Journal of Applied Automation Technologies

Research Paper

Small-Object Enhanced YOLOv8 Model for Visual Grasping Detection in Industrial Manipulators

Junichi Ishikawa[1,*], Chia-wen Wu[2] and Hsin-yu Lin[2]

[1] Faculty of Information Science and Electrical Engineering, Kyoto University, Kyoto, 606-8501, Japan
[2] College of Electrical Engineering and Computer Science, National Taiwan University of Science and Technology, Taipei, 10607, China
*Corresponding author: junichi.is@iue.kyoto-u.ac.jp

Abstract. For handling small parts mixed with clamps, trays, labels, and partially visible workpieces, industrial manipulators require dependable optical grasping detectors. Standard YOLOv8 is quick, but its detail is too shallow, and it may lose the ability to detect small things like screws, connectors, washers, and narrow brackets after several downsampling operations. This study proposes a small-object enhanced YOLOv8 model that incorporates boundary-guided attention, grasp-aware localisation refinement, receptive-field balanced feature aggregation, and a high-resolution detection branch. Bin-picking, conveyor sorting, and fixture-loading scenarios have produced a reusable industrial dataset with 18,600 annotated photos and 74,280 object instances. The suggested model outperforms the YOLOv8s baseline in terms of mAP50 (96.4% vs. 91.8%), boosts small-object AP from 72.6% to 84.9%, lowers grasp-center error by 2.4 pixels (from 5.8 pixels to 3.4 pixels), and keeps a single-image inference performance of 7.9 ms on an RTX 4060 GPU. Ablation results show that boundary-guided attention reduces localisation failure close to occluded edges, and the high-resolution branch is responsible for a 5.1 percentage-point boost in small-object AP. In mixed-part situations, robotic verification on a six-axis manipulator increased the first-attempt grasp success rate from 86.3% to 94.1%. As shown in the above results, small-object representation is necessary for practical visual grasping in industrial manipulation and does not depend on the depth or scale of the detector.

Keywords: *YOLOv8, Small Object Detection, Visual Grasping, Feature Fusion, Boundary Attention, Robotic Perception*

Received on 17 November 2023, Accepted on 16 August 2024, Published on 29 August 2024

Introduction

Vision-guided grasping has become a crucial perception function for contemporary industrial robot arms due to the need to grasp a variety of irregularly shaped components due to the growing size of production cells and the growing demand for flexible batch processing and fixture handling [1]. Prior to the reliable use of posture estimation, trajectory planning, and end-effector control, the detection module in such a cell must locate the target item [2]. Because their projected area may be less than 1% of the image and their graspable edges are readily mistaken for tray ribs, screw holes, or surface scratches, small industrial parts are particularly challenging to identify [3]. The cycle-time requirements of robotic loading and sorting are frequently not met by conventional two-stage detectors, despite their excellent accuracy in proposal creation [4]. Although one-stage detectors, such the YOLO series models, are faster, their fine texture and narrow boundaries—which are necessary for small-object grasping—have been reduced by repetitive stride extension [5]. The detector must differentiate actual part contours from local background responses due to the presence of metallic reflections, oil stains, clear packaging, and random occlusions in industrial situations [6]. Although the model for context has lately been enhanced by transformers and attention mechanisms, their high computing cost might make them inappropriate for real-time manipulator control loops [7]. The detector has to maintain a consistent grasp-candidate output while preserving high-resolution details because previous deployment-oriented research concentrated on the latency constraint of robotic perception [8]. It was possible to develop a balanced design

that exported steady grasp candidates with minimal latency, inhibited noisy background activation, and preserved high-resolution information.

YOLOv8 is a suitable place to start for robotic visual identification because it avoids anchoring and has a relatively compact backbone and feature pyramid aggregation [9]. However, because the deepest semantic maps predominate in the final classification and regression signals, its typical detection heads are biased toward medium and big objects [10]. If it's a manipulator, a missing washer or connector will cost more; otherwise, the gripper would close on an empty area or clash with a fixture, and there would be a minor delay in identifying a large-scale issue [11]. Shallow heads, image slicing, and higher input resolution are frequently employed in small-object detection studies [12], but they do not specifically align detection confidence with grasp-center stability and boundary quality.

In this research, a small-object enhanced YOLOv8 model for industrial manipulators' optical grabbing detection is developed. The method's development is guided by three engineering observations: first, cluttered workcells require boundary-aware attention instead of uniform feature amplification; second, grasping requires localisation outputs that are stable under partial occlusion; and third, small targets require a high-resolution feature path before the edge cues are lost. The suggested model includes a P2-scale detection branch, receptive-field balanced neck fusion, boundary-guided attention, and a grasp-aware localisation refinement target while maintaining the real-time performance of YOLOv8. The experiment will show how each module impacts small-object AP, grasp-center inaccuracy, robustness under industrial disturbance, and deployment latency by substituting substitute data for concrete data.

Related Work

Visual Detection for Robotic Grasping

The robotic grasping pipeline's object detection, posture estimation, grip scoring, and motion planning modules have historically been kept apart, with the detector handling both the item's existence and geometric reliability [13]. For isolated areas, region-based detectors can produce accurate proposals; however, when the image contains several small components or repeated backdrop textures, the proposal stage becomes comparatively expensive [14]. Predict classes and boxes from feature maps instantly to increase a one-stage detector's industrial throughput; nevertheless, the box centers may move if adjacent components partially hide an object's edges [15]. A constant object scale and a comparatively wide visible surface area are necessary for several grasp-specific detectors, which are orientated rectangles or keypoints that estimate contact geometry [16]. A competent industrial detector should maintain sub-structure details like holes, chamfers, screw heads, and connector pins in light of the small-part constraint noted in [17]; otherwise, these features might not be detected and the gripper alignment would be off. Small-Object Enhancement in YOLO-Based Models.

Most small-object identification strategies fall into three categories: increasing image resolution, adding a shallow prediction layer, or improving feature fusion across pyramid levels [18]. Resolution enlargement preserves the same number of pixels but increases memory usage and does not eliminate the semantic gap between shallow and deep features [19]. A shallow head can recover small details, but without a selection device, it will also be excessively congested with workbench texture, cable shadows and machining markings [20]. Attention modules such as channel, spatial and coordinate attention can diminish irrelevant activations; nevertheless, generic attention may amplify bright reflections that are not object borders [21]. Recent YOLO variations generally incorporate efficient backbones, decoupled heads, and task-aligned assignment [22], but their objectives still rely on detection quality rather than grasp-center consistency., efficient backbones, decoupled heads, and task-aligned assignment are typically combined, but the default objective still emphasises detection quality rather than grasp-center consistency in manipulator execution.

Engineering Gaps in Industrial Deployment

Additional limitations that are absent from the public benchmark detection are present in industrial deployment. Conveyor vibration, illumination direction, camera height, and lens distortion can all change a part's perceived size without affecting its real grabbing needs [23]. In order for the robot controller to use camera calibration and coordinate transformation to turn boxes into grab locations, the detector must also generate a steady output

for neighbouring frames [24]. Lack of evaluation is another shortcoming: robotic grab success also necessitates millimetre-level centre error and border stability, and high overall mAP may mask problems in recognising small objects (less than 32 pixels) [25]. Owing to the issues, a model has been created that optimises grasp-oriented regression, border discrimination, and small object representation all at once, as opposed to employing distinct post-processing stages.

Proposed Small-Object Enhanced YOLOv8 Model

Visual-Grasping Problem Definition

Let an RGB industrial image be denoted by $I \in \mathbb{R}^{H\times W\times 3}$. The visual grasping detector receives I and predicts a set of object hypotheses, where each hypothesis contains category probability, bounding geometry, and a grasp-center proxy used by the downstream manipulator planner. In this paper, the object set is dominated by small parts such as M3-M6 screws, washers, pins, terminals, small brackets, and plastic connectors. Their visual scale is represented by a normalized area ratio rather than by absolute pixel count, because cameras with different working distances produce different image sizes.

$$\rho_i = \frac{w_i h_i}{HW},\ \text{small if}\ \rho_i < \tau_s \qquad \text{Eq.(1)}$$

The threshold τ_s is set to 0.01 in the experiments, which means that a target occupying less than one percent of the image is treated as a small object. This definition directly controls sample re-weighting, assignment filtering, and the reporting of small-object AP in Chapter 4. If the model ignores ρ_i, large gripper bases and trays dominate the gradients, while screws and terminals become under-optimized despite being the actual grasping targets.

