EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

AGV monocular vision localization algorithm based on Gaussian saliency heuristic

Heng Fu[1], Yakai Hu[1], Shuhua Zhao[1], Jianxin Zhu[1*], Benxue Liu[2] and Zhen Yang[2]

*Correspondence:
1044792533@qq.com

[1] Anyang Cigarette Factory of Henan China Tobacco Industry Co., Ltd, Anyang 455000, China
[2] Zhengzhou University, Zhengzhou 450001, China

Abstract

To address the issues of poor detection accuracy and the large number of target detection model parameters in existing AGV monocular vision location detection algorithms, this paper presents an AGV vision location method based on Gaussian saliency heuristic. The proposed method introduces a fast and accurate AGV visual detection network called GAGV-net. In the GAGV-net network, a Gaussian saliency feature extraction module is designed to enhance the network's feature extraction capability, thereby reducing the required output for model fitting. To improve the accuracy of target detection, a joint multi-scale classification and detection task header are designed at the stage of target frame regression to classification. This header utilizes target features of different scales, thereby enhancing the accuracy of target detection. Experimental results demonstrate a 12% improvement in detection accuracy and a 27.38 FPS increase in detection speed compared to existing detection methods. Moreover, the proposed detection network significantly reduces the model's size, enhances the network model's deployability on AGVs, and greatly improves detection accuracy.

Keywords: AGV monocular vision location, Gaussian saliency enhancement, GAGV-net

1 Introduction

At present, AGV vision-aided localization is a research hotspot, which plays a vital role in industry and service industry [1, 2]. However, due to the complexity of application scenarios, AGV vision-aided localization technology is facing huge challenges [3–5].

The vision localization system of AGV plays an important role in AGV localization. In the vision navigation and localization system, the navigation method based on local vision to install vehicle cameras in robots is widely used at present. In this navigation mode, control equipment and sensing devices are loaded on the robot body, and high-level decisions such as image recognition and path planning are completed by the onboard control computer.

The working principle of the visual navigation and localization system is simply to perform optical processing on the surrounding environment of the robot. First, use the camera to collect image information, compress the collected information, and then feed it back to a learning subsystem composed of neural networks and statistical methods. Then, the learning subsystem connects the collected image information with the actual

position of the robot, Complete the autonomous navigation and localization function of the robot. At present, most AGV vision localization systems are divided into three steps: visual information collection, target detection, and localization algorithm. Among them, target detection is the key to affect the localization accuracy of visual localization. For the existing AGV-based visual object detection, it is very difficult to detect accurately and quickly in complex environments.

Therefore, in order to solve these problems, this paper proposes a new AGV vision localization method. In this method, AGV visual detection network (GAGV-net) based on Gaussian saliency heuristic is proposed. First of all, in order to improve the feature extraction ability of the network to the target, we introduced the Gaussian salient target feature extraction module in the feature extraction part of the network to enrich the feature expression of the target. Through the excellent feature extraction capability of the feature extraction module, the model parameters of the network are greatly reduced, which realizes the purpose of model lightweight. Secondly, in the part of network classification decision, we introduce the joint multiscale classification module to improve the classification accuracy of the network. The experimental results show that the proposed detection method has better detection performance than the existing advanced detection methods, and the model size is much smaller than the existing methods, which greatly improves the detection speed. The contributions of this paper are as follows.

(1) A new AGV monocular vision localization framework based on Gaussian saliency heuristic is proposed.
(2) An AGV vision detection network (GAGV-net) based on Gaussian saliency heuristic is proposed. Compared with the existing methods, the network has a higher detection accuracy and detection speed, which provides technical support for rapid and accurate AGV vision-aided localization. The experimental results show that, compared with the existing detection methods, the detection accuracy of the proposed detection network is improved by 12%, and the detection speed is improved by 27.38 FPS.
(3) In the GAGV-net network, an efficient feature extraction module of target saliency is proposed. Through this feature extraction module, the feature extraction ability of the network is greatly improved, thus reducing the parameters required for model fitting.
(4) In the GAGV network, a joint multi-scale classification module is proposed, which greatly improves the classification accuracy of the network.

