

Comparative Performance of YOLO Versions for Pallet Detection SCARA Robot

Dilham Hidayatul Fajri¹, Risfendra¹

¹Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Padang, Padang, Indonesia

Article Info	ABSTRACT
<i>Article history:</i> Received April 27, 2026 Revised May 23, 2026 Accepted June 01, 2026 <i>Keywords:</i> YOLO Comparative Study Pallet Detection Machine Learning	The rapid evolution of the You Only Look Once (YOLO) object detection family from YOLOv8 to YOLO26 has produced architecturally diverse models whose relative suitability for structured industrial detection tasks remains unevaluated. This study aims to conduct a controlled comparative performance evaluation of six YOLO medium variants (YOLOv8, YOLOv9, YOLOv10, YOLOv11, YOLOv12, and YOLO26) on an industrial pallet detection dataset to identify the optimal model for Vision-Guided Robotics (VGR) pick and place automation. All models were trained under identical conditions 50 epochs, batch size 16, 640×640 input resolution on a dataset of 4,366 annotated images across three classes (pallet_pick, pallet_place_1, pallet_place_2) using a 70/20/10 train/validation/test split. Performance was assessed using Precision, Recall, mAP@0.5, mAP@0.5:0.95, and training time. All six models achieved perfect Precision (1.0000), Recall (1.0000), and mAP@0.5 (0.9950). Differentiation emerged at mAP@0.5:0.95, where YOLOv9m achieved the highest final-epoch value (0.9776) and YOLO26m achieved the highest peak value (0.9805) with the fastest training time (80.8 minutes, 26.3% faster than YOLOv8). YOLO26 is recommended as the optimal model for industrial pallet detection, offering the best combination of localization accuracy, training efficiency, and deployment readiness through its NMS-free and DFL-free architecture.
<i>Corresponding Author:</i> Risfendra Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Padang Kampus UNP Pusat, Jl. Prof. Hamka, Air Tawar, Padang 25131, Indonesia Email: risfendra@ft.unp.ac.id	

1. INTRODUCTION

The rapid expansion of Industry 4.0 has placed increasing demand on intelligent manufacturing systems to perform complex tasks such as automated pick and place with high speed, precision, and adaptability. Vision-Guided Robotics (VGR), which equips robots with real-time object localization capabilities through computer vision, has emerged as the primary enabler of flexible robotic automation, particularly in structured industrial environments involving palletized workpieces [1]–[5]. Among the various computer vision algorithms employed in VGR systems, the You Only Look Once (YOLO) family has established itself as the most widely adopted framework for real-time object detection, offering a single-pass convolutional neural network inference pipeline that combines high detection accuracy with low latency properties essential for time-critical manufacturing cycles [6]–[8]. Since its introduction in 2016, the YOLO family has undergone continuous architectural evolution, with each successive version introducing targeted innovations in backbone design, loss functions, label assignment strategies, and deployment efficiency [9].

The recent period from 2023 to 2025 has witnessed the most rapid and architecturally diverse progression in YOLO's history, producing six significant versions that each represent distinct design philosophies. YOLOv8 (2023), developed by Ultralytics, established the modern anchor-free baseline with an enhanced CSPNet backbone, offering broad multi-task support and extensive deployment compatibility [10]. YOLOv9 (February 2024) introduced the Generalized Efficient Layer Aggregation Network (GELAN) and

Programmable Gradient Information (PGI) to address information loss in deep networks, improving representational efficiency without increasing inference cost [11]. YOLOv10 (May 2024), from Tsinghua University, pioneered NMS-free training through consistent dual-label assignments, reducing post-processing latency while achieving competitive accuracy [12]. YOLO11 (October 2024) further refined architectural efficiency through compact C3k2 bottlenecks and a C2PSA spatial attention module, achieving 22% fewer parameters than YOLOv8 while maintaining or exceeding its mAP on standard benchmarks [13]. YOLOv12 (February 2025) marked a shift toward attention-centric design, introducing Area Attention (A2) modules and Residual Efficient Layer Aggregation Networks (R-ELAN) to maximize detection accuracy through enhanced global context modeling [14]. Most recently, YOLO26 (September 2025) implemented the most radical architectural simplifications in the family's history removing both the Distribution Focal Loss (DFL) decoding module and the Non-Maximum Suppression (NMS) post-processing step, introducing Progressive Loss Balancing (ProgLoss), Small-Target-Aware Label Assignment (STAL), and the MuSGD optimizer delivering up to 43% faster CPU inference than predecessor models while improving small-object detection accuracy [15], [16].

Despite the rich architectural diversity across these six YOLO versions, the industrial robotics community lacks a systematic and controlled comparative evaluation of their performance on domain-specific tasks such as pallet detection for SCARA robot pick and place automation. Existing comparative studies of YOLO versions have predominantly focused on general benchmarks such as MS COCO [7], agricultural applications [17], or traffic and autonomous driving scenarios [18],[19], with limited attention to structured industrial pallet detection where the visual characteristics fixed geometric layouts, high inter-class similarity between placement zones, and consistent lighting present a unique and distinct detection challenge. Furthermore, most VGR studies employing YOLO-based detection have adopted a single model version without justifying the selection relative to alternatives, leaving practitioners without a principled basis for model choice in industrial deployment contexts [20],[21]. A recent comparative study on YOLO object detectors specifically for robotic vision tasks confirmed that model suitability varies significantly depending on the target application and hardware constraints, further reinforcing the need for domain-specific benchmarking [22].