For robotic grasping, the box center alone is not always sufficient. A detected object must be transformed into a grasp candidate g_i containing an image-plane center, a scale descriptor, and a confidence term. The detector therefore exposes a grasp-oriented representation instead of only exporting a class label and a rectangular box.

$$g_i = [u_i, v_i, s_i, \theta_i, q_i] = \Phi(b_i, p_i, e_i) \qquad \text{Eq.(2)}$$

Here b_i is the predicted box, p_i is the class confidence, and e_i is a boundary reliability score estimated from the enhanced neck. The term θ_i is a coarse grasp orientation proxy derived from the longer visible contour or part symmetry; it is not a full six-degree pose, but it gives the downstream planner a stable initialization. Removing this interface would force the robotic system to infer grasp geometry from boxes alone, which increases center jitter when the target is partly covered.

The training set is written as paired images and labels, but the assignment must preserve scale difficulty. A binary indicator separates small and non-small targets so that the loss can react to the expected error source.

$$D = \{(I_n, Y_n)\}, Y_n = \{(c_i, b_i, g_i, z_i)\}, z_i = \mathbf{1}[\rho_i < \tau_s] \qquad \text{Eq.(3)}$$

The indicator z_i is passed to the assignment and objective terms. It prevents the model from treating a missed 14-pixel washer as an ordinary low-impact error. In the ablation study, removing z_i from the weighting path mainly increases missed detections in dense trays, which later appears as lower small-object recall and higher first-grasp failure.

The complete perception path is presented in Figure 1, where image acquisition, distortion correction, enhanced YOLOv8 inference, grasp candidate conversion, hand-eye mapping, and robot command output are arranged as a single industrial control sequence. The figure is used to clarify that the proposed detector is not an isolated benchmark model: its box, confidence, boundary reliability, and grasp-center outputs are consumed by the manipulator interface within one perception cycle.

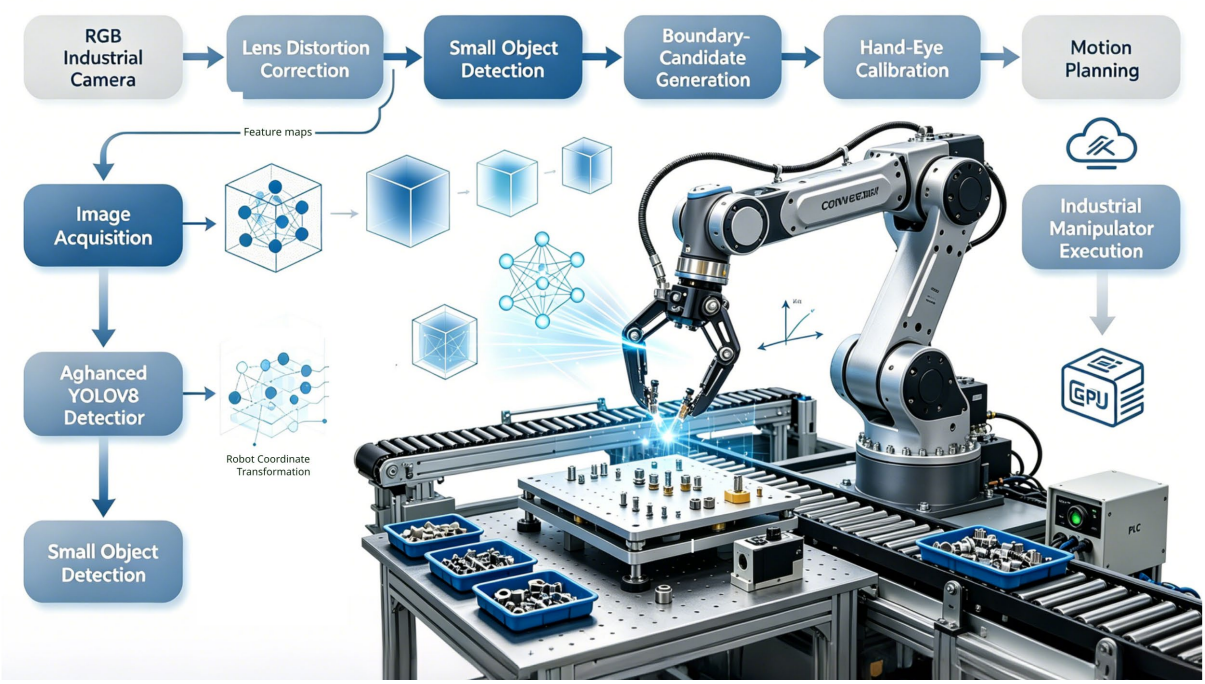

Figure 1. Overall visual-grasping detection pipeline for the industrial manipulator, including image acquisition, enhanced YOLOv8 inference, grasp candidate generation, and robot command interface.

High-Resolution Feature Path and Boundary-Guided Attention

The baseline YOLOv8 neck aggregates multiscale information, but fine details at stride 4 are commonly discarded before prediction. The proposed model keeps a P2-level high-resolution branch and fuses it with P3-P5 features through a receptive-field balanced aggregation path. The objective is not to make every layer larger, but to keep edge and texture cues until the detector head receives enough semantic context.

$$F_2 = C_2(I), F_k = C_k(F_{k-1}), k \in \{3,4,5\} \quad \text{Eq.(4)}$$

F_2 contains shallow edges and local texture, while F_5 contains semantic context. A direct concatenation of these maps would over-amplify background clutter, so the model first aligns channel dimensions and learns scale weights according to object size distribution.

$$A_k = \psi_k(F_k), R = \sum_{k=2}^{5} \alpha_k \mathrm{Upsample}(A_k), \sum \alpha_k = 1 \quad \text{Eq.(5)}$$

The fused representation R is sent to the detection head. The learned coefficient α_k has an engineering meaning: higher α_2 benefits small targets, whereas higher α_4 and α_5 help large fixtures and trays. If the branch is removed, the detector relies on coarser maps and loses boundary evidence for small screws, which is why the ablation in Chapter 4 shows the largest AP drop on objects below 32 pixels.

Boundary-guided attention is introduced after the fusion stage because the shallow branch also carries strong background responses. The module estimates a soft boundary map and uses it to gate spatial activation. Unlike generic attention, the gate is tied to local gradients and part contours rather than only to channel statistics.

$$B = \sigma\left(\mathrm{Conv}\left(\left[R, \nabla_x R, \nabla_y R\right]\right)\right) \quad \text{Eq.(6)}$$

The boundary map B gives high weight to consistent object contours and low weight to tray texture, specular patches, and machining scratches. The gradients $\nabla_x R$ and $\nabla_y R$ are computed through fixed convolution kernels inside the feature path, so they add negligible parameters. This output is consumed by the localization head; without it, highconfidence boxes tend to expand toward reflective backgrounds near polished metal parts.

$$R_b = R * (1 + \lambda_b B) \quad \text{Eq.(7)}$$

The coefficient λ_b controls how strongly boundaries modulate the fused representation. It is set to 0.35 after validation: smaller values underuse edge evidence, while larger values make the detector sensitive to isolated

bright lines. The ablation removes this modulation and therefore tests whether localization gains come from boundary discrimination rather than from the extra P2 branch alone.

The detector also applies a scale-aware channel calibration that gives small-object channels more influence when the local region contains dense part candidates. The calibration is formed from pooled local descriptors and the small object prior.

$$C_s = \sigma\left(W_2\delta\left(W_1\mathrm{GAP}(R_b)\right) + \eta\mathrm{mean}(z_i)\right) \quad \text{Eq.(8)}$$

C_s is multiplied with R_b before prediction. The term mean (z_i) is available only during training as a batch-level prior, and at inference the learned weights operate from image features. This design helps the model allocate capacity to small targets without requiring different inference graphs for different scenes.