2 Related work

2.1 AGV visual positioning

In general, accurate target detection algorithm is the key of AGV vision localization technology. Therefore, a large number of scholars have conducted research on AGV visual localization algorithm, hoping to achieve more accurate target localization through accurate target detection technology. In these methods, Kang et al. [6] used cameras to capture tags and used SVM predictors to classify tags. Ding et al. [7] used the decision tree model to pre classify the targets and used the long short-memory (LSTM) network

to distinguish uncertain state data to improve the fault detection accuracy. Kuang et al. [8] proposed a Hough-based fuzzy inference algorithm transformation to solve the problem of slow inspection speed, so as to improve the real-time performance of the entire system. Yang et al. [9] improved YOLOv5 model to achieve more accurate target inspection. This method introduces the attention mechanism in yolo-v5 to improve the feature representation of the target, and realizes the effective constraint of the network by improving the loss function. Liu et al. [10] proposed an end-to-end edge detection method based on traditional adaptive threshold method and depth learning to overcome the problem of non-uniformity and achieve accurate target detection. Dong et al. [11] proposed a vision-aided localization and navigation system to enhance the intelligence and capability of traditional AGVS equipped with 2D LIDAR sensors and make it more robust in various environments, which integrates the advantages of cameras and 2D LIDAR. Li et al. [12] proposed a vision-based adaptive localization algorithm for global attitude correction and visual servo motion controller, which realized the automatic driving function of AGV. Although the existing AGV visual inspection algorithms have achieved good detection performance, as the application scenarios become more complex, the detection performance of existing methods is poor. In addition, the existing detection methods often have a large number of parameters, which leads to poor real-time detection.

2.2 Deep learning target detection

Target detection is a vital problem in computer vision, focusing on identifying and localizing specific objects in images or videos [13–16]. In recent years, the performance of target detection has significantly improved with the advancement of deep learning technology. Currently, deep learning-based target detection algorithms have become the mainstream approach. Classic algorithms like Fast R-CNN [13], YOLO [14], SSD [15], and RetinaNet [16] have gained popularity in this field. These algorithms utilize different network structures and loss functions, each with its own strengths and limitations. The R-CNN series, including R-CNN, Fast R-CNN, and Faster R-CNN [17], are well-known algorithms in target detection. The fundamental concept behind these algorithms is the conversion of target detection into candidate region extraction and classification. Initially, a set of candidate regions is extracted using techniques like selective search. Then, each candidate region is subjected to extraction and classification, resulting in the output of the target's location and category [18]. The YOLO series is a single-stage target detection algorithm, represented by YOLO, YOLOv3 [19], YOLOv5 [20], and others. These algorithms transform the target detection problem into a regression problem. The image is divided into multiple grids, and each grid predicts the target's location and category. This approach offers fast detection speed and high real-time performance, but it may not yield optimal accuracy for small targets. SSD is another single-stage target detection algorithm that utilizes multi-scale feature maps. SSD predicts the target's location and category on feature maps of varying scales and fuses these predictions to obtain the final target detection outcome. This algorithm provides fast detection speed and robust performance for small targets. RetinaNet [16] addresses the issue of category imbalance in target detection by employing focal loss. It replaces the traditional cross-entropy loss function, prioritizing challenging samples for classification. RetinaNet offers improved

detection accuracy and better generalization capability while maintaining fast detection speed. Mask R-CNN [21] is a target detection algorithm developed based on the R-CNN series. In addition to detecting the target's position and category, Mask R-CNN generates semantic segmentation results for the target. It achieves this by adding a segmentation branch to the R-CNN algorithm. Mask R-CNN provides improved semantic segmentation accuracy and higher detection accuracy. In conclusion, deep learning-based target detection algorithms offer fast detection speed and high accuracy. However, they still face challenges such as ineffective detection of small targets and category imbalance. Researchers are actively exploring new algorithms and technologies to enhance the performance of target detection.

3 Method

In this part, a new AGV visual localization framework based on Gaussian saliency heuristic will be introduced. Figure 1 shows the proposed AGV visual localization framework based on Gaussian saliency. The target will first be imaged by a monocular camera to obtain the target image. Secondly, the obtained image will enter the proposed GAGV-net network for target detection. Finally, the detection result image will be found through the feature point and the target will be located using PnP [22] algorithm. Figure 2 shows the flowchart of the AGV visual localization algorithm pro-posed in this paper. The acronyms appearing in the article are shown in Table 1.

AGV vision-aided localization is a key technology using vision localization. This technology includes visual imaging technology, target detection technology, and visual localization technology. Among them, target detection technology is the key to AGV visual localization, and accurate target detection can greatly improve the localization accuracy. Therefore, in order to greatly improve the accuracy of AGV visual localization, this paper proposes a Gaussian saliency inspired AGV visual detection network (GAGV-net) to improve the localization accuracy. Figure 3 shows the GAGV-net detection network framework proposed in this paper. In the feature extraction stage of the target, the input image will first go through the backbone network for initial feature extraction. In order to obtain better initial feature extraction effect and lightweight network, GAGV-net

