Article

Not peer-reviewed version

# Fault-Tolerant Control of AGVs via Deep Feature Enhancement and Multi-Source Verification in Complex Industrial Environments

Yazhou Zhou , Shanshan Peng , Zhennan Zhou , Yun Wang * , Nan Zhou , Biao Zhou , Fei Shan

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1455.v1

Keywords: automated guided vehicle; YOLOv8; anomaly detection; adaptive decision making; robust sensing; industrial material handling

 doi:10.20944/preprints202604.1455.v1



*Article*

# Fault-Tolerant Control of AGVs via Deep Feature Enhancement and Multi-Source Verification in Complex Industrial Environments

**Yazhou Zhou [1], Shanshan Peng [2], Zhennan Zhou [2], Yun Wang [1,*], Nan Zhou [2,3], Biao Zhou [2] and Fei Shan [2,3]**

[1] School of Mechanical Engineering, Jiangsu University, 301 Xuefu Road, Zhenjiang 212013, China
[2] Suzhou Yuanzi Intelligent Technology Co., Ltd., 381 Suzhou Avenue, Suzhou 215000, China
[3] College of Oceanography, Jiangsu University of Science and Technology, 2 Mengxi Road, Zhenjiang 212134, China
* Correspondence: wangyun@ujs.edu.cn

**Abstract**

To address the issue of 2D laser-guided automated guided vehicles (AGVs) in industrial intelligent material handling scenarios being susceptible to interference from changes in lighting and complex obstacles, leading to abnormal positioning and mapping and frequent false stops, this paper designs a lightweight, multi-dimensional perception and anti-false-stop YOLOv8 anomaly recognition network, achieving accurate identification of various interferences in complex environments. An adaptive decision-making fault-tolerant control algorithm is proposed, introducing a temporal logic verification and dynamic threshold adjustment mechanism to achieve real-time dynamic switching of obstacle avoidance levels, ensuring efficient coordination between perception decision-making and control execution. An AGV anomaly detection sample set suitable for complex industrial scenarios is constructed, providing reliable data support for model optimization and accuracy evaluation. Finally, real-world deployment verification in a real electronics factory environment shows that this method reduces the vehicle false-stop rate and improves task handling efficiency. This research effectively solves the robust perception problem of AGVs in complex industrial environments and has significant engineering application value.

**Keywords:** automated guided vehicle; YOLOv8; anomaly detection; adaptive decision making; robust sensing; industrial material handling

## 1. Introduction

Against the backdrop of deep transformation in intelligent manufacturing, automated guided vehicles (AGVs), as the core carrier of flexible logistics, directly determine the operational efficiency of factories based on the robustness of their perception [1,2]. However, in industrial scenarios with complex lighting and dense equipment, such as electronics factories, mainstream laser guidance solutions face severe challenges: due to their high dependence on geometric features, they are prone to echo voids or "false wall" effects when encountering specular reflection, drastic changes in lighting, or fence structures [3]. This ambiguity at the perception level leads to frequent false stops, becoming a bottleneck restricting the efficient operation of the system [4].

To address the aforementioned limitations, researchers have conducted extensive research in the field of multi-sensor fusion. Yuan et al. [5] proposed a multi-source fusion positioning model integrating odometer, IMU, lidar and ultra-wideband (UWB) to address the problem of cumulative errors that easily occur between two-dimensional lidar and inertial navigation systems indoors. They used the unscented Kalman filter (UKF) algorithm to achieve deep coupling of multi-dimensional data. Song et al. [6] developed a hybrid sensing pipeline, which used a fully convolutional neural network

 doi:10.20944/preprints202604.1455.v1



(FCN) to perform semantic segmentation on camera and lidar data and combined it with Kalman filtering to achieve asynchronous sensor state tracking, verifying the effectiveness of cross-modal feature complementarity in improving sensing reliability. Zhou et al. [7] systematically reviewed various navigation methods and pointed out that multi-sensor information fusion (MSIF) is an inevitable trend to overcome the navigation stability of AGVs in complex industrial environments. They also discussed the balance between efficiency and real-time performance in heterogeneous data processing.

However, there are still three technical gaps in practical engineering applications: First, insufficient dataset support. Although the KITTI dataset established by Geiger et al. [8] laid the foundation for outdoor perception, it focuses on conventional road environments and lacks a special sample library for industrial sites (such as mirror images and light and shadow in narrow alleys), which limits the generalization ability of the model under extreme conditions. Second, insufficient decoupling between physical and semantic features. Although Yang et al. [9] and Qian et al. [10] optimized the recognition accuracy of YOLOv5 through attention mechanism and gated convolution, such models still perform visual tasks in isolation. Due to the lack of a consistency constraint mechanism for deep coupling between laser physical echo and visual semantics, the system is difficult to eliminate "pseudo-obstacle" interference from a physical nature. Finally, the lack of coordination between perception uncertainty and control execution. Existing systems generally lack fault tolerance processing for perception uncertainty. Vassilev et al. [11] pointed out that perception models have significant inference variability under severe lighting conditions, while traditional logic triggers forced braking as soon as it detects an anomaly. As Li et al. [12] stated, rigid control without a multi-stage deceleration strategy can lead to a conflict between the reference path and the execution capability; furthermore, Han et al. [13] pointed out that such an architecture lacking robust fault tolerance mechanism can exacerbate mechanical wear and lead to path tracking errors.

To address the aforementioned challenges, this paper proposes a robust AGV perception framework that integrates multi-dimensional feature constraints and adaptive decision-making. The main contributions of this paper are as follows:

1. A sample set for AGV anomaly detection in complex industrial scenarios is constructed, including real-world scene images of high-dynamic obstacles, ground dynamic interference, geometrical ambiguity, mirror image interference, and severe light and shadow interference, providing a data foundation for subsequent model optimization and accuracy evaluation.

2. Based on the stability of YOLOv8, a lightweight, multi-dimensional perception and anti-false-stop YOLOv8 anomaly recognition network is constructed by deeply integrating visual semantics, laser physical characteristics, and temporal motion constraint criteria. This network achieves accurate identification of various environmental disturbances in complex industrial scenarios, significantly improving the continuity of AGV operations.

3. An adaptive decision-making fault-tolerant control algorithm is proposed. Addressing the "discontinuity" and "uncertainty" in the perception environment, a temporal logic verification and dynamic threshold adjustment mechanism is introduced, enabling the system to switch obstacle avoidance levels in real time based on target confidence. This algorithm effectively avoids frequent emergency braking of the system under boundary conditions, ensuring efficient coordination between perception decision-making and control execution in complex dynamic scenarios.