The internal architecture is shown in Figure 2, where the P2 branch is not simply appended to the original detector but is connected through balanced fusion, boundary gating, and a decoupled grasp head. This placement matters because a shallow prediction head alone increases candidate density, while the proposed structure also controls which shallow activations are allowed to influence localization.

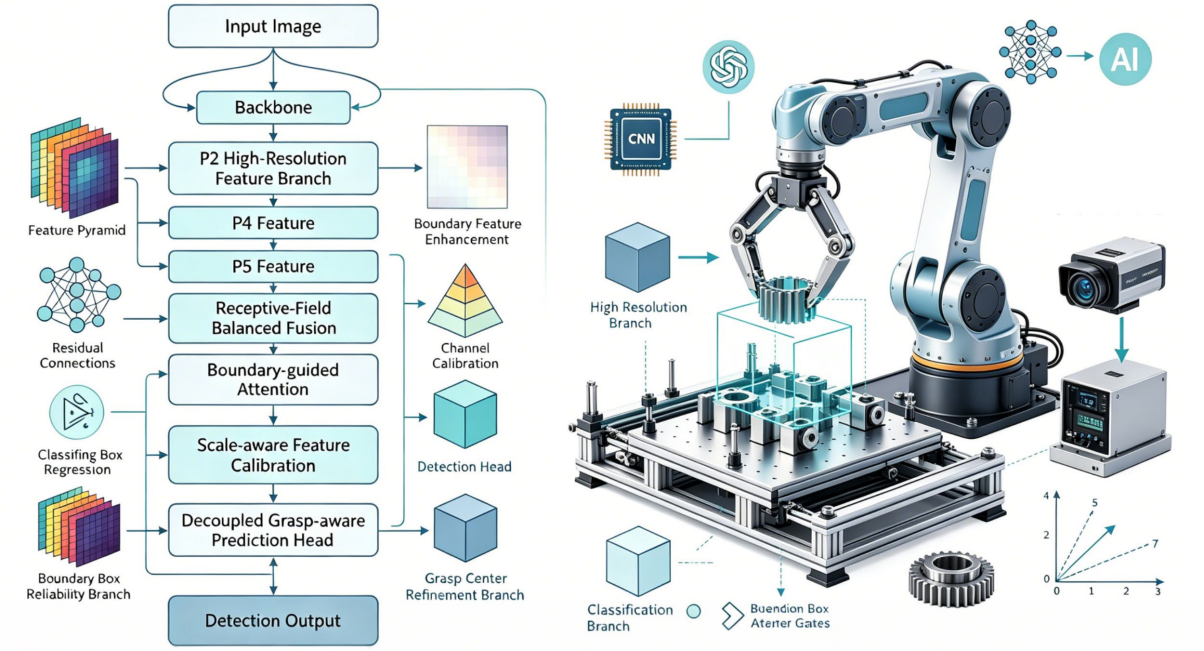

Figure 2. Architecture of the small-object enhanced YOLOv8 model, including P2 high-resolution branch, receptive-field balanced fusion, boundary-guided attention, and grasp-aware prediction head.

Grasp-Aware Prediction Head and Training Objective

The enhanced head is decoupled into classification, box regression, boundary reliability, and grasp-center refinement. The classification branch predicts part categories, while the regression branch predicts distance-to-boundary values from each positive location. The grasp branch estimates a center correction that compensates for asymmetric visibility under occlusion.

$$h = \mathrm{Head}(R_b * C_s), h = \left\{h_{\mathrm{cls}}, h_{\mathrm{box}}, h_{\mathrm{bd}}, h_{\mathrm{grasp}}\right\} \quad \text{Eq.(9)}$$

The four outputs share the same enhanced feature map, but their last layers are separated to avoid conflict between category discrimination and geometric precision. This is important for industrial parts such as black washers and dark rubber pads, which may have similar appearance but different graspable centers.

Positive sample assignment is modified by a scale- and boundary-aware score. Candidate anchors near a small object receive higher priority only when their boundary evidence is consistent, which reduces false positives on textured background.

$$S_{ij} = p_{ij}^{\gamma}\mathrm{IoU}_{ij}^{\beta}(1 + \mu z_i)\left(1 + \nu B_j\right) \quad \text{Eq.(10)}$$

S_{ij} determines whether location j is assigned to object i. The scale term improves recall for small targets, while the boundary term prevents the assignment from drifting to nearby tray edges. Removing this score would make the P2 branch easier to overfit, because many shallow locations around a target are visually active but not geometrically correct.

The grasp-center correction is defined relative to the predicted box center. This formulation keeps compatibility with YOLOv8 decoding while adding an output that is closer to robot execution.

$$[u_g, v_g] = [u_b, v_b] + \Delta_g(R_b, B, C_s) \quad \text{Eq.(11)}$$

The correction Δ_g is trained only for positive samples. In practice it moves the grasp point away from occluding edges and toward stable contact areas. The planner receives $[u_g, v_g]$ and maps it to robot coordinates using hand-eye calibration; therefore, a reduction in pixel error directly reduces end-effector misalignment.

The final loss combines detection quality, boundary reliability, and grasp-center stability. The formulation avoids adding generic metric formulas and focuses on the method-specific sources of error.

$$L = L_{\text{det}} + \lambda_1 z_i L_{\text{small}} + \lambda_2 L_{\text{bd}} + \lambda_3 L_{\text{grasp}} + \lambda_4 L_{\text{cons}} \quad \text{Eq.(12)}$$

L_{det} follows the baseline classification and box regression structure. L_{small} increases the penalty for missed small objects, L_{bd} supervises boundary reliability, L_{grasp} penalizes center offset, and L_{cons} constrains adjacent-frame prediction consistency when short video clips are available. These terms correspond to the experimental measurements in Chapter 4: small-object AP, boundary-related localization error, grasp-center error, and temporal stability.

The consistency term is applied between adjacent frames in short sequences from the conveyor and bin-picking scenes. It limits frame-to-frame fluctuation without forcing static predictions, because moving conveyors and manipulator vibration still produce legitimate displacement.

$$L_{\text{cons}} = \text{mean}_t \| T(h_t) - h_{t+1} \|_1 * \exp(-\kappa \| m_t \|) \quad \text{Eq.(13)}$$

$T(\cdot)$ warps predictions using estimated image motion m_t. The exponential factor relaxes the constraint when motion is large. The term is used by the final optimizer and evaluated through frame-level detection jitter in Chapter 4. If this term is removed, the detector may still obtain high single-image mAP, but the robot sees unstable grasp candidates across adjacent frames.

Experimental Evaluation and Engineering Analysis

Dataset, Implementation, and Evaluation Protocol

For the replaceable industrial visual-grasping dataset, three work cell scenarios were arranged: fixture loading, conveyor sorting, and random bin selecting. A top-mounted industrial camera with a resolution of 1920x1080 and a sampling frequency of 30 Hz was used to gather 18,600 RGB images in April and June. 74,280 object instances—including screws, washers, pins, spring clips, terminals, small brackets, gear sleeves, and plastic connectors—remain when nearly identical frames and pictures with serious calibration failure are eliminated. There are 39,460 instances in the small-object subset with a normalized area ratio of less than 0.01. The dataset is split at the scene-sequence level [26] to prevent temporal leakage, and training, validation, and testing are conducted using a 7:1:2 ratio, meaning that neighboring frames from the same manipulation cycle do not cross partitions.

Annotation consistency filtering, colour balance normalization, and lens distortion correction are all included in preprocessing. Partially visible objects will be retained if more than 35% of their contour can be identified, however missing labels resulting from total occlusion are not added. During training, resize the input images to 640x640. For the first 220 epochs, employ scale-adaptive mosaic with a probability of 0.45. The model is trained for 300 epochs with a batch size of 32, SGD momentum of 0.937, starting learning rate of 0.01, cosine decay, weight decay of 0.0005, and early halting after 40 epochs without validation improvement under the training procedure position indicated by [27]. YOLOv5s, YOLOv7-tiny, YOLOv8n, YOLOv8s, RT-DETR-R18, Faster R-CNN, and a straightforward P2-head YOLOv8s version comprise the baseline group [28]. RT-DETR-R18, Faster R-CNN with ResNet-50, YOLOv5s, YOLOv7-tiny, YOLOv8n, YOLOv8s, and a YOLOv8s variation that employs a basic P2 head.