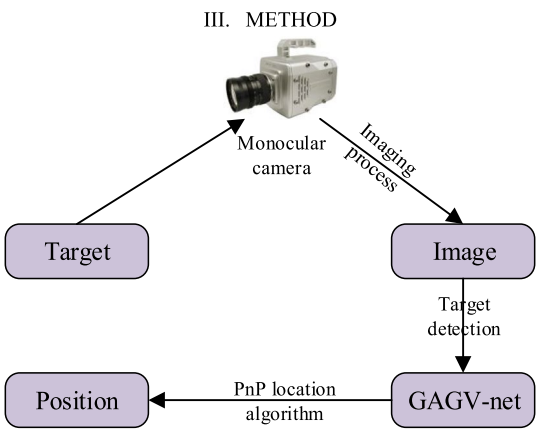

Fig. 1 The proposed AGV vision localization framework

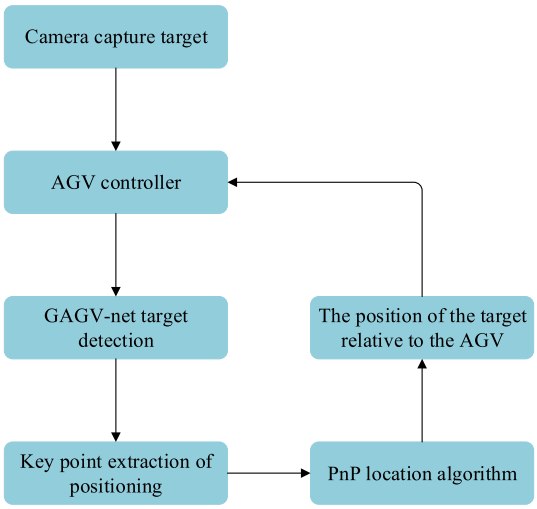

Fig. 2 The flowchart of proposed AGV vision localization algorithm

Table 1 Acronyms description

AGV	Automated-guided vehicle
GAGV-net	Gaussian saliency AGV vision detection network
UWB	Ultrawide band
PnP	Perspective-n-points
Conv	Convolutional
BN	Batch normalization
AGV	Automated-guided vehicle

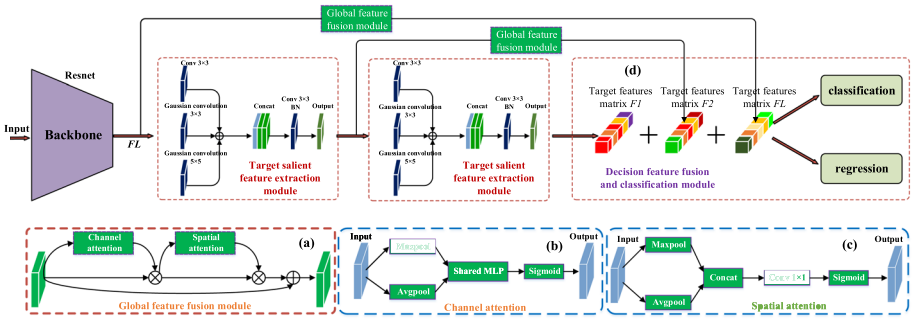

Fig. 3 The proposed GAGV-net. The joint multi-scale classification module includes two parts: **a**, **d**, where **a** is the global feature fusion module and **d** is the decision feature fusion and classification module. Global feature fusion module integrates global features by introducing attention mechanism to improve the feature expression of the target. **b**, **c** are channel attention operation and spatial attention operation in global feature fusion module, respectively

selects ResNet [23] network as the backbone network. After the input image passes through the backbone network, the feature matrix of the target is obtained.

Then, the obtained initial target features will be extracted through two target salient feature extraction modules to obtain the salient features of the target. In the classification and regression stage, in order to enrich the feature expression of the target and

obtain excellent detection performance, the extracted initial features and salient features of the target will be classified and regressed through the joint multi-scale classification module, and finally the detection results will be obtained. The network parameters of the backbone network are shown in Table 2.

3.1 Problem definition

In order to provide a more intuitive analysis of the AGV visual localization problem we are addressing, we have formulated it as follows.

$I_{\text{in}} \in R^{H\times W\times C}$ represents an input image, where H, W, and C represent the height, width, and number of channels of the image, respectively. $y \in \{1, 2, 3, \ldots K\}$ represents the category of the target in the input image, where K is the total number of target categories; $b \in R^4$ represents the bounding box of the target in the input image, where $b = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ defines the coordinates of the top left and bottom right corners of the target. The target detection network can be represented as function $f_{\text{det}}(\cdot)$, which detects the target in the input image I_{in} and accurately marks its bounding box b. The PnP localization algorithm can be represented as a function $f_{\text{loc}}(\cdot)$, which receives the prediction of the target boundary box b and outputs the coordinate position P of the target in the real world. P_1 is the actual target location. Therefore, our goal is to find an optimal model $f_{\text{det}}(\cdot)$ that minimizes the error between P and P_1. The specific formula for our objective function is as follows.

$$f_{\text{det}} = \arg\min L\{P, P_1\}, \tag{1}$$

$$P = f_{\text{loc}}(b), \tag{2}$$

$$y, b = f_{\text{det}}(I_{\text{in}}), \tag{3}$$

where $L\{\cdot\}$ is the mean square error function.