## 2. System Overall Framework

### *2.1. Framework Overview and Operating Logic*

The system operation logic proposed in this paper is shown in Figure 1, and is mainly divided into the following three functional stages:

 doi:10.20944/preprints202604.1455.v1



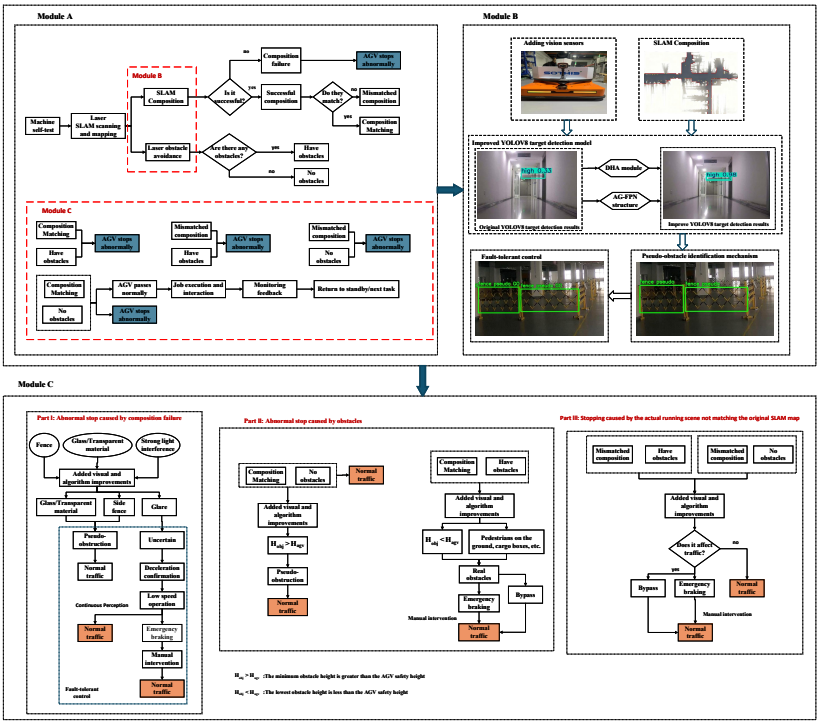



**Figure 1.** System overall framework diagram.

Basic navigation and status monitoring stage (Module A): The vehicle relies on 2D LiDAR to execute the SLAM algorithm to achieve environmental mapping and autonomous localization. The system monitors the raw sensor data and positioning status in real time. Once an anomaly in the LiDAR point cloud or a potential obstacle signal is detected, the visual-assisted perception mode is immediately triggered.

Visual Enhancement and Multidimensional False Obstacle Recognition Stage (Module B): This module is the core recognition layer of the system, designed to eliminate perceptual ambiguity. First, an improved lightweight YOLOv8 algorithm (integrating AG-FPN and DHA modules) is used to extract high-dimensional semantic features, achieving accurate targeting of multi-scale targets in complex industrial environments. Subsequently, the system performs consistency verification between visual semantics and laser echo intensity, geometric height, and time series. Through this deep coupling of physical characteristics and semantic features, the system can accurately isolate "false obstacles" such as strong light and virtual images, achieving essential decoupling from environmental interference.

Adaptive Decision-Making and Hierarchical Fault-Tolerant Control Stage (Module C): This stage is responsible for executing a closed-loop response to the perceptual decision results. Based on a risk assessment model, the system triggers a hierarchical response strategy according to target attributes and position confidence. By dynamically switching between adaptive deceleration, path detour, and emergency braking, the system ensures that the AGV maintains operational continuity under uncertain conditions, effectively reducing unnecessary stoppages.

### *2.2. Improve Detection Network*

The detection network based on YOLOv8 designed in this paper mainly consists of the following three parts:

1. **Backbone layer**: Extracts basic features of the AGV operating environment through multi-layer convolution and residual structure [14], including edge, texture and geometric information, to provide multi-level semantic representation for subsequent feature fusion and detection.

 doi:10.20944/preprints202604.1455.v1



2. **Neck layer**: Adopts improved AG-FPN structure and improves the transmission and fusion effect of cross-scale information by reducing computational complexity through efficient feature reuse [15].
3. **Head front end**: Embeds DHA module in the front end of the detection head and improves the feature discrimination and capture ability of the model for abnormal regions and difficult-to-detect targets by parallel fusion of saliency, channel and spatial attention mechanisms. It adopts learnable weights [16] to adaptively enhance the semantic representation and spatial structure information of multi-scale features.

#### 2.2.1. AG-FPN Module

To address the feature degradation problem of small targets or abnormal areas caused by complex background interference and uneven illumination in industrial intelligent material handling tasks, this paper designs an Adaptive Ghost Feature Pyramid Network (AG-FPN), as shown in Figure 2. The core logic of AG-FPN lies in reducing the computational load through "redundancy suppression" and enhancing the semantic consistency of cross-scale features through "adaptive fusion". This module consists of three key components: a multi-level adaptive ghost pyramid (MAGP), cross-layer adaptive fusion (CAF), and semantic consistency reconstruction (SCR).

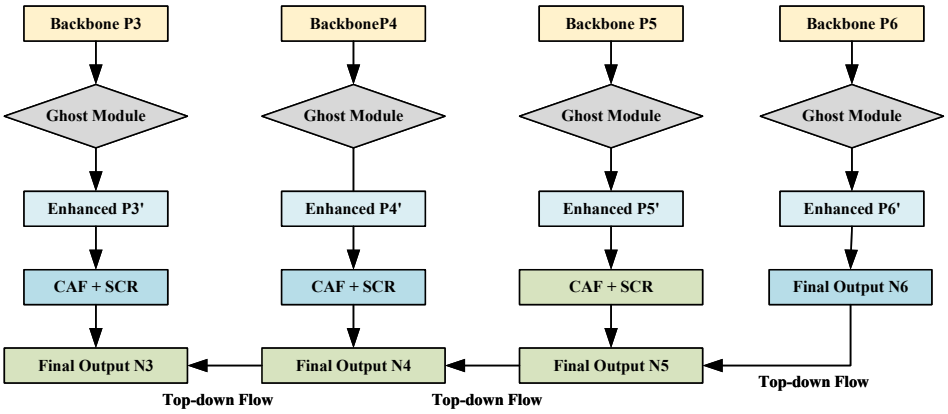



**Figure 2.** AG-FPN structure.

1. MAGP: Efficient Feature Generation Based on Redundancy Suppression

In industrial scenarios, there is significant semantic overlap between feature channels. To maintain feature richness at a minimal computational cost, this paper introduces the MAGP module. The feature extraction process is reshaped through the coupling logic of "intrinsic + ghost" features:

$$P_{enhanced} = \text{Concat}\left(\underbrace{f_{conv}(P_{in})}_{Y'}, \underbrace{\Phi(Y')}_{Y''}\right) \tag{1}$$

where $P_{enhanced}$ represents the enhanced feature layer. Through the "cheap feature reuse" mechanism, it ensures that edge devices possess high-dimensional spatial feature expression without increasing computational overhead. $P_{in}$ denotes the original feature layer output by the backbone network; $f_{conv}$ represents the intrinsic feature extraction operation; and $\Phi$ denotes the linear mapping operator.

 doi:10.20944/preprints202604.1455.v1



2. CAF: Cross-layer Adaptive Fusion Based on Weight Games

Traditional feature pyramids use equal weight stacking [17], which often causes high-level semantic information to overshadow the fine details of small obstacles in shallow layers. To address this, this paper designs the Cross-layer Adaptive Fusion (CAF) module, which transforms the feature fusion process into a self-adaptive game logic:

$$F_{fused} = \omega_1 \cdot P_i + \omega_2 \cdot \text{Upsample}(P_{i+1}) \tag{2}$$

where $F_{fused}$ represents the self-adaptive fusion feature flow, achieving dynamic allocation of perceptual focus. When detecting small anomalies, $\omega_1$ is automatically increased to enhance spatial details; when identifying large-scale targets, $\omega_2$ is increased to lock in semantic information. $P_i$ denotes the shallow features of the current scale; $\omega_1$ and $\omega_2$ are the adaptive weight coefficients, satisfying $\omega_1 + \omega_2 = 1$, which are dynamically generated through the global perception excitation network; Upsample represents the upsampling operator used to align the spatial resolution of different scales; and $P_{i+1}$ represents the deep features of the previous level.