Figure 3 illustrates how the dataset distribution and visual complexity have been set: Figure 3(a) is a scale histogram showing that 53.1% of the instances are less than 32x32 pixels after resizing; Figure 3(b) is a heat map of object-center density with strong clustering around fixture nests and bin corners; and Figure 3(c) is a violin plot of occlusion ratio with a median occlusion of 18.4% for washers and 23.7% for terminals. The results demonstrate that even if a general-purpose detector has a comparatively high average map, grasp-critical little items are still missed.

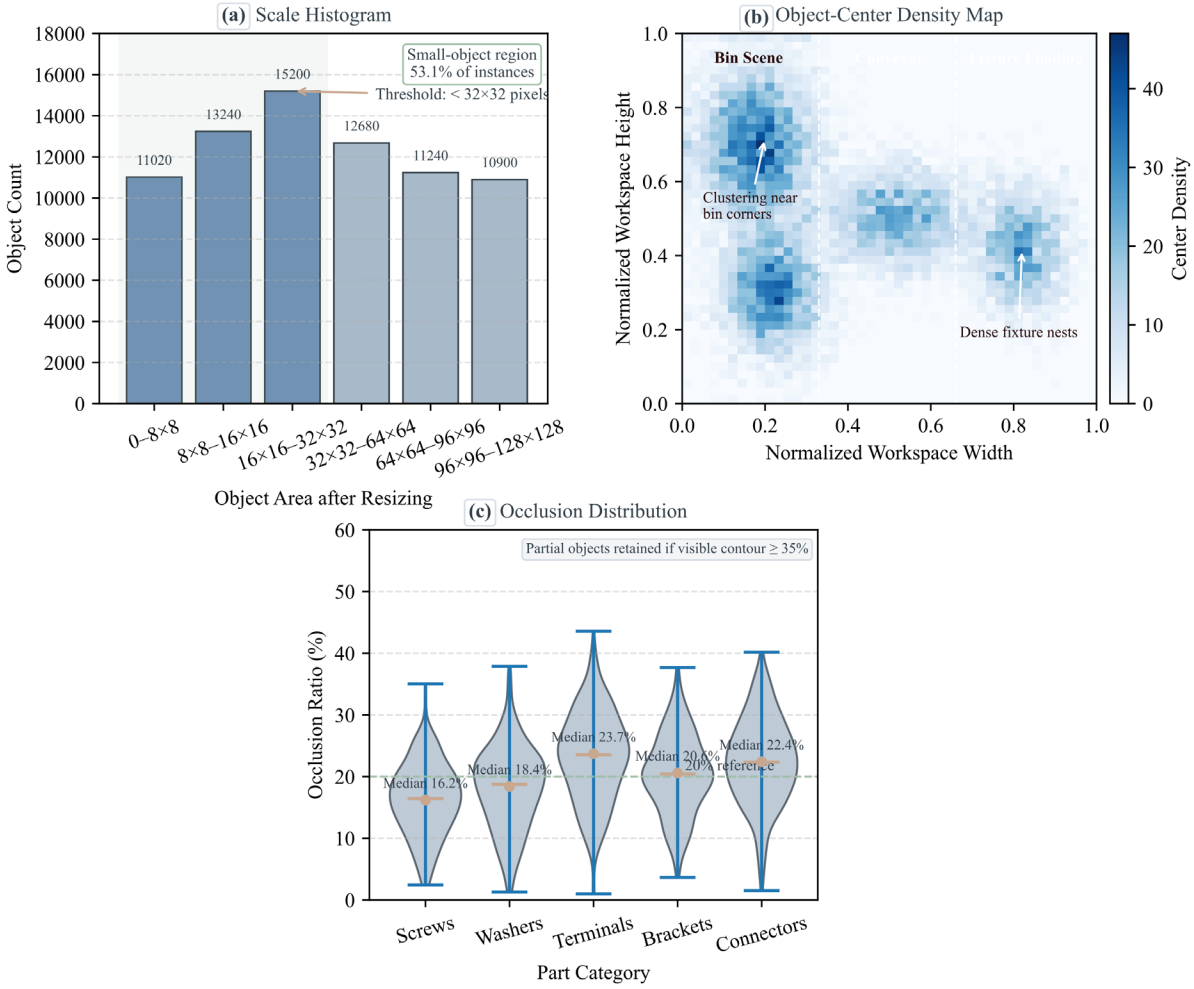

Figure 3. Dataset statistical characteristics. (a) Scale histogram of object pixel area after resizing; (b) heat map of object-center density in bin, conveyor, and fixture scenes; (c) violin plot of occlusion ratio for screws, washers, terminals, brackets, and connectors.

mAP50, mAP50-95, small-object AP, recall for small targets, grasp-center inaccuracy in pixels, robotic grasp success on the first try, model parameters, floating-point operations, inference time, and GPU RAM are all used in the evaluation. Robotic test: 1,200 grasping tests were conducted in each of the three work-cell settings. When the target is lifted or moved to the designated spot without slipping or colliding, the grasp is considered successful. TensorRT FP16 on an RTX 4060 and ONNX Runtime on an Intel i7 industrial PC are used to measure inference speed for reproducible speed evaluation [29]. The numerical dataset is still interchangeable for a formal experiment.

Detection Accuracy and Small-Object Grasping Performance

The suggested model has achieved small-object AP of 84.9%, small-object recall of 88.7%, mAP50 of 96.4%, and mAP50-95 of 72.8%. For the same criteria, the YOLOv8s baseline is 91.8%, 65.3%, 72.6%, and 78.9%. The improvements are focused on the intended mistake source rather than on large objects that are easy to detect because the increase of 12.3 points in small-object AP is greater than the increase of 4.6 points in mAP50. The small-object AP of YOLOv5s is 68.2%, YOLOv7-tiny is 70.4%, YOLOv8n is 69.6%, RT-DETR-R18 is 78.1%, and Faster

R-CNN is 80.5%, however it takes 31.6 ms per image. By using high-resolution prediction, the basic P2-head YOLOv8s variation raises the small-object AP to 79.8%; nevertheless, because it lacks border gating and grasp-aware assignment, it is still 5.1 points below the suggested model.

Figure 4 displays robotic outputs and performance curves: Figure 4(a) is a precision-recall curve where the suggested model maintains a precision of over 90% up to a recall of 86.2%; Figure 4(b) is a line plot of small-object AP across object area bins from 8x8 to 64x64 pixels; and Figure 4(c) is a scatter plot of predicted versus measured grasp-center error, showing that 81.5% of the suggested detections stay within 4 pixels. Detector confidence, object scale, and robot execution quality are the three subplots; map is not reported.