3.2 Target salient feature extraction module

In the field of target detection [24–28], the performance of the detection network is limited by its ability to extract target features. Deeper networks generally have better target feature extraction ability, but they often result in larger model sizes and slower detection speeds. These drawbacks are not suitable for AGV visual detection tasks. To address this,

Table 2 The network parameters of the backbone network

Level	Size/filter	Pool size	Stride	Padding	Output (input)
Input layer	–				1
Layer 1	3 × 3/6	2	1	1	6
Layer 2	3 × 3/12	2	1	1	12
Layer 3	3 × 3/20	2	1	1	20
Layer 4	3 × 3/40	2	1	1	40
Layer 5	3 × 3/64	2	1	1	64
Layer 6	3 × 3/128	2	1	1	128
Layer 7	3 × 3/256	2	1	1	256

we propose a feature extraction module for object saliency in this paper. This module extracts the salient features of the target, enriching the feature expression and improving the network's feature extraction capability. Instead of deepening the network depth, which can lead to parameter redundancy, this module reduces the number of parameters required for model fitting while maintaining excellent feature extraction ability.

Inspired by the human visual mechanism, where the target stands out from the surrounding environment, we simulate this using Gaussian convolution. In convolutional networks, as the receptive field decreases, the scale of the target in the feature matrix decreases but its significance increases. We leverage Gaussian convolution to extract salient features of the target, enhancing the contrast between the target and the background.

By using different scales of Gaussian convolution kernels, the proposed object salient feature extraction module improves the contrast of the target. This allows the network to focus more on the target rather than the background. The module employs Gaussian convolutions with different kernel sizes to extract saliency features of different scales, enriching the feature expression of the object. Figure 4 illustrates the structure of the target salient feature extraction module, and the specific feature extraction process can be described as follows.

$$F_{\mathrm{L}} = \mathrm{Backbone}(I_{\mathrm{in}}) \tag{4}$$

$$F_{\mathrm{o}} = \mathrm{Conv}_{3\times3}(F_{\mathrm{l}}) \tag{5}$$

$$F_{\mathrm{S1}} = \mathrm{Gconv}_{3\times3}(F_{\mathrm{L}}) \tag{6}$$

$$F_{\mathrm{S2}} = \mathrm{Gconv}_{5\times5}(F_{\mathrm{L}}) \tag{7}$$

$$G_{\mathrm{out}} = \mathrm{Conv}_{3\times3}(\mathrm{Concat}(F_{\mathrm{o}}, F_{\mathrm{S1}}, F_{\mathrm{S2}})) \tag{8}$$

where I_{in} is the input image, Backbone() is the backbone network, and F_{L} is the initial target feature. F_{o} is a feature obtained by ordinary convolution, and F_{S1} and F_{S2} are significant features. Concat() is the splicing operation, and $\mathrm{Gconv}_{5\times5}(\cdot)$ and $\mathrm{Gconv}_{3\times3}(\cdot)$

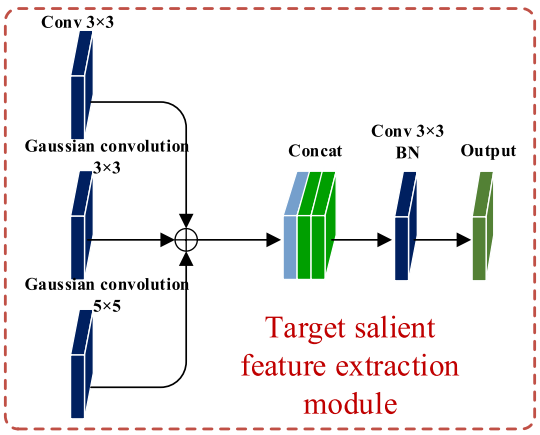

Fig. 4 Target salient feature extraction module

are Gaussian convolutions of different scales. G_{out} is the output of the target salient feature extraction module. The kernel function of Gaussian convolution is shown below.

$$G(m,n) = T \cdot \exp\left\{-\frac{1}{2}\left[\left(\frac{m-m_{\text{c}}}{\theta_m}\right)^2 + \left(\frac{n-n_{\text{c}}}{\theta_n}\right)^2\right]\right\} \tag{9}$$

where $G(m,n)$ is the kernel function of the Gaussian convolution kernel, T is the amplitude of the kernel function, and θ_m and θ_n are the scale parameters of the Gaussian convolution kernel function. m_{c} and n_{c} are the central coordinates of the kernel function.