3. SCR: Semantic Consistency Reconstruction Based on Residual Mapping

Although the dynamic weighting of the CAF module improves flexibility, it also disrupts the numerical distribution of signals, which can lead to numerical instability during the training process. To address this, this paper introduces the SCR module, which is responsible for normalizing and correcting the fused signals:

$$F_{out} = BN(F_{fused} + \text{Res}(F_{fused})) \tag{3}$$

where $F_{out}$ represents the reconstructed output feature. It provides the subsequent detection head with feature support that possesses both multi-scale semantic coupling and high numerical stability, significantly improving the system's discrimination sensitivity for difficult-to-detect targets. $F_{fused}$ denotes the dynamically weighted fusion signal from the previous module; Res represents the residual operator; and $BN$ represents the Batch Normalization layer.

2.2.2. DHA Module

In the anomaly detection task of AGV, small targets and local subtle anomalies are often difficult to capture effectively. To this end, this paper proposes a Dynamic Hybrid Attention (DHA) module, as shown in Figure 3. This module integrates saliency attention [18], channel attention and spatial attention [19], and traditional feature pyramids weight stacking [17], and introduces learnable dynamic fusion weights to enhance the response of feature maps to abnormal regions. Its physical operation logic is as follows:

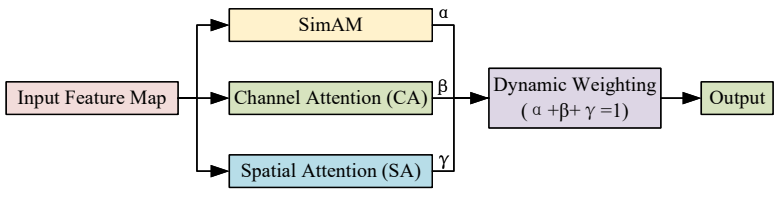



**Figure 3.** Structure of the Dynamic Hybrid Attention (DHA) module.

The input feature map $F \in \mathbb{R}^{C \times H \times W}$ is first fed into three parallel perceptual branches for multi-dimensional feature extraction:

 doi:10.20944/preprints202604.1455.v1



1. SimAM Saliency Attention Branch

To highlight local high-contrast targets without increasing the number of model parameters, this paper utilizes the "spatial inhibition" principle from neuroscience. The SimAM saliency energy function is introduced as:

$$e_t(w_t, b_t, y, x_i) = \frac{1}{M-1} \sum_{i=1}^{M-1} (y_0 - (w_t x_i + b_t))^2 + (y_t - (w_t t + b_t))^2 \tag{4}$$

The formula calculates the linear separability between the current neuron and its neighborhood by minimizing the energy function $e_t$. A lower energy value indicates a higher discriminability of the point relative to its background. $x$ and $x_i$ represent the current target neuron and other neurons within its neighborhood, respectively. $w_t$ and $b_t$ denote the linear transformation weight and bias of the neuron.

2. Channel Attention Branch ($M_c$)

Different channels of the feature map carry different semantic information (e.g., color, edges, specific textures). To suppress redundant noise in industrial environments and strengthen key channels, the channel calibration logic is introduced:

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \tag{5}$$

This aims to suppress channels carrying redundant background noise and significantly enhance the response of feature channels carrying key information about abnormal targets. AvgPool denotes global average pooling; MaxPool denotes global max pooling; and MLP represents the shared Multi-Layer Perceptron used to learn the non-linear dependency relationships between various channels.

3. Spatial Attention Branch ($M_s$)

To preserve precise location information of abnormal targets during cross-channel interactions, spatial weight mapping is introduced:

$$M_s(F) = \sigma\left(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])\right) \tag{6}$$

The generated weight map enables pixel-wise positional weighting of the feature map. By emphasizing the geometric topological features of abnormal regions, it significantly enhances the system's spatial localization robustness under complex structural occlusions. $[;]$ denotes the feature concatenation operation along the channel dimension; $f^{7\times7}$ represents a convolution operator with a large receptive field.

4. Dynamic Weight Fusion Logic: Adaptive Scheduling of Perceptual Strategies

To achieve optimal allocation of perceptual resources in different environments, the DHA module introduces learnable dynamic fusion weights $\alpha, \beta, \gamma$:

$$F' = (\alpha \cdot M_{Sim} + \beta \cdot M_c + \gamma \cdot M_s) \otimes F \tag{7}$$

A self-adaptive scheduling mechanism is constructed. For example, $\alpha$ (saliency branch) is automatically increased when lighting changes abruptly, and $\gamma$ (spatial branch) is increased when target overlap is severe. $\alpha, \beta, \gamma$ represent the contribution coefficients of the three attention branches; $\otimes$ represents the element-wise multiplication feature broadcasting operation.

*2.3. Multidimensional Pseudo-Obstacle Discrimination Model*

In industrial intelligent material handling scenarios, AGVs are often affected by environmental factors such as glass reflection, strong light interference, dynamic suspended objects, or repetitive barriers, resulting in false target recognition and frequent erroneous shutdowns. To address this

 doi:10.20944/preprints202604.1455.v1



issue, this paper constructs a multi-dimensional false obstacle discrimination mechanism based on improvements to the semantic output of the YOLOv8 detection network. This mechanism performs secondary logical verification of the detection results through deep coupling of visual, spatial, and temporal dimensions. The discrimination process is as follows:

**(1) Visual Feature Discrimination Based on Saliency Response**

This method utilizes the physical continuity differences in surface texture to filter out specular glare noise. Specular glare or reflective virtual images often manifest as fragmented edge topology in low-level features, whereas real obstacles possess closed contours. The discrimination criterion is defined as:

$$S_{vis} < \tau_v \tag{8}$$

where $S_{vis}$ represents the regional saliency weighted score output by the DHA module, which reflects the degree of aggregation of texture features and edge contrast within the candidate region. $\tau_v$ is an adaptive threshold. When the saliency score is lower than the threshold, the system initially determines the target as non-solid optical interference.

**(2) Spatial-Geometric Consistency Verification**

To address visual false alarms caused by transparent glass or specular reflections, the physical complementarity between LiDAR and visual sensors is utilized for verification [20]. While vision captures color semantics, LiDAR only detects physical entities. The discrimination criterion is defined as:

$$Cont(P_{lidar} \cap V_{proj}) < \rho_{\min} \tag{9}$$

where $V_{proj}$ represents the projected frustum region of the visual detection bounding box in 3D space; $P_{lidar}$ denotes the synchronized LiDAR point cloud set; and $\rho_{\min}$ is the minimum point cloud density threshold. If an object is detected by vision but the point cloud feedback in the corresponding physical space is sparse, the target is determined to be a visual false alarm lacking a physical entity, thus achieving essential alignment between semantics and spatial reality.

**(3) Temporal State Stability Discrimination**

False alarms caused by environmental noise (such as flying dust or instantaneous lighting fluctuations) exhibit significant randomness in the time dimension, whereas the motion trajectories of real obstacles follow physical continuity. The discrimination criterion is defined as:

$$\|z_k - \hat{z}_k\| > \varepsilon \text{ or } L < L_{\min} \tag{10}$$

In this process, Kalman filtering is employed for target state tracking [21]. $z_k$ denotes the actual observed position at frame $k$, while $\hat{z}_k$ represents the predicted position based on historical trajectories. $\varepsilon$ is the residual threshold used to identify targets with abnormal position jitter, and $L_{\min}$ is designed to filter out bursty, instantaneous interference. This step ensures the temporal robustness of the perception results.

**(4) Multi-Criterion Joint Decision Fusion**

A single level of verification is often insufficient to cover all complex industrial scenarios. To avoid the risk of unnecessary downtime caused by misjudgment, the system constructs a weighted consistency decision model that serves as a logical switch for the control layer:

$$O_{final} = (C_1 \wedge C_2) \vee (C_1 \wedge C_3) \vee (C_2 \wedge C_3) \tag{11}$$

where $O_{final}$ denotes the final decision operator output to the control layer. $C_1$, $C_2$, and $C_3$ represent the sub-decision operators from the aforementioned visual, spatial, and temporal dimensions, respectively. When the target is judged as "True" in a specific dimension, the operator takes a value of 1, otherwise 0. Only when the authenticity of the target is confirmed by at least two dimensions simultane-

 doi:10.20944/preprints202604.1455.v1



ously will $O_{final}$ output 1 and trigger the braking command. This non-linear fusion method effectively mitigates the risk of false alarms triggered by single-sensor failure through mutual constraints between modalities, ensuring high-reliability operation of AGVs in complex industrial environments.