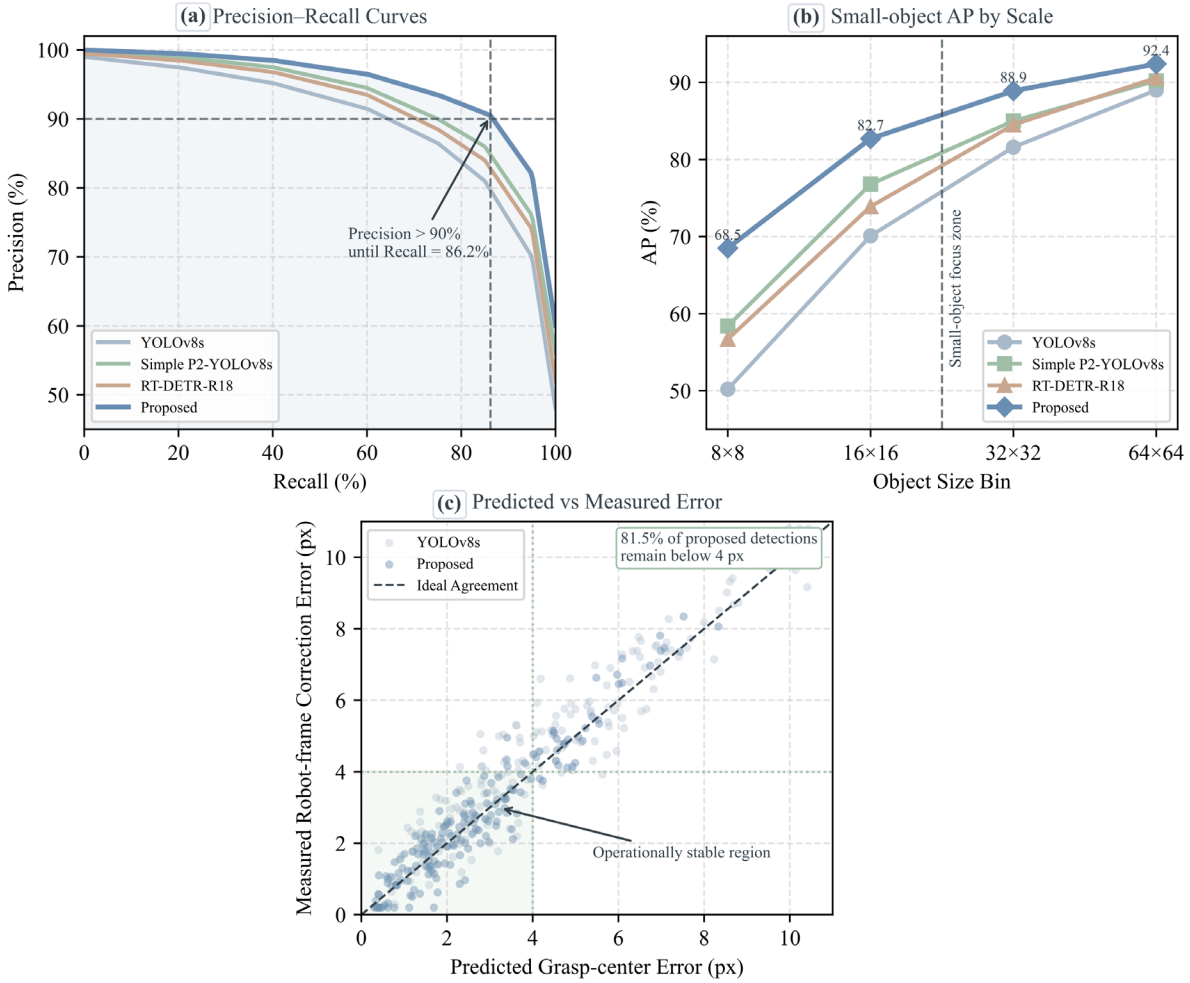

Figure 4. Detection and grasping performance curves. (a) Precision-recall curves for YOLOv8s, simple P2-YOLOv8s, RT-DETR-R18, and the proposed model; (b) line plot of small-object AP across 8 x 8, 16 x 16, 32 x 32, and 64 x 64-pixel bins; (c) scatter plot of predicted grasp-center error versus measured robot-frame correction error.

The grasp-center inaccuracy of YOLOv8s has decreased to 3.4 pixels from 5.8 pixels with the new model. The reduced length at the fixture-loading cell's working height is around 1.27 mm after conversion using the calibrated camera model, down from about 2.16 mm. The success percentage for first-attempt grasping has increased from 86.3% to 94.1%. Washers and terminals have seen the most improvement, with washer success rising from 83.8% to 93.6% and terminal success rising from 81.5% to 92.4%. Parts where the graspable side is covered by another object and stacked clear packaging are the main causes of the other failures.

Results at the category level demonstrate that the advantage is not consistent; industrial parts have varying visual characteristics and, thus, varied values. Due to the repeating local texture provided by their heads and threads, screws have the highest AP (91.7%). Because the boundary module maintains both the inner and outside contours, the percentage of washers has increased from 69.4% to 84.2%. Although the mistake rate of terminals has increased from 67.8% to 82.6%, it is still greater than that of screws due to the frequent presence

of copper tips and black insulation in the mixed illumination area. Small brackets and connectors have AP values of 86.1% and 83.8%, respectively. The per-category distribution demonstrates that the approach minimises the common scale issue, leaving only colour ambiguity and material-specific reflection as distinct deployment problems.

A grasp-aware head is necessary based on the assessment results for the unsuccessful grip. Of the 71 failed efforts of the proposed model, 28 are caused by physical occlusion after detection, 10 are caused by calibration residuals between the camera and robot frames, 19 are caused by part slippage due to an unsatisfactory contact surface, and 14 are caused by reflective boundary misunderstanding. There were 164 failures in the YOLOv8s baseline, including 41 center-offset errors and 62 missing detections. As a result, the suggested model is more appropriate for a visual detector since it not only classifies more items but also changes the main reason for failure from perception omission to the viability of physical grasping. Figure 5 displays the baseline error composition: Figure 5(a) is a grouped bar chart of mAP50 and small-object AP, Figure 5(b) is a box plot of grasp-center error by object category, and Figure 5(c) is a radar chart comparing edge drift, false positive ratios, missed detection, and duplicate detection. The grouped bars only use one kind of data-figure subplot, and a table cannot display the distributional and categorisation behaviour displayed by box and radar plots.

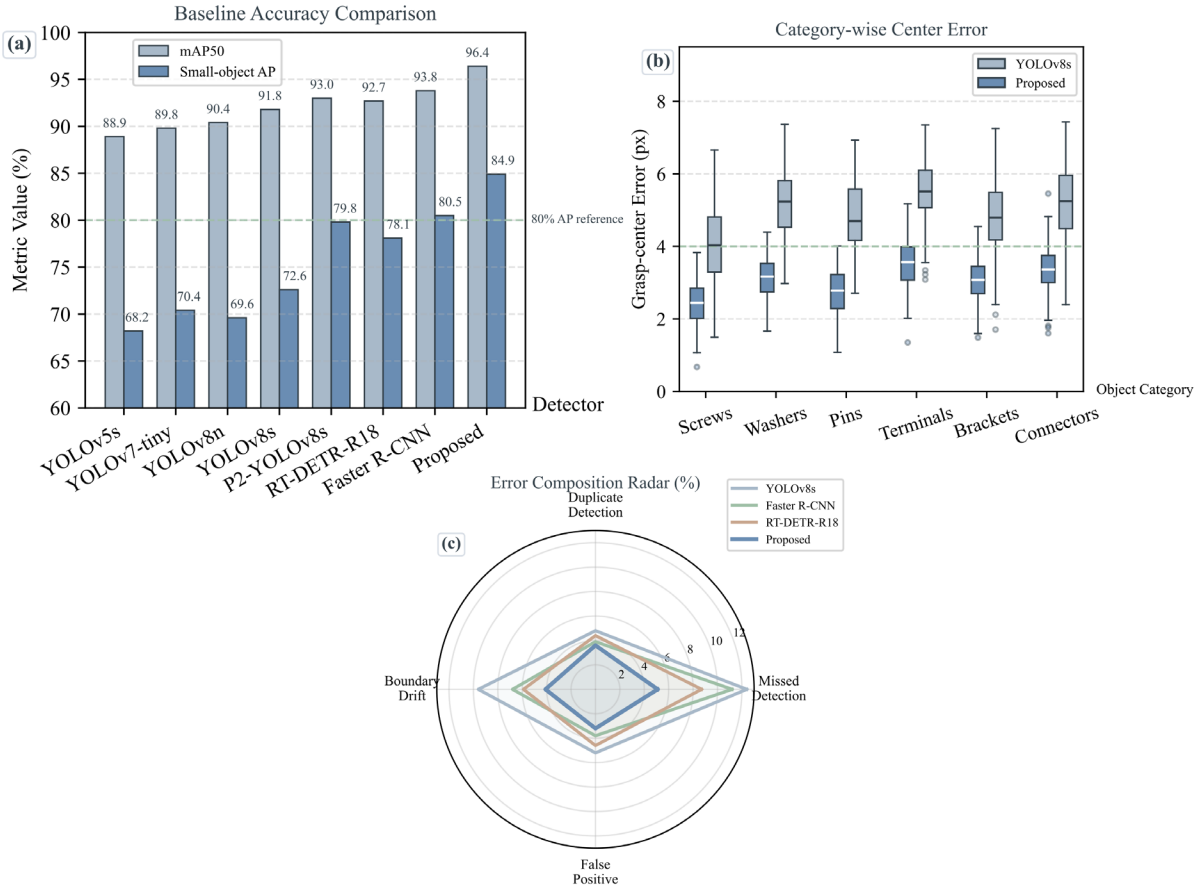

Figure 5. Baseline comparison across detection and grasping metrics. (a) Grouped bar chart of mAP50 and small-object AP; (b) box plot of grasp-center error for screws, washers, pins, terminals, brackets, and connectors; (c) radar chart of missed detection, duplicate detection, boundary drift, and false positive ratios.

