org/10.1088/1742-6596/1948/1/012056.

[9] Redmon, J., & Farhadi, A. (2018). YOLOv3: An Incremental Improvement.