### *2.4. Fault-Tolerant Control Strategy Based on Risk Perception*

In dynamic industrial environments, the detection results output by the sensing system are subject to certain uncertainties due to sudden changes in illumination, local occlusion, and random noise from sensors [22]. If the AGV relies solely on instantaneous detection results to perform braking actions, it is prone to decision oscillations and frequent false stops. To address this, this paper designs a fault-tolerant control strategy based on risk perception, which ensures the continuity of the system's operation under interference conditions through threshold redundancy, multi-source verification, and hierarchical execution logic.

**(1) Threshold Redundancy Mechanism**

Traditional single-threshold discrimination often leads to frequent starting and stopping of the actuator when the target confidence is near the critical value. This paper introduces a dual-threshold hysteresis discrimination mechanism to establish a decision buffer zone. The system categorizes risk levels based on the target's real-time confidence $P$, using a primary detection threshold $\tau_{high}$ and a fault-tolerant buffer threshold $\tau_{low}$:

$$Level = \begin{cases} High\_Risk, & P \geq \tau_{high} \\ Uncertain, & \tau_{low} \leq P < \tau_{high} \\ Low\_Risk, & P < \tau_{low} \end{cases} \tag{12}$$

This mechanism reserves the interval $[\tau_{low}, \tau_{high}]$ as a non-linear transition zone. Within this interval, the system does not perform drastic braking; instead, it maintains a low-speed state for continuous observation. This "soft switching" logic effectively filters out jump noise during edge detection by the algorithm.

**(2) Multi-Source Consistency Verification**

For targets initially screened as "Uncertain," visual semantics alone are insufficient to support safety decisions. Therefore, spatial occupancy features from LiDAR point clouds are introduced to eliminate visual false alarms through the complementarity of heterogeneous data:

$$C_{fusion} = IoU(V_{proj}, L_{bbox}) \tag{13}$$

where $V_{proj}$ represents the spatial projection area of the visual detection frame onto the physical plane via the camera's extrinsic matrix; $L_{bbox}$ denotes the physical bounding box of the obstacle extracted by the LiDAR point cloud clustering algorithm. When $C_{fusion} > \theta$ (a preset overlap threshold), the risk is locked as a "True Physical Obstacle." If vision shows a response but there is no LiDAR return, the system automatically determines it as a visual artifact (e.g., reflections or virtual images), thereby avoiding unnecessary braking actions.

**(3) Hierarchical Response and Fault-Tolerant Execution Logic**

This logic transforms the uncertainty output by the perception layer into a smooth velocity curve for the control layer. By replacing the traditional "binary" stop-start mode with a hierarchical strategy, operational efficiency is optimized. Based on the final verified risk factors, the execution velocity $v_{cmd}$ generated by the controller is as follows:

$$v_{cmd} = \begin{cases} 0, & Target = High\_Risk \\ \lambda \cdot v_{target}, & Target = Uncertain \\ v_{target}, & Target = Pseudo\_Obstacle \end{cases} \tag{14}$$

where $v_{target}$ represents the target cruising speed set by the mission; $\lambda$ denotes the speed attenuation coefficient (typically within the range $[0.3, 0.5]$). When a target is judged as "Uncertain," the AGV

 doi:10.20944/preprints202604.1455.v1



does not stop completely but proceeds at a low speed of $\lambda$ times the target velocity, while actively increasing the sensor sampling frequency for re-verification. This "degraded operation" logic, rather than "complete interruption," significantly enhances the vehicle's transport efficiency.

## 3. Experimental Verification and Result Analysis

### *3.1. Experimental Platform and Deployment Environment*

This study uses an industrial-grade stealth AGV as a verification platform, integrating a high-precision LiDAR, a depth camera, and an NVIDIA Jetson edge computing terminal. The system is based on Ubuntu and ROS architecture, and uses TensorRT to accelerate model inference, ensuring millisecond-level synchronization of perception and control commands. Detailed hardware, software, and training configurations are shown in Table 1.

**Table 1.** Experimental Environment and Parameter Settings.

| Category | Component | Core Parameters / Specifications |
|---|---|---|
| Hardware Platform | AGV Chassis | Industrial latent AGV (Dual-wheel differential, 500kg, 1.5m/s) |
| | LiDAR | Hokuyo UST-10LX (270° scanning, ±30mm accuracy) |
| | Depth Camera | Intel RealSense D435i |
| | Edge Computing | NVIDIA Jetson AGX Orin (2048 CUDA cores, 32GB RAM) |
| | Dispatch Server | Intel Xeon Gold (Multi-machine dispatching and data storage) |
| Software Environment | Operating System | Ubuntu 20.04 LTS |
| | Middleware | ROS Noetic / TensorRT |
| | Language | Python 3.8 / PyTorch 2.0 / CUDA 12.1 |
| Algorithm Training | Training Hardware | NVIDIA RTX A6000 (48GB VRAM) Workstation |
| | Input Size | 640 × 640 pixels |
| | Optimizer | SGD (Batch Size = 4, Epochs = 200) |
| | Hyperparameters | Momentum, weight decay, etc., follow YOLOv8 official config |

### *3.2. Dataset Construction and Preprocessing*

#### 3.2.1. Definition of Anomaly Categories in Industrial Scenarios

In complex industrial environments, unplanned AGV downtime primarily stems from physical space conflicts and the failure of the sensing system's feature representation. Through field research and operational condition analysis, this paper categorizes typical abnormal operating conditions into the following five types:

(i) **High-dynamic obstacles:** Labeled "high," these are objects located above the AGV's path and within the blind zone of the 2D LiDAR scanning plane (such as open fire doors, height-limiting barriers, and protruding pipes). These obstacles are highly likely to cause collisions in the cargo area.

(ii) **Ground dynamic interference:** Labeled "Ground," this includes randomly intruding workers, temporarily stacked pallets, and scattered objects. These targets exhibit strong randomness, placing stringent demands on the real-time response of the sensing system.

(iii) **Geometric ambiguity:** Labeled "fence," this refers to structures such as fences or perforated metal mesh, where the laser beam easily penetrates, creating a "sparse point cloud" phenomenon, preventing the formation of an effective closed physical contour.

(iv) **Mirror image interference:** Labeled "mirror," this covers visual virtual images generated by highly reflective mirrors. Such operating conditions are highly likely to induce SLAM localization drift and false obstacle recognition in the perception layer.

(v) **Severe light and shadow interference:** Aligned with the label "light," including backlight, strong direct light, and sudden changes in ambient lighting. This type of environmental noise can cause oversaturation or feature loss in the image sensor.

 doi:10.20944/preprints202604.1455.v1



#### 3.2.2. Data Collection, Annotation and Augmentation

For the long-tailed distribution features of abnormal samples in industrial sites [23], this paper constructs an integrated dataset construction scheme of "core scene extraction + multi-dimensional data enhancement." 76 real vehicle images containing the above five typical working conditions were selected as the original benchmark set, Figure 4. In order to alleviate the overfitting problem caused by the scarcity of samples, the sample size was expanded to 760 images by using operators such as Mosaic-9 stitching [24], random contrast adjustment and affine geometric transformation. All samples were labeled at the pixel level using LabelImg [25] and the semantic consistency of the labels was ensured by a multi-person cross-validation mechanism.