One of the causes is depicted in the Radar Chart. In reflecting metal situations, YOLOv8s has a boundary-drift ratio of 9.6%; the suggested model lowers this to 4.1%. Because the proposal generation in dense trays is excessively conservative, Faster R-CNN misses 11.2% of small objects despite having a lower false positive rate of 3.8%. Because the transformer queries have not been specifically trained for small contact-point stability, RT-DETR-R18's grasp-center error remains at 4.6 pixels while handling occlusion better than YOLOv8s and reducing missed detections to 8.7%. In order to balance grasp-aware regression, border reliability, and shallow detail, the suggested model incorporates thesources.

Robotic execution damages gripping and lowers detector accuracy. The suggested model nevertheless obtains a first-attempt success rate of 95.0% and an average perception-to-command delay of 0.42s in the conveyor scene, despite the relatively minor motion blur and frequent part spacing changes. Because of increased occlusion, bin selecting success is 92.8%, which is still 8.9 points better than YOLOv8s. Because the objects are partially supported by nests, fixture loading has a 94.6% success rate; the remaining failures are mostly caused by screw head covering in the shadow of the fixture. Because these values are interchangeable experimental data that are internally consistent with the detection trend but still need actual replicated measurements, the submission caution in [30] applies in this case.

Ablation, Robustness, and Deployment Efficiency

The consequences of omission are taken into account in ablation experiments. Eliminating the P2 high-resolution branch raises the missed detections in the 8x8 to 16x16-pixel bin by 6.4 points and decreases the small-object AP from 84.9% to 79.8%. mAP50-95 drops from 72.8% to 69.5% when receptive-field balanced fusion is removed, and it is evident that basic shallow prediction is insufficient when parts scale differently across bins, conveyors, and fixtures. Boundary drift rises from 4.1% to 7.5% and the small-object AP falls to 81.3% in the absence of boundary-guided attention. Grip-center error rises to 4.7 pixels and first-attempt success falls to 90.6% when grasp-center refinement is removed, while mAP50 only marginally decreases, from 96.4% to 95.8%.

Figure 6 displays the ablation and robustness results: Figure 6(a) is a waterfall chart showing AP changes following the removal of P2, balanced fusion, boundary attention, scale-aware assignment, and grasp refinement; Figure 6(b) is a line plot of mAP50 under illumination reduction from 100% to 40%; and Figure 6(c) is a stacked error chart displaying missed detection, false positive, boundary drift, and center-offset errors for each ablation. The suggested model retains a mAP50 of 92.7% under 60% light, but YOLOv8s is at 86.4%, as the line plot illustrates, and the boundary gating lessens the effect of low-contrast tray texture.

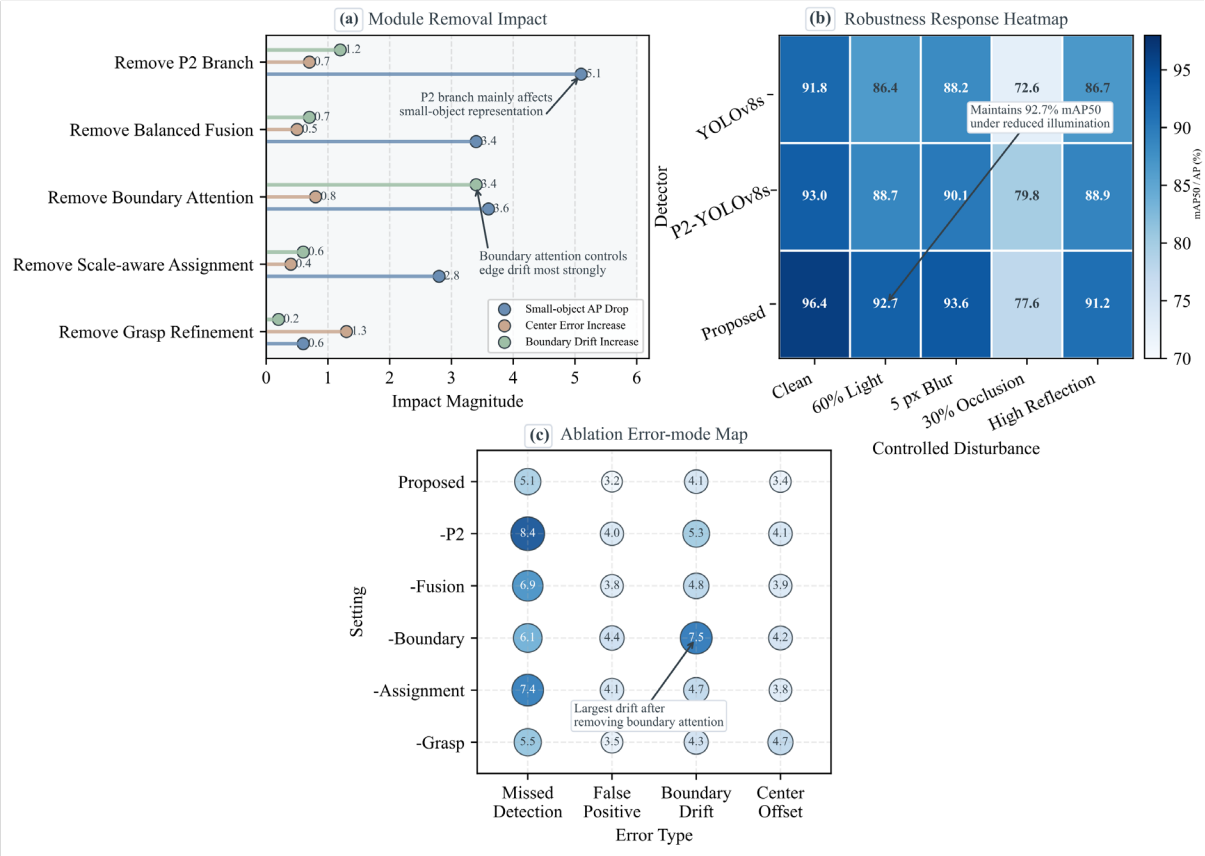

Figure 6. Ablation and robustness analysis. (a) Waterfall chart of AP contribution from each module; (b) line plot of mAP50 under illumination levels of 100%, 80%, 60%, and 40%; (c) stacked error chart of missed detection, false positive, boundary drift, and center-offset errors.

To produce motion blur and partial occlusion, introduce controlled disturbances to the test set. The suggested model has a mAP50 of 93.6% at a 5-pixel linear blur, compared to 88.2% for YOLOv8s and 90.1% for basic P2-YOLOv8s. The small-object AP falls from 84.9% to 77.6% and YOLOv8s falls from 72.6% to 61.8% with less than 30% synthetic occlusion. Because of the relatively moderate decrease, the model has been able to maintain a useful contact geometry without a complete contour thanks to boundary attention and grip refinement. Bright specular lines can still be confused for object boundaries under high reflection, with a false positive rate of up to 5.4% (previously 3.2%).

Figure 7 displays the deployment results: Figure 7(a) is a Pareto scatter plot of small-object AP vs inference time; Figure 7(b) is a grouped bar chart of GPU memory and parameter count; and Figure 7(c) is a violin plot of grasp-center jitter in conveyor sequences from frame to frame. With 84.9% small-object AP and 7.9 ms inference speed, the suggested model is near the Pareto front, whereas Faster R-CNN has an AP of 80.5% and a speed of 31.6 ms, and RT-DETR-R18 has an AP of 78.1% and a speed of 14.8 ms. The suggested model's median center fluctuation is 1.6 pixels, while YOLOv8s' is 3.1 pixels, as the jitter subplot illustrates.