In the GAGV-net, considering the computational cost, we set the size of the Gaussian convolution kernel to 3×3 and 5×5. It is worth noting that θ_m and θ_n in $G(m,n)$ are constantly updated and optimized in the network training process and do not need to be set manually. Compared with ordinary convolution, in the training process of the network, only the scale parameters θ_m and θ_n of the Gaussian saliency convolution are constantly updated, hence the kernel function of the Gaussian convolution is always a two-dimensional Gaussian distribution. This enables GAGV-net to effectively extract the salient features of the target.

3.3 Joint multi-scale classification module

After the backbone network and target salient feature extraction module, the rich features of the target are fully extracted. In order to further enrich the target feature expression and improve the detection performance, we designed a joint multi-scale classification module to achieve the final detection. The ability of object feature expression of network is the most important factor to determine the detection accuracy. Therefore, different from the existing target detection methods, we perform feature fusion on different levels of target features after target feature extraction in the network. We use the proposed joint multi-scale classification module to fuse the target features, so as to enrich the target feature representation.

Our joint multi-scale classification module is shown in Fig. 3. As can be seen from the figure, the joint multi-scale classification module includes two parts: global feature fusion module and decision feature fusion and classification module. First of all, the global feature fusion module integrates the global features by introducing the attention mechanism [29]. The feature expression of the target in this module is greatly enriched, which provides the basis for accurate classification. In addition, in order to further enrich the feature expression of the target, we fused the global fusion features of different depths in the decision feature fusion and classification module, and finally realized the detection of the target.

Attention mechanism has been widely used in the field of target detection, so in this paper, we use attention mechanism to design a global feature fusion module to enrich the feature expression of the target. Figure 3b shows the proposed global feature fusion module, which includes channel attention and spatial attention. Its specific operation can be expressed as follows.

$$F_{\text{c}} = F_{\text{in}} \times C_{\text{attention}}(F_{\text{in}}), \quad F_{\text{s}} = F_{\text{c}} \times S_{\text{attention}}(F_{\text{c}}) \tag{10}$$

$$F_{\text{out}} = F_{\text{s}} + F_{\text{in}} \tag{11}$$

where F_{out} is the global feature after fusion, F_{in} is the input features, F_{c} is the feature matrix after channel attention operation, and F_{s} is the feature matrix after spatial attention operation. $C_{\text{attention}}()$ and $S_{\text{attention}}()$ are channel and spatial attention operations, respectively.

The feature expression of the target has been greatly improved after the global feature extraction by a module. However, in order to fully utilize the feature of the target to improve the detection performance, we designed the decision feature fusion and classification module to fuse the global features of different depths, and then more fully utilize the feature of the target.

The specific operation process of the decision feature fusion and classification module can be described as follows.

$$F_{\text{J}} = \text{add}(F_{\text{S1}}, F_{\text{S2}}, F_{\text{L}}) \tag{12}$$

where F_{J} is the multi-scale feature of the fused target. F_{S1}, F_{S2} and F_{L} are shallow features, which can be seen in Fig. 5.

The feature matrix after feature fusion will carry out the frame regression and target classification of the target, and finally obtain the detection results.

3.4 Loss function

In order to constrain the training of the network, we choose the cross entropy [24, 30] and IOU loss function. The IOU loss function is used to constrain the border regression of the network, and its calculation method can be shown in Fig. 6 and Formula (8). The cross-entropy loss function is used to constrain the classification task of the network.

$$\text{IOU} = \left| \frac{G \cap D}{G \cup D} \right| \tag{13}$$

$$L_{\text{Box}} = 1 - \text{IOU} \tag{14}$$

$$L = W_{\text{box}} L_{\text{box}} + W_{\text{cls}} L_{\text{cls}} \tag{15}$$

where L_{Box} is the border loss, L_{cls} is the classification loss, and L is the total loss. W_{box} is the weight of border loss, and W_{cls} is the weight of classification loss.

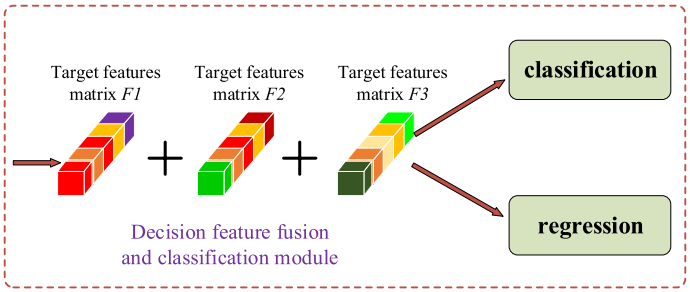

Fig. 5 Decision feature fusion and classification module

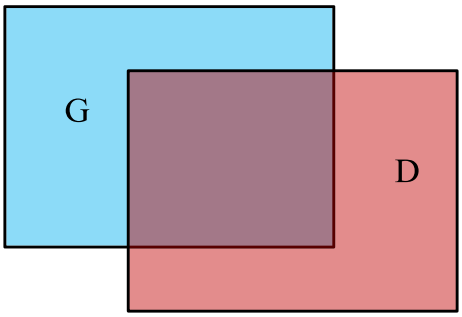

Fig. 6 Schematic diagram of IOU calculation, G is the groundtruth, D is the detection result