**Figure 4.** Dataset Example.

#### 3.2.3. Transfer Learning Strategy and Weight Initialization

Considering the limitations of the scale of the self-built industrial dataset, this paper introduces a transfer learning strategy [26] to optimize model training: the model first loads pre-trained weights based on the large-scale COCO dataset [27], and through transfer learning, the model learns in advance the ability to represent general low-level features such as edges and textures. Then, fine-tuning is performed on the self-built industrial scene dataset. This strategy significantly accelerates the convergence of the loss function and guides the network to accurately capture specific long-tail targets such as suspended cables and mirror shadows, effectively improving the generalization robustness of the model in complex dynamic environments.

### *3.3. Evaluation Indicators*

In this study, several conventional object detection evaluation metrics are employed to assess the detection and recognition of abnormal behaviors caused by AGV mapping failures in complex environments. Specifically, the metrics include Precision ($P$), Recall ($R$), Average Precision ($AP$), mean Average Precision ($mAP$), inference speed (Frames Per Second, FPS), number of parameters (Parameters), and Giga Floating-point Operations (GFLOPs). The formulas are defined as follows:

$$P = \frac{TP}{TP + FP} \tag{15}$$

$$R = \frac{TP}{TP + FN} \tag{16}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i = \frac{1}{N}\sum_{i=1}^{N}\int_0^1 P(R)dR \tag{17}$$

where $N$ denotes the total number of detection categories, and $AP_i$ represents the average precision for the $i$-th category of targets. $TP$, $FP$, and $FN$ correspond to true positives, false positives, and false negatives, respectively.

 doi:10.20944/preprints202604.1455.v1



### *3.4. Detection Model Training and Results*

#### 3.4.1. Training Process and Convergence Analysis

As shown in Figure 5, the model demonstrated excellent convergence speed within the first 100 epochs, with the localization loss ($\mathcal{L}_{box}$), classification loss ($\mathcal{L}_{cls}$), and distribution focal loss ($\mathcal{L}_{dfl}$) all exhibiting a step-wise downward trend. This indicates that the initial feature representation capabilities endowed by transfer learning enabled the model to rapidly localize anomalous targets during the early stages of training. As training progressed into the mid-to-late stages (after 100 epochs), the curves for the various loss functions smoothed out and gradually converged, signifying that the model had successfully completed the transfer mapping from generic features to features specific to the industrial scenario, and that the model weights had reached a steady state.

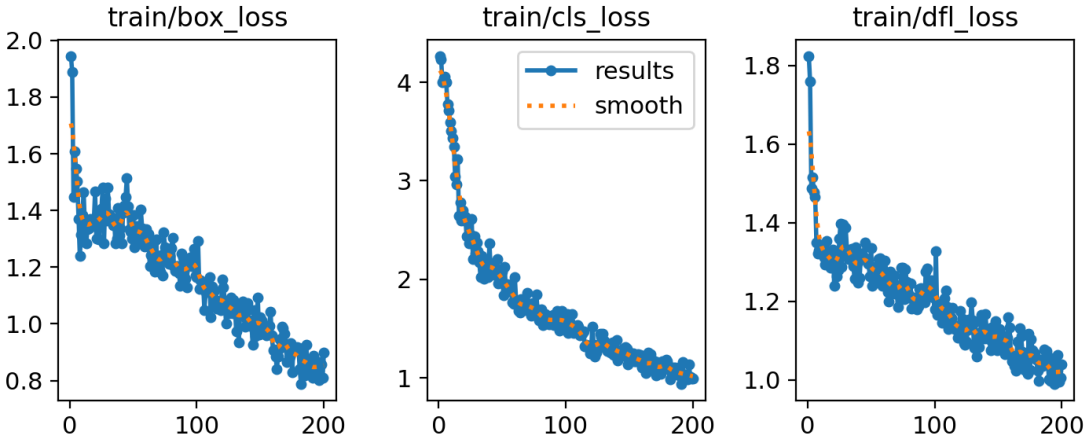



**Figure 5.** Training Loss Curve.

#### 3.4.2. Ablation Study Analysis

To validate the effectiveness of the AG-FPN and DHA modules in enhancing the overall performance of the detection model, this study designed four sets of comparative experiments. Experiment 1 served as the baseline using the standard YOLOv8 model, while Experiments 2 through 4 evaluated model performance after introducing a single improvement module—individually—and after simultaneously incorporating both modules. The quantitative metrics for each experimental group are presented in Table 2.

**Table 2.** Ablation Study Results.

| Exp | YOLOv8 | AG-FPN | DHA | P | R | mAP@0.5 | mAP@50-95 | Params/M | FLOPs/G | FPS |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ✓ | | | 0.920 | 0.802 | 0.912 | 0.690 | 3.01 | 8.1 | 330 |
| 2 | ✓ | ✓ | | 0.926 | 0.815 | 0.921 | 0.702 | 3.16 | 8.4 | 315 |
| 3 | ✓ | | ✓ | 0.931 | 0.820 | 0.928 | 0.715 | 3.46 | 9.0 | 295 |
| 4 | ✓ | ✓ | ✓ | **0.945** | **0.835** | **0.939** | **0.732** | 3.62 | 9.5 | 280 |

Ablation study results indicate that the standalone introduction of **AG-FPN** boosts mAP@0.5 by 0.9%, thereby enhancing multi-scale feature fusion capabilities; meanwhile, **DHA** contributes a 2.5% improvement to mAP@50-95, significantly optimizing bounding box localization accuracy. In Experiment 4, integrating both modules yielded optimal performance, with mAP@0.5 and mAP@50-95 increasing by **2.7%** and **4.2%**, respectively, compared to the baseline—demonstrating the excellent compatibility between feature enhancement and perception head optimization in addressing complex industrial anomalies.

 doi:10.20944/preprints202604.1455.v1



3.4.3. Comparative Experimental Analysis

To objectively evaluate the detection performance of the improved algorithm proposed in this paper—specifically regarding anomalous AGV mapping failures in complex industrial environments—several currently mainstream lightweight object detection models were selected for comparative testing. All models were trained and validated using the same small-sample anomaly dataset; the experimental results are presented in Table 3.

**Table 3.** Performance comparison between the improved model and other YOLO-series models.

| Model | P | R | mAP@0.5 | mAP@50-95 | Params/M | FLOPs/G | FPS |
|---|---|---|---|---|---|---|---|
| YOLOv5n | 0.882 | 0.744 | 0.854 | 0.582 | 1.90 | 4.5 | 450 |
| YOLOv8n | 0.920 | 0.802 | 0.912 | 0.690 | 3.01 | 8.1 | 330 |
| YOLOv9n | 0.925 | 0.810 | 0.915 | 0.695 | 2.15 | 7.8 | 395 |
| YOLOv10n | 0.932 | 0.812 | 0.918 | 0.698 | 2.30 | 6.4 | 385 |
| YOLOv11n | 0.935 | 0.818 | 0.925 | 0.702 | 2.60 | 6.5 | 360 |
| **Ours** | **0.945** | **0.835** | **0.939** | **0.732** | **3.62** | **9.5** | **280** |

Experimental results demonstrate that the model proposed in this paper achieves optimal performance metrics: an mAP@0.5 of 93.9% and an mAP@50-95 of 73.2%, representing improvements of 2.7% and 4.2%, respectively, over the baseline YOLOv8n. This algorithm yields significant gains in accuracy at the cost of only a marginal increase in computational overhead. Furthermore, while maintaining a high frame rate of 280 FPS, it successfully strikes a profound balance between detection accuracy and inference speed.