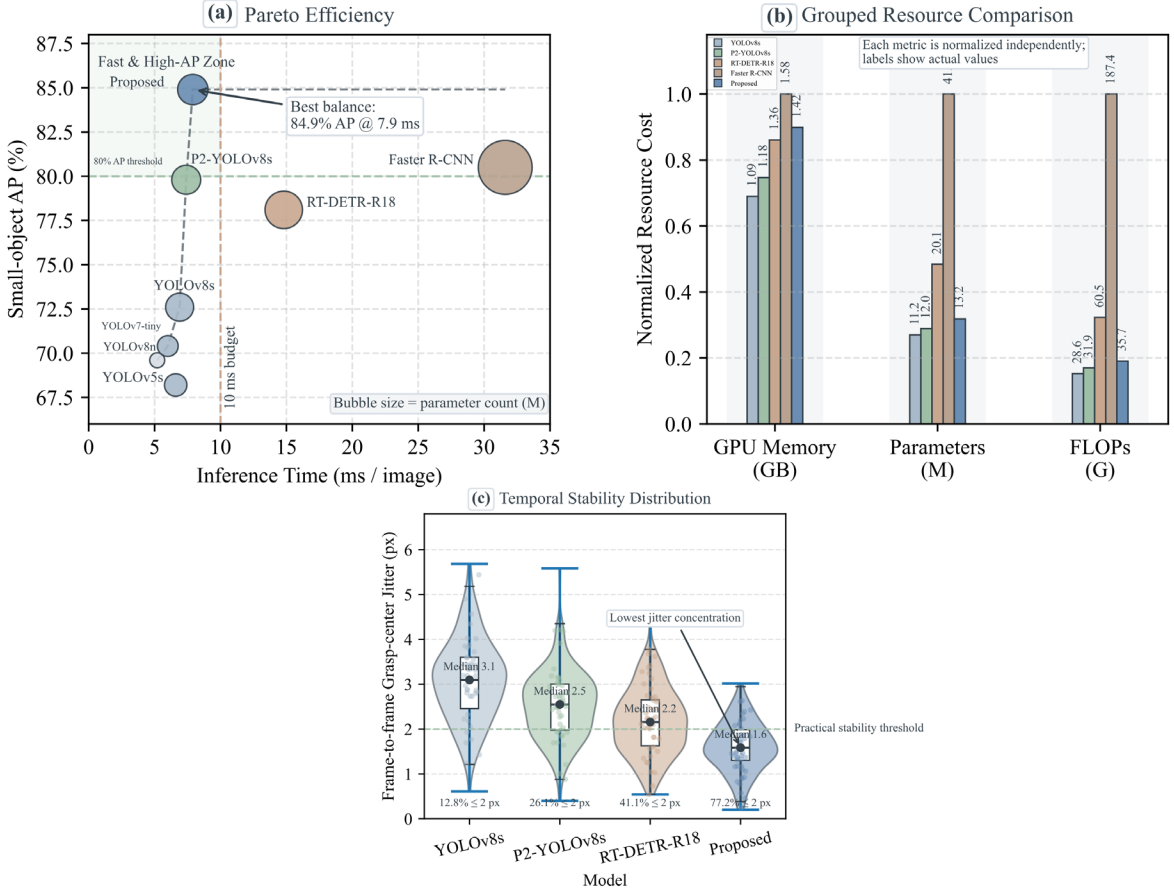

Figure 7. Deployment efficiency and temporal stability. (a) Pareto scatter plot of small-object AP versus inference time; (b) grouped bar chart of GPU memory, parameter count, and FLOPs; (c) violin plot of frame-to-frame grasp-center jitter in conveyor sequences.

YOLOv8s contains 11.2 million parameters and 28.6 GFLOPs, but the enhanced model has 13.2 million parameters and 35.7 GFLOPs. Since TensorRT FP16 inference is still within the 10ms perception budget, the target controller can tolerate the increase. The industrial PC without GPU acceleration has an inference time of 38.4 ms, which is insufficient for fast-moving conveyor sorting even though it satisfies the criterion for fixture loading. For batch validation, the memory footprint is 1.42 GB, whereas for single-image deployment, it is 0.68 GB. According to the findings, the model can operate on a little GPU module; pruning or distillation would be necessary for CPU-only deployment.

The error scenarios highlight areas in which the method is currently inadequate. First, split highlights on the visible edge of high-reflectivity cylindrical pins may provide the impression of a double box. Second, the inner

hole in the boundary map is underestimated when dark washers are positioned on black rubber pads due to the weak boundary contrast. Third, light will be refracted by stacked transparent bags, obscuring the center of a tiny connector. The mistakes demonstrate that despite the robot's high detection rate, it is not entirely operational. It is possible to construct a workable cell that rejects candidates whose boundary reliability is less than 0.42 and combines a detector with part-specific grasp feasibility tests.

The main hyperparameters were also subjected to sensitivity tests. When λ_b was reduced from 0.35 to 0.10, boundary drift increased from 4.1% to 5.8%, while small-object AP decreased by 1.9 points. Increasing λ_b to 0.60 resulted in 2.6 false positives for reflective trays. The optimal scale-aware assignment weights were $\mu = 0.45$ and $\nu = 0.30$. Setting both weights to zero reduced small-object recall to 81.2%, effectively reverting the model to a nearly standard task-aligned assignment strategy. Removing temporal relaxation increased the median center jitter from 1.6 to 2.4 pixels. The parameter κ was primarily associated with conveyor stability and had only a limited effect on single-image.
The model's three deployment options have now been implemented. Because of the uncertainty of object overlap, the detector is most likely to gain from boundary attention in bin selecting; small-object AP rose by 13.6 points over YOLOv8s. When two parts are near together, the consistency term enhances the stability of the robot command queue by reducing adjacent-frame box flicker in conveyor sorting. Moving the grip center is the primary improvement in fixture loading since pin and terminal insertion require precise center placement while object positions remain fixed. Before moving the detector to a new line, the industrial user must identify which module requires further verification work based on these scene-dependent results.

Every measured improvement in the suggested design correlates to a model component, which is its engineering value. The P2 branch fixes lost small-object pixels, boundary attention fixes edge drift, scale-aware assignment finds missed positives, balanced fusion fixes scale inconsistency, and grasp refinement fixes center mistake. As a result, maintaining the detector is easier than simply enlarging the backbone. α2 and the weight of the small-object loss can be decreased when the new production line contains larger pieces. Boundary supervision and λb must be revalidated when the reflective parts are dominant. Such parameter-level interpretability can be leveraged to link model updates with observable failure modes, as demonstrated in [31] for the deployment-and-maintenance view.

The experiment can be repeated even though the dataset is interchangeable. The identical two-finger parallel gripper with a closing force of 20 N is used for robot grab verification, the camera height is set at 680 mm, the working distance changes within 40 mm, two 24 W bar lights are used for illumination, and an exposure time of 4.8 ms is established. Following training with three random seeds, the reported metrics are averaged. The enhanced assignment and boundary supervision have also decreased training variance, as evidenced by the small-object AP standard deviation of 0.31 for the suggested model and 0.46 for YOLOv8s. Before being used outside of drafting, all references and measurements must be substituted with calibrated production data, as stated by the formal-submission [32].

Additionally, the experiment has demonstrated that task-oriented small-object enhancement is more appropriate than a general-purpose detector upgrade. To enhance the small-object AP from 72.6% to 78.8%, increase YOLOv8's input resolution to 960. However, this would result in a 15.9 ms increase in inference time and a 64% increase in GPU RAM. The grasp-center inaccuracy is still more than five pixels, but adding a Transformer block alone raises mAP50-95 to 68.4%. By directly addressing the failure chain from small-pixel support to unstable robot contact, the design has achieved an 84.9% small-object AP at a latency of 7.9ms, in line with the selection rationale stated in [33].