Fig. 7 Images in the training dataset

4 Experimental results

4.1 Experimental setup

In order to verify the effectiveness and progressiveness of the proposed methods, we conducted experimental verification. To validate the proposed method, we collected a large amount of data to train our detection model. The training set used in the experiment consists of 2000 images collected by AGV's visual camera, containing many scenes. Specifically, we use the camera on AGV to capture images and create a standard dataset for model training. The collected dataset targets include electricity meters, key, 0–9 and F digital displays, totaling 13 categories of targets. Figure 7 shows some of the images in the dataset. When training the network, we use 50% of the data set to train the network, 30% as test data, and 20% as verification data. Our experiment was carried out on a computer with NVIDIA 3080ti graphics card. The software is installed with Python 3.7, Python 1.1, and Pycharm 2021.2.5.

In the field of target detection, recall (R) and precision (P) [24–28, 30] are two important indicators to verify the performance of the method. P and R are defined as follows.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{16}$$

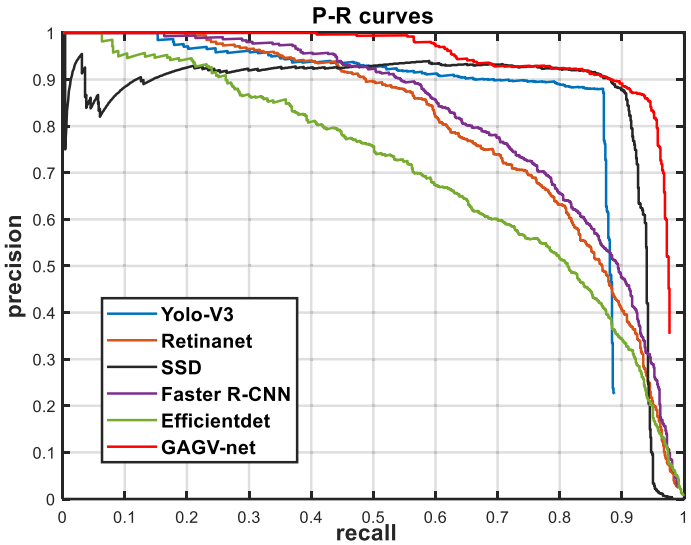

Fig. 8 *P–R* curves of different comparison methods

Table 3 *P* and *R* of different detection methods

Method	*P*	*R*	FPS
YOLO-V3	0.75	0.72	11.01
SSD	0.68	0.79	7.80
Retinanet	0.83	0.83	18.41
Efficientdet	0.81	0.77	19.91
Faster R-CNN	0.74	0.82	7.29
OUR	0.93	0.88	47.29

$$R = \frac{TP}{TP + FN} \tag{17}$$

where TP is true positive, FP is false positive, FN is false negative.

4.2 Comparative experimental results and discussion

In order to verify the effectiveness and progressiveness of the proposed method, we compared it with the current state of the art (SOTA) method. When conducting comparative experiments, for fairness, all comparative methods were retrained on our dataset. Moreover, the experimental settings for these comparative methods are the same as those in their original literature. These methods include Faster R-CNN [24], SSD [31], Yolo-V3 [32], Retinanet [33], Efficientdet [34]. The experimental results are shown in Fig. 8 and Table 3. Figure 8 shows the PR curve of each comparison method, from which we can clearly see that the proposed method has the best detection performance. Table 3 shows the *P* and *R* values of different comparison methods. From the table, we can clearly see that compared with the existing methods, the proposed methods are superior to the existing advanced detection methods in all performance indicators. For the detection accuracy *P*, the detection method proposed in this paper is improved by at least 12%. For the recall rate *R*, the detection method proposed

in this paper has been improved by at least 5%. In addition, Table 3 also shows the detection speed of different comparison methods, from which we can see that the proposed method has the highest detection speed, the detection speed is improved by at least 27.38 FPS. Figure 9 shows the detection results of the detection algorithm on the AGV trolley. It is worth noting that in Fig. 9, the detection results were obtained through the camera on the AGV, and we captured the detection results on the upper computer software for display. From the figure, we can clearly see that our algorithm has a high detection accuracy and can effectively deal with targets of different sizes. This also proves that the proposed Gaussian convolution can effectively extract the salient features of the target, enabling the detection network to focus on the target itself, thereby greatly improving detection performance. In addition, since the proposed method utilizes Gaussian convolution $G(m, n)$ to extract the salient features of the target, theoretically, the detection performance will be affected by Gaussian noise. However, due to the dynamic adaptive setting of θ_m and θ_n in Gaussian convolution in the detection network, this greatly reduces the impact of Gaussian noise.