3.4.4. Visualization of Detection Results

To intuitively validate the improved algorithm's capability to perceive complex industrial anomalies, this section selects typical scenarios for a visual comparison (as shown in Figure 6). The analysis reveals that the improved model demonstrates significant advantages across the following dimensions: In the "Fence" scenario, the bounding box generated by the baseline YOLOv8 model exhibits an offset, and its confidence score is merely 0.66. In contrast, the improved model proposed in this paper achieves a tighter bounding box regression, with the confidence score rising substantially to 0.97. Regarding interference from "Light" sources and "Mirror" reflections, the baseline model is highly prone to false positives (e.g., misidentifying reflective areas as "Light" sources, with confidence scores ranging only from 0.26 to 0.32). The improved model is capable of accurately distinguishing between physical obstacles and optical virtual images; false detection boxes are completely eliminated, thereby significantly enhancing perceptual robustness in extreme lighting and shadow environments. In multi-target scenarios involving both "Ground" surfaces and distant windows ("Mirrors"), the improved model demonstrates superior stability in perceptual recognition, achieving confidence scores of 0.96 and 0.92 for the two respective target categories. This visual comparison confirms that the proposed model effectively resolves the issues of low confidence and high false detection rates inherent in the baseline algorithm, generating highly clean and reliable detection signals that provide a critical safeguard for AGV obstacle avoidance in complex environments.

 doi:10.20944/preprints202604.1455.v1



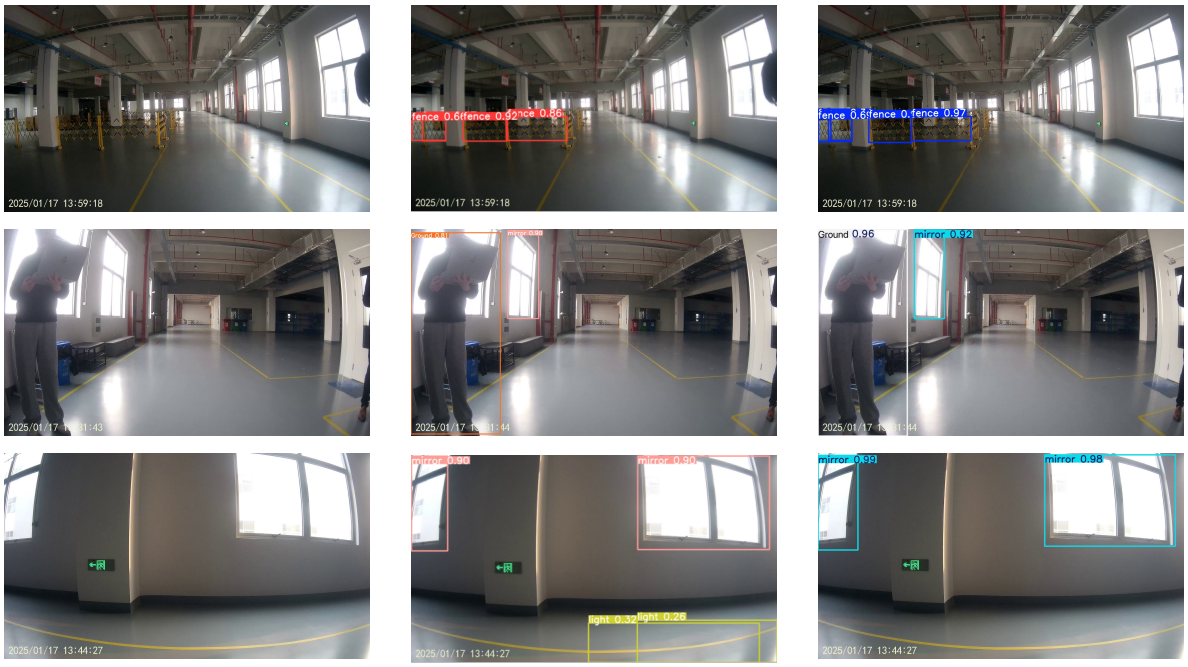



**Figure 6.** Visualization of experimental results in different scenarios.

### *3.5. Pseudo-Obstacle Recognition Experiment*

#### 3.5.1. Discriminant Criteria and Logical Models

To quantitatively verify the pseudo-obstacle recognition mechanism proposed in this paper, a joint decision-making model based on spatial topology and temporal features is constructed, utilizing the semantic probabilities output by YOLOv8. The discrimination logic is formally defined by the piecewise function $C(O)$ as follows:

$$C(O) = \begin{cases} Real, & (P_{obj} < \tau_p) \wedge (H_{min} < H_{safe}) \wedge (S \geq \tau_s) \\ Pseudo, & (H_{min} \geq H_{safe}) \vee (Type \in \{Mirror, Fence\}) \\ Uncertain, & (P_{obj} \geq \tau_p) \wedge (S < \tau_s) \end{cases} \tag{18}$$

Where the parameters are defined as follows:

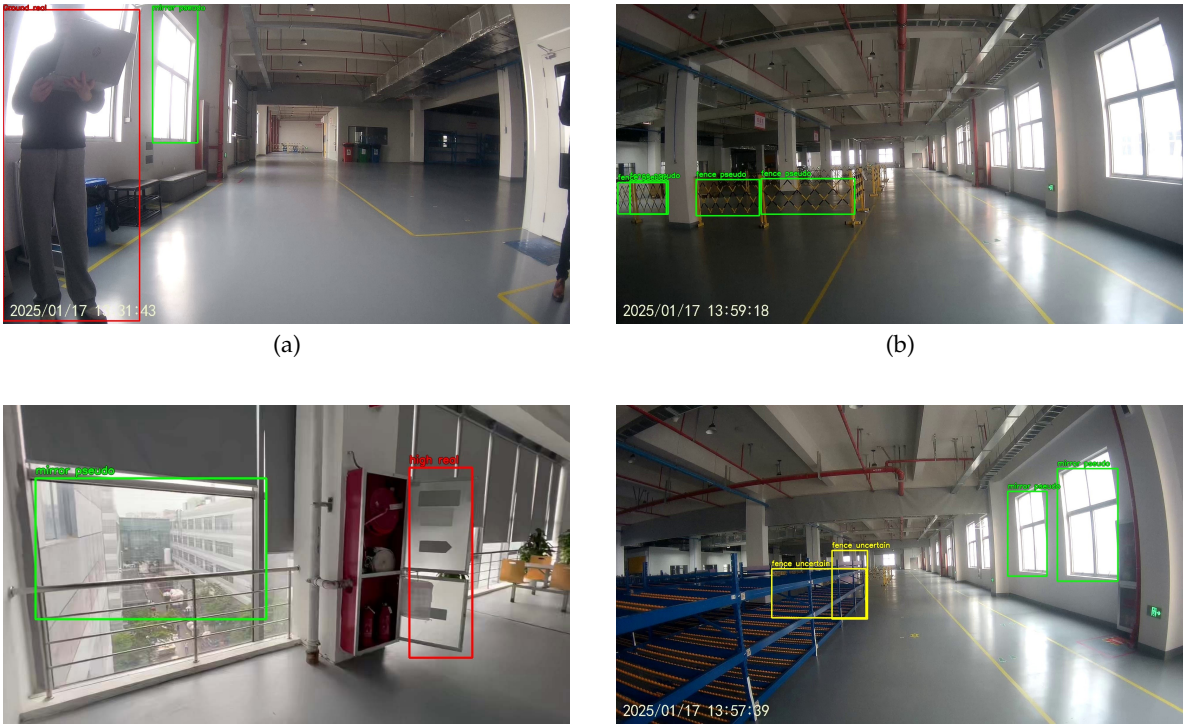
- $H_{min}$ represents the real-time vertical height from the bottom of the target to the ground.
- $H_{safe}$ denotes the safe passage height threshold for the AGV.
- $S$ represents the temporal stability factor of the target.
- $\tau_s$ is the stability threshold, which is set to 0.8 in this study. This design is motivated by the use of time-window filtering [28] to forcibly filter out non-persistent false alarms caused by ambient light flicker or instantaneous sensor noise.
- $P_{obj}$ and $\tau_p$ represent the classification confidence of the detection network and its corresponding threshold, respectively.
- **Real** indicates that the target possesses both physical spatial occupancy and temporal stability, triggering the highest priority emergency braking.
- **Pseudo** covers fences outside the path (no path occupancy) and optical-induced phantom images (no physical entity); the system executes "imperceptible passage."
- **Uncertain** refers to transient targets whose confidence reaches the standard but whose temporal performance fluctuates. These are marked as uncertain states, triggering the fault-tolerant deceleration mechanism described in Section 2.4.