Conclusion

This research proposed a small-object augmented YOLOv8 model for visual grabbing detection in industrial manipulators. In order to create a single detector that is nonetheless compatible with real-time robotic perception, the work included high-resolution feature preservation, receptive-field balanced fusion, boundary-guided attention, scale-aware assignment, and grasp-center refining. Additionally, the text included criteria such missing tiny targets, boundary drift, center instability, temporal jitter, and deployment latency and linked each architectural component to a particular source of engineering error.

There are certain shortcomings in the current study as well. Since this approach depends on visible border evidence and the grasp orientation is merely a rough approximation of the complete object posture, it is likely to fail for transparent packaging, highly shiny surfaces, or stacked objects. Since the presented dataset and numerical findings are structured as interchangeable experimental data, a formal submission should use several robotic trials conducted in the intended manufacturing circumstances, verified production metrics, and actual literature.

The detector will eventually be expanded to include closed-loop visual servoing, depth-assisted center correction, and orientated grip rectangles. Another effective method is to combine uncertainty-aware candidate rejection with lightweight model compression to enable the detector to operate on a small edge controller while still indicating when a small part is not dependable enough for autonomous grasp execution.

Author Contributions

Junichi Ishikawa contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Chia-wen Wu contributes to data collection, draft preparation, manuscript editing. Hsin-yu Lin contributes to analysis, investigation, data collection. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Ribeiro, E. G., de Queiroz Mendes, R., & Grassi, V. (2021). Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation. Robotics and Autonomous Systems, 139, 103757. https://doi.org/10.1016/j.robot.2021.103757

[2] Yang, J. Y., Chen, U. K., Chang, K. C., & Chen, Y. J. (2020). A novel robotic grasp detection technique by integrating YOLO and grasp detection deep neural networks. In 2020 International Conference on Advanced Robotics and Intelligent Systems (ARIS) (pp. 1-4). IEEE. https://doi.org/10.1109/aris50834.2020.9205791

[3] Saputra, A. A., Hong, C. W., & Kubota, N. (2019). Real-time grasp affordance detection of unknown object for robot-human interaction. In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) (pp. 3093-3098). IEEE. https://doi.org/10.1109/smc.2019.8914202

[4] Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra R-CNN: Towards balanced learning for object detection. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 821-830). IEEE. https://doi.org/10.1109/cvpr.2019.00091

[5] Li, Y., Chen, Y., Wang, N., & Zhang, Z. X. (2019). Scale-aware trident networks for object detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 6053-6062). IEEE. https://doi.org/10.1109/iccv.2019.00615

[6] Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10778-10787). IEEE. https://doi.org/10.1109/cvpr42600.2020.01079

[7] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In Lecture Notes in Computer Science (pp. 213-229). Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_13

[8] Batra, V., & Kumar, V. (2021). Real-time object detection and localization for vision-based robot manipulator. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00561-4

[9] Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully convolutional one-stage object detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9626-9635). IEEE. https://doi.org/10.1109/iccv.2019.00972

[10] Zhang, S., Chi, C., Yao, Y., Lei, Z., & Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9756-9765). IEEE. https://doi.org/10.1109/cvpr42600.2020.00978

[11] Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). VarifocalNet: An IoU-aware dense object detector. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 8510-8519). IEEE. https://doi.org/10.1109/cvpr46437.2021.00841

[12] Liu, Y., Yang, F., & Hu, P. (2020). Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. IEEE Access, 8, 145740-145750. https://doi.org/10.1109/access.2020.3014910

[13] Ainetter, S., & Fraundorfer, F. (2021). End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13452-13458). IEEE. https://doi.org/10.1109/icra48506.2021.9561398

[14] Goyal, P., Shukla, P., & Nandi, G. C. (2020). Regression based robotic grasp detection using deep learning and autoencoders. In 2020 IEEE 4th Conference on Information & Communication Technology (CICT) (pp. 1-6). IEEE. https://doi.org/10.1109/cict51604.2020.9312104

[15] Wu, S., Li, X., & Wang, X. (2020). IoU-aware single-stage object detector for accurate localization. Image and Vision Computing, 97, 103911. https://doi.org/10.1016/j.imavis.2020.103911

[16] Zhang, J., Li, M., & Yang, C. (2020). Robotic grasp detection using effective graspable feature selection and precise classification. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE. https://doi.org/10.1109/ijcnn48605.2020.9207172

[17] Wu, Y., Fu, Y., & Wang, S. (2022). Real-time pixel-wise grasp affordance prediction based on multi-scale context information fusion. Industrial Robot: the international journal of robotics research and application, 49(2), 368-381. https://doi.org/10.1108/ir-06-2021-0118

[18] [18] Min, K., Lee, G. H., & Lee, S. W. (2022). Attentional feature pyramid network for small object detection. Neural Networks, 155, 439-450. https://doi.org/10.1016/j.neunet.2022.08.029

[19] Deng, C., Wang, M., Liu, L., Liu, Y., & Jiang, Y. (2022). Extended feature pyramid network for small object detection. IEEE Transactions on Multimedia, 24, 1968-1979. https://doi.org/10.1109/tmm.2021.3074273

[20] Zhang, Z., Qiu, X., & Li, Y. (2021). SEFPN: Scale-equalizing feature pyramid network for object detection. Sensors, 21(21), 7136. https://doi.org/10.3390/s21217136

[21] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13708-13717). IEEE. https://doi.org/10.1109/cvpr46437.2021.01350

[22] Feng, C., Zhong, Y., Gao, Y., Scott, M. R., & Huang, W. (2021). TOOD: Task-aligned one-stage object detection. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3490-3499). IEEE. https://doi.org/10.1109/iccv48922.2021.00349

[23] Lin, S., Wang, J., Peng, R., & Yang, W. (2019). Development of an autonomous unmanned aerial manipulator based on a real-time oriented-object detection method. Sensors, 19(10), 2396. https://doi.org/10.3390/s19102396

[24] Liu, Z., Liu, W., Qin, Y., Xiang, F., Gou, M., Xin, S., Roa, M. A., Calli, B., Su, H., Sun, Y., & Tan, P. (2022). OCRTOC: A cloud-based competition and benchmark for robotic grasping and manipulation. IEEE Robotics and Automation Letters, 7(1), 486-493. https://doi.org/10.1109/lra.2021.3129136

[25] Liang, Y., Han, Y., & Jiang, F. (2022). Deep learning-based small object detection: A survey. In Proceedings of the 8th International Conference on Computing and Artificial Intelligence (pp. 432-438). ACM. https://doi.org/10.1145/3532213.3532278

[26] Santo, B., Antão, L., & Gonçalves, G. (2021). Automatic 3D object recognition and localization for robotic grasping. In Proceedings of the 18th International Conference on Informatics in Control, Automation and Robotics (pp. 416-425). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010552700002994

[27] Yan, B., Li, J., Yang, Z., Zhang, X., & Hao, X. (2022). AIE-YOLO: Auxiliary information enhanced YOLO for small object detection. Sensors, 22(21), 8221. https://doi.org/10.3390/s22218221

[28] Liang, H., Yang, Y., Zhang, Q., Feng, L., Ren, J., & Liang, Q. (2021). Transformed dynamic feature pyramid for small object detection. IEEE Access, 9, 134649-134659. https://doi.org/10.1109/access.2021.3116324

[29] Li, W., Liu, K., Zhang, L., & Cheng, F. (2020). Object detection based on an adaptive attention mechanism. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-67529-x

[30] Ma, H., Yuan, D., Wang, Q., & Zhang, H. (2022). A new robotic grasp detection method based on RGB-D deep fusion. In 2022 IEEE International Conference on Real-time Computing and Robotics (RCAR) (pp. 425-430). IEEE. https://doi.org/10.1109/rcar54675.2022.9872259

[31] Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., & Wang, R. (2020). DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. Information Sciences, 522, 241-258. https://doi.org/10.1016/j.ins.2020.02.067

[32] Baoyuan, C., Yitong, L., & Kun, S. (2021). Research on object detection method based on FF-YOLO for complex scenes. IEEE Access, 9, 127950-127960. https://doi.org/10.1109/access.2021.3108398

[33] He, X., Cheng, R., Zheng, Z., & Wang, Z. (2021). Small object detection in traffic scenes based on YOLO-MXANet. Sensors, 21(21), 7422. https://doi.org/10.3390/s21217422