To verify the performance of the proposed method in positioning accuracy, we conducted experimental verification. As shown in Fig. 10, the detection performance of the proposed method in visual positioning is shown, with the blue curve representing the actual target position and the red curve representing our prediction results. It can be clearly seen from the figure that the proposed method can locate the target with minimal error. It is worth noting that for the convenience of display, x and y in Fig. 10 are the normalized coordinates of the target position.

4.3 Ablation experiment

In order to further verify the effectiveness of the proposed method, we conducted ablation experiments. The experimental results are shown in Table 4. From the table, we can clearly see that the target salient feature extraction module and the joint multi-scale classification module proposed in GAGV-net are both effective in improving the detection performance.

Fig. 9 The detection result image of the GAGV-net deployed on the AGV trolley

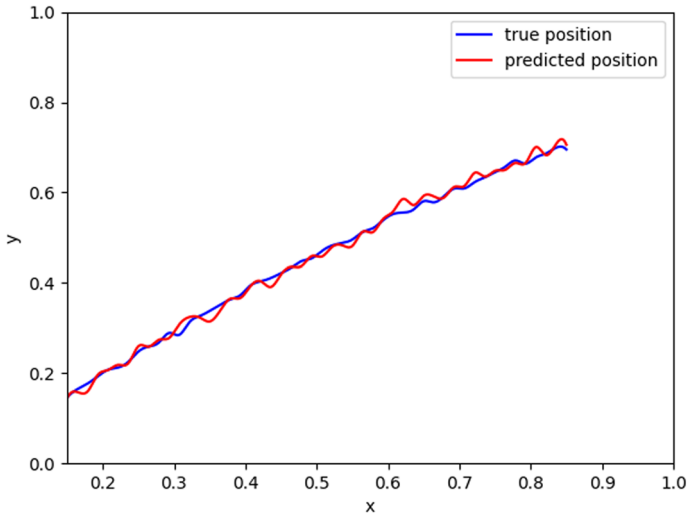

Fig. 10 Visual localization performance of the proposed method

Table 4 Ablation results

Target salient feature extraction module	Joint multi-scale classification module	*R*	*P*
√		0.69	0.72
	√	0.74	0.79
√	√	0.88	0.93

5 Conclusion

In order to solve the problems of poor detection performance and slow detection speed of existing AGV-based visual localization methods. This paper proposes an AGV visual inspection network GAGV-net based on visual saliency. The network enriches the feature representation of the target through the designed feature extraction module of the target saliency, thereby reducing the parameters required for model fitting. At the same time, in order to improve the detection accuracy, a joint multi-scale classification module is proposed in the GAGV-net network, which improves the detection accuracy by fusing features of different depths of the target. The experimental results show that the proposed method has better detection performance than the existing advanced methods.

Acknowledgements
This work was supported by the China Tobacco Henan Science and Technology Project (AW202122).

Author contributions
To address the issues of poor detection accuracy and the large number of target detection model parameters in existing AGV monocular vision location detection algorithms, this paper presents an AGV vision location method based on Gaussian saliency heuristic. The proposed method introduces a fast and accurate AGV visual detection network called GAGV-net. In the GAGV-net network, a Gaussian saliency feature extraction module is designed to enhance the network's feature extraction capability, thereby reducing the required output for model fitting. To improve the accuracy of target detection, a joint multi-scale classification and detection task header is designed at the stage of target frame regression to classification. This header utilizes target features of different scales, thereby enhancing the accuracy of target detection. Experimental results demonstrate a 12% improvement in detection accuracy and a 27.38 FPS increase in detection speed compared to existing detection methods. Moreover, the proposed detection network significantly reduces the model's size, enhances the network model's deployability on AGVs, and greatly improves detection accuracy.

Funding
China Tobacco Henan Science and Technology Project (AW202122).

Availability of data and materials
The data supporting the findings of this study are included in the article. Please contact the author if any help needs.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
We declare that the funding "China Tobacco Henan Science and Technology Project (AW202122)" provided in this article does not affect the results or performance of this study.