 doi:10.20944/preprints202604.1455.v1



3.5.2. Visual Analysis of Experimental Results

As shown in Figure 7, in Scenario (a), the detection network simultaneously extracts features of pedestrians and windows. The discrimination logic correctly labels the Ground truth as **Real** (confidence 0.96) while identifying the window under strong light interference as **Pseudo**.




(a) (b)

(c) (d)

**Figure 7.** Visualization of Experimental Results for Pseudo-Obstacle Detection Mechanisms Across Different Scenarios.

In Scenario (b), regarding the multi-segment fence structure, the system recognizes that it does not intrude into the core driving path and labels it uniformly as **Pseudo**. This avoids unnecessary AGV downtime when traversing narrow aisles.

In Scenario (c), for the suspended fire door, the system measures its minimum height $H_{min}$ to be lower than $H_{safe}$, resulting in a **Real** classification. Conversely, the glass curtain wall reflection area is determined to be **Pseudo**.

In Scenario (d), for complex edges severely affected by light and shadow interference, the system cautiously marks them as **Uncertain**, providing a discriminative basis for subsequent fault-tolerant control.

*3.6. Fault-Tolerant Control Strategy Verification Experiment*

3.6.1. Tiered Response Strategies and Action Mapping

To verify the robustness of the system in uncertain environments, this paper integrates the pseudo-obstacle identification results with a fault-tolerant control strategy to construct a graded response decision model [29]. The model transforms perception states into specific motion commands through a non-linear mapping function $A(O)$:

$$v_{cmd} = A(O) = \begin{cases} 0, & O = Real \\ \lambda \cdot v_{target}, & O = Uncertain \\ v_{target}, & O = Pseudo \end{cases} \tag{19}$$

The specific decision-making logic is defined as follows:

 doi:10.20944/preprints202604.1455.v1



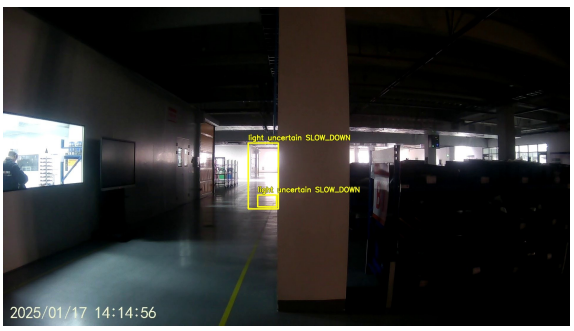
- **Emergency Braking (STOP, $O = Real$):** When a target is verified as a real obstacle with physical occupancy risk (e.g., ground cargo, dynamic pedestrians), the system outputs a velocity $v_{cmd} = 0$, triggering the highest priority braking to ensure absolute safety.
- **Deceleration Confirmation (SLOW_DOWN, $O = Uncertain$):** For targets with fluctuating confidence or temporal instability (e.g., instantaneous strong light, complex textures), the system executes a degraded operation strategy. Through the velocity attenuation factor $\lambda$, the AGV enters a low-speed perception mode, aiming to obtain more stable temporal features by increasing observation duration and avoiding false stops caused by blind decision-making.
- **Normal Passage (GO, $O = Pseudo$):** For pseudo-obstacles determined to have no spatial occupancy risk (e.g., high-level pipelines, mirror phantoms, fences outside the path), the system maintains the preset cruise speed $v_{target}$, achieving "imperceptible filtering" of interference items and ensuring the continuity of the logistics rhythm.

#### 3.6.2. Experimental Visualization Analysis

As shown in Figure 8, real-vehicle experiments conducted across four typical industrial scenarios validate the significant advantages of the hierarchical decision-making mechanism in enhancing operational efficiency: In scenario (e), repetitive fence structures are identified by the mechanism as **Pseudo**; the control strategy outputs a **GO** command, allowing the AGV to pass smoothly through the narrow passage without interference from structural textures. In scenario (f), the system simultaneously identifies personnel ahead (**Real**) and large-scale specular reflections on the side (**Pseudo**); the strategy accurately outputs a **STOP** command to avoid personnel while successfully filtering out "mirror phantoms" that commonly trigger false alarms in traditional LiDAR, ensuring the purity of decision commands. In scenario (g), affected by complex ambient lighting, the system produces unstable detection results, and the target is marked as **Uncertain**; the strategy outputs a **SLOW_DOWN** command, causing the AGV to decelerate for a more stable observation perspective. In scenario (h), within a corridor environment, the system captures the edge of a cargo box below the safety height threshold via the $H_{\min}$ perception operator, identifying it as **Real**. After verification through multi-source features, a **STOP** command is issued.


(e)


(f)




(g)


(h)

**Figure 8.** Experimental results of fault-tolerant control in industrial scenarios

 doi:10.20944/preprints202604.1455.v1



### *3.7. System-Level Performance Evaluation and Verification*

#### 3.7.1. Experimental Environment and Scheme Design

To verify the robustness and engineering feasibility of the improved algorithm in complex dynamic industrial scenarios, this study conducted a week-long real-vehicle deployment test in the PCB assembly workshop of a large-scale electronics manufacturing plant.

The experimental platform utilizes an industrial-grade differential stealth AGV, with an NVIDIA Jetson AGX Orin (32GB) serving as the onboard computing unit. Sensors include an Intel RealSense D435i depth camera, deployed at a height of 450 mm with a downward tilt angle of 5°. The total test path length is 220 m, covering three typical extreme working conditions: Mirroring Area (high reflectivity interference), Fencing Area (perceptual redundancy caused by metal grid fences), and areas with abrupt light and shadow transitions.

The experiment was conducted with two comparative groups to evaluate the performance of the proposed system:

- **Group A (Baseline Group):** This group utilized the baseline YOLOv8n detector combined with conventional obstacle avoidance logic. In this configuration, the system triggers an immediate emergency brake (*Emergency Braking*) as soon as any target is detected within the safety zone.
- **Group B (Improved Group):** It incorporates the improved algorithm presented in this paper and integrates pseudo-obstacle detection logic and a graded response fault tolerance strategy.

Each group completed 300 cycles of testing over a total distance of 132 km. This large-scale sampling was designed to eliminate contingency and rigorously verify the long-term stability of the proposed system in dynamic industrial environments.

#### 3.7.2. Comprehensive Performance Index Analysis

A total of 132 km of operational trajectories and 1,452 obstacle avoidance events were recorded via the onboard Black-box Logger. False Stop Rate (FSR), Continuous Range (CR), and Average Process Time (APT) were introduced as core evaluation metrics. The results are shown in Table 4.