Received: 19 September 2023 Accepted: 15 January 2024
Published online: 19 March 2024

References

1. T. Xue, P. Zeng, H. Yu, A reinforcement learning method for multi-agv scheduling in manufacturing, in *2018 IEEE International Conference on Industrial Technology (ICIT)*, pp. 1557–1561 (2018)
2. Y. Mao, *Research and Software Implementation of Electromagnetic Guided AGV Single Machine Control* (Kunming University of Science and Technology, Kunming, 2006)
3. Y. Shen, *Research on Laser Guided AGV Vehicle Control System* (Hefei University of Technology, Hefei, 2007)
4. X. Lin, Structure design and control strategy of magnetic navigation AGV. J. Jilin Inst. Chem. Technol. **036**(007), 30–35 (2019)
5. B. Wang, *Research and Implementation of Magnetic Navigation AGV Vehicle Control System* (Guilin University of Electronic Technology, Guilin, 2019)
6. J. Kang, J. Lee, H. Eum, C.-H. Hyun, M. Parks, An application of parameter extraction for AGV navigation based on computer vision, in *2013 10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 622–626 (2013)
7. X. Ding, D. Zhang, L. Zhang, L. Zhang, C. Zhang, B. Xu, Fault detection for automatic guided vehicles based on decision tree and LSTM, in *2021 5th International Conference on System Reliability and Safety (ICSRS)*, pp. 42–46 (2021)
8. P. Kuang, Q. Zhu, G. Liu, Real-time road lane recognition using fuzzy reasoning for AGV vision system, in *2004 International Conference on Communications, Circuits and Systems (IEEE Cat. No.04EX914)*, pp. 989–993 (2004)
9. D. Yang, C. Su, H. Wu, X. Xu, X. Zhao, Shelter identification for shelter-transporting AGV based on improved target detection model YOLOv5. IEEE Access **10**, 119132–119139 (2022)
10. S. Liu, M. Xiong, W. Zhong, H. Xiong, Towards industrial scenario lane detection: vision-based AGV navigation methods, in *2020 IEEE International Conference on Mechatronics and Automation (ICMA)*, pp. 1101–1106 (2020)
11. J. Dong, X. Ren, S. Han, S. Luo, UAV vision aided INS/odometer integration for land vehicle autonomous navigation, in *IEEE Transactions on Vehicular Technology*, pp. 4825–4840 (2022)
12. L. Li, Y. -H. Liu, M. Fang, Z. Zheng, H. Tang, Vision-based intelligent forklift automatic guided vehicle (AGV), in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 264–265 (2015)
13. R. Girshick, "Fast R-CNN", in *2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile*, pp. 1440–1448 (2015)
14. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 779–788 (2016)
15. W. Liu, et al., SSD: single shot multibox detector, in *ECCV* (2016)
16. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, pp. 2999–3007 (2017)
17. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149 (2017)
18. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA*, pp. 580–587 (2014)
19. J. Redmon, A. Farhadi, YOLOv3: an incremental improvement. arXiv e-prints (2018)
20. X. Zhu, S. Lyu, X. Wang, Q. Zhao, TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios, in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada*, pp. 2778–2788 (2021)
21. K. He, G. Gkioxari, P. Dollár, R. Girshick, "Mask R-CNN", in *2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy*, pp. 2980–2988 (2017)
22. V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: an accurate O(n) solution to the PnP problem. Int. J. Comput. Vis. **81**, 155–166 (2009)
23. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *IEEE CVPR*, pp. 770–778 (2016)

24. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1137–1149 (2017)
25. T. Chen, Z. Lu, Y. Yang, Y. Zhang, B. Du, A. Plaza, A siamese network based U-net for change detection in high resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. **15**, 2357–2369 (2022)
26. T. Kong, A. Yao, Y. Chen, F. Sun, HyperNet: towards accurate region proposal generation and joint object detection, in *IEEE CVPR*, pp. 845–853 (2016)
27. T. Shi, N. Boutry, Y. Xu, T. Géraud, Local intensity order transformation for robust curvilinear object segmentation. IEEE Trans. Image Process. **31**, 2557–2569 (2022)
28. H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y. Chen, J. Wu, UNet 3+: a full-scale connected UNet for medical image segmentation, in *IEEE ICASSP*, pp. 1055–1059 (2020)
29. W. Wang, X. Tan, P. Zhang, X. Wang, A CBAM based multiscale transformer fusion approach for remote sensing image change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **15**, 6817–6825 (2022)
30. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: exceeding yolo series in 2021. arXiv:2107.08430 (2021)
31. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A.C. Berg, SSD: single shot multibox detector. arXiv:1512.02325 [cs], 9905:21–37 (2016)
32. J. Redmon, A. Farhadi, *Yolov3: An Incremental Improvement*, CoRR, vol. abs/1804.02767 (2018)
33. T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 318–327 (2020)
34. M. Tan, R. Pang, Q.V. Le, EfficientDet: scalable and efficient object detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 10778–10787 (2020)

Publisher's Note