**Table 4.** Comparison of experimental results between Group A and Group B.

| Performance Dimension | Evaluation Metrics | Group A | Group B | Improvement |
|---|---|---|---|---|
| Perception Accuracy | Fence Recognition Accuracy | 92.4% | 96.7% | ↑ 4.3% |
| | Mirror Recognition Accuracy | 76.5% | 89.6% | ↑ 13.1% |
| Operational Continuity | FSR | 23.7% | 2.1% | ↓ 91.1% |
| | CR | 45.2 m | 406.8 m | ↑ 8.99 times |
| Operational Efficiency | APT | 148.5 s | 112.4 s | ↓ 24.3% |
| | Daily Cumulative Delay | 13.6 h | 8.6 h | ↓ 5.0 h |
| | Overall System Pass Efficiency | / | / | ↑ 36.8% |

The field deployment results provide strong evidence that, owing to the precise extraction of complex light, shadow, and subtle features by the **DHA** and **AG-FPN** modules, the system's recognition accuracy for reflective phantoms improved by 13.1%. This improvement facilitated a significant reduction in the **False Stop Rate (FSR)** from 23.7% to 2.1%.

By introducing pseudo-obstacle identification and adaptive hierarchical response strategies, the **Continuous Range (CR)** increased nearly 9-fold, effectively releasing 5.0 hours of labor time per unit daily. Field deployment confirms that this algorithm significantly enhances the perception robustness

 doi:10.20944/preprints202604.1455.v1



and operational continuity of AGVs with zero additional hardware cost. Consequently, it effectively addresses the efficiency bottleneck caused by overly sensitive obstacle avoidance decisions in industrial scenarios, demonstrating high value for large-scale engineering promotion.

## 4. Summary and Outlook

To address the issues of perception instability and false stops for AGVs in complex industrial environments, this study developed a perception-decision closed-loop system and achieved several key results. Specifically, an industrial anomaly dataset covering five typical conditions, such as abrupt illumination changes and specular reflections, was constructed to support model training for long-tail events. Furthermore, the designed AG-FPN and DHA modules enabled the improved model to achieve a mAP@0.5 of 93.9% while maintaining an inference speed of 280 FPS, which significantly enhances the discrimination of subtle and optical pseudo-obstacles while remaining lightweight. The proposed pseudo-obstacle identification mechanism and hierarchical fault-tolerant strategy further reduced the False Stop Rate (FSR) from 23.7% to 2.1% and increased the Continuous Range (CR) by 8.99 times. Field deployment successfully demonstrated that the system can reduce unnecessary delays by 5.0 hours per day, highlighting its substantial engineering value.

Despite the breakthroughs achieved in single-vehicle perception, further research is required regarding multi-vehicle collaboration and the impact of communication latency. Future work will explore quantization techniques to adapt the algorithm for lower-power edge computing platforms and investigate online incremental learning methods for perception models in extreme environments, such as heavy dust, to reduce maintenance costs caused by environmental fluctuations. Additionally, Transformer or LSTM architectures will be introduced for temporal modeling to strengthen predictive capabilities for dynamic anomalous behaviors. Ultimately, the research scope will be extended to multi-AGV collaborative scenarios to construct an end-to-end intelligent closed-loop system evolving from individual perception to swarm decision-making.

**Author Contributions:** Conceptualization, Y.Z. and Y.W.; methodology, Y.Z.; software, Y.Z., S.P. and Z.Z.; validation, Y.Z., S.P., Z.Z., N.Z., B.Z. and F.S.; formal analysis, Y.Z. and S.P.; investigation, Y.Z., N.Z., B.Z. and F.S.; resources, Y.W. and F.S.; data curation, Y.Z., S.P. and Z.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.W.; visualization, Y.Z.; supervision, Y.W.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 51575245; the Open Research Fund of the State Key Laboratory of Marine Critical Materials, grant number 2025K15; and the Zhenjiang Key Research and Development Program, grant number GY2023013. The APC was funded by the authors.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author.

**Conflicts of Interest:** The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Fragapane, G.; De Koster, R.; Sgarbossa, F.; Strandhagen, J.O. Planning and control of autonomous mobile robots for intralogistics: Literature review and research agenda. *Eur. J. Oper. Res.* **2021**, *294*, 405–426.
2. Vlachos, I.; Pascazzi, R.M.; Ntotis, M.; Mykoniatis, K. Smart and flexible manufacturing systems using Autonomous Guided Vehicles (AGVs) and the Internet of Things (IoT). *Int. J. Prod. Res.* **2024**, *62*, 5574–5595.
3. Damodaran, D.; Mozaffari, S.; Alirezaee, S.; Paiva, A.L.S. Experimental analysis of the behavior of mirror-like objects in LiDAR-based robot navigation. *Appl. Sci.* **2023**, *13*, 2908.

 doi:10.20944/preprints202604.1455.v1




4. Grewal, R.; Tonella, P.; Stocco, A. Predicting safety misbehaviours in autonomous driving systems using uncertainty quantification. In *Proceedings of the 2024 IEEE Conference on Software Testing, Verification and Validation (ICST)*, Toronto, ON, Canada, 27–31 May 2024; pp. 70–81.
5. Yuan, C.; Liu, J.; Wang, Y. Research on indoor positioning and navigation method of AGV based on multi-sensor fusion. *Highlights Sci. Eng. Technol.* **2022**, *7*, 206–213.
6. Song, D.; Tian, G.M.; Liu, J. Real-time localization measure and perception detection using multi-sensor fusion for Automated Guided Vehicles. In *Proceedings of the 2021 40th Chinese Control Conference (CCC)*, Shanghai, China, 26–28 July 2021; pp. 3219–3224.
7. Zhou, S.; Cheng, G.; Meng, Q.; Chen, G. Development of multi-sensor information fusion and AGV navigation system. In *Proceedings of the 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 12–14 June 2020; Volume 1, pp. 2043–2046.
8. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237.
9. Yang, D.; Su, C.; Wu, H.; Zheng, Y. Shelter identification for shelter-transporting AGV based on improved target detection model YOLOv5. *IEEE Access* **2022**, *10*, 119132–119139.
10. Qian, L.; Zheng, Y.; Cao, J.; Li, Z. Lightweight ship target detection algorithm based on improved YOLOv5s. *J. Real-Time Image Process.* **2024**, *21*, 3.
11. Vassilev, A.; Hasan, M.; Griffor, E.; Hany, J. On the Assessment of Sensitivity of Autonomous Vehicle Perception. *arXiv* **2026**, arXiv:2602.00314.
12. Li, W.; Qiu, J. An improved autonomous emergency braking algorithm for AGVs: Enhancing operational smoothness through multi-stage deceleration. *Sensors* **2025**, *25*, 2041.
13. Han, J.; Zhang, J.; Lv, C.; Wang, H. Robust fault tolerant path tracking control for intelligent vehicle under steering system faults. *IEEE Trans. Intell. Veh.* **2024**, doi:10.1109/TIV.2024.xxxxxxx.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589.
16. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
17. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
18. Yang, L.; Zhang, R.Y.; Li, L.; Xie, X. Simam: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, Online, 18–24 July 2021; pp. 11863–11874.
19. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8–14 September 2018; pp. 3–19.
20. Cui, Y.; Chen, R.; Chu, W.; Chen, L.; Tian, D.; Li, Y.; Cao, D. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 722–739.
21. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 17–20 September 2017; pp. 3645–3649.
22. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5574–5584.
23. Van Horn, G.; Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv* **2017**, arXiv:1709.01450.
24. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
25. Tzutalin. LabelImg. Available online: https://github.com/tzutalin/labelImg (accessed on 20 May 2024).
26. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359.
27. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.

 doi:10.20944/preprints202604.1455.v1



28. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360.
29. Yan, R.; Dunnett, S.J.; Jackson, L.M. Model-based research for aiding decision-making during the design and operation of multi-load automated guided vehicle systems. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108264.