This study addresses the identified gap by conducting a controlled comparative performance evaluation of six YOLO model variants YOLOv8, YOLOv9, YOLOv10, YOLO11, YOLOv12, and YOLO26m on a custom industrial pallet detection dataset comprising 4,366 annotated images across three object classes: `pallet_pick`, `pallet_place_1`, and `pallet_place_2`. All models are trained under strictly identical experimental conditions 50 epochs, batch size 16, input resolution 640×640 pixels using the same dataset split (70% training / 20% validation / 10% test) to ensure a fair and reproducible comparison. Performance is evaluated using standard detection metrics (Precision, Recall, mAP@0.5, mAP@0.5:0.95) supplemented by inference efficiency indicators (inference speed and parameter count) to provide a complete trade-off analysis for industrial deployment.

The novelty and contributions of this work are threefold: (1) it presents the first controlled comparative benchmark of YOLOv8 through YOLO26 on an industrial pallet detection task, covering the full span of modern YOLO architectures from 2023 to 2025; (2) it provides a domain-specific performance ranking and trade-off analysis accuracy versus inference efficiency that directly guides model selection for Vision-Guided Robotics systems in industrial pick and place automation; and (3) it delivers empirical evidence on how architectural innovations across successive YOLO versions including NMS-free inference, attention mechanisms, and edge-optimized training strategies translate into real-world detection performance on structured industrial objects, contributing practical knowledge to the growing body of literature on intelligent manufacturing automation.

2. METHOD

This study employs a controlled experimental design to ensure a fair and reproducible comparison across all six YOLO model variants. All models are trained and evaluated on the same dataset, under identical training configurations, and assessed using the same set of evaluation metrics. The overall research workflow consists of four sequential phases.

2.1. Dataset Preparation

The dataset used in this comparative study was constructed from video recordings captured within the Teaching Factory environment at Universitas Negeri Padang, using a Logitech C920 webcam mounted at a fixed height of 65 cm above the working area. Three video sequences were recorded from different viewpoints to introduce sufficient variation in perspective, pallet orientation, and lighting condition. Individual frames were extracted from each video, yielding a total of 4,366 annotated images distributed across three object classes: `pallet_pick`, `pallet_place_1`, and `pallet_place_2`. Dataset example and classes can be view at figure 1.

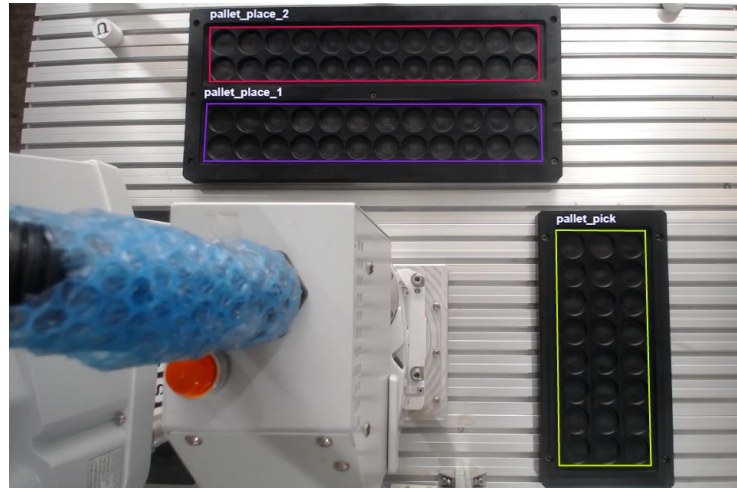


Figure 2. Dataset annotation with roboflow

All annotation and dataset management were performed using the Roboflow platform. Each image was manually labeled with bounding boxes assigned to the corresponding class. To enhance model generalization and robustness against real-world variation encountered during robot operation, the following offline augmentation techniques were applied to the training subset: random rotation ($\pm 15^\circ$), brightness and contrast adjustment, horizontal flipping, and random scaling. The augmented dataset was partitioned following a 70/20/10 strategy, as summarized in Table 1.

Table 1. Dataset Distribution

Subset	Proportion	Number of Images
Training Set	70%	3,075
Validation Set	20%	865
Train Set	10%	426
Total	100%	4,366

The same dataset including identical train/validation/test splits and augmentation was used across all six models to ensure that observed performance differences are attributable solely to architectural differences between YOLO versions, not to data variation.

2.2. YOLO Model Variants

Six YOLO model variants representing the complete span of modern YOLO architecture development from 2023 to 2025 were selected for comparison. The medium-size variant (suffix m) was chosen for all models to provide a balanced and consistent basis for comparison in terms of capacity and computational cost. Table 2 summarizes the key architectural characteristics of each model.

Table 2. YOLO Model Variants and Key Architectural Characteristics

Model	Year	Backbone/Key Architecture
YOLOv8	2023	Anchor-free CSPDarknet, C2f module
YOLOv9	2024	GELAN + Programmable Gradient Info
YOLOv10	2024	Dual-head consistent label assign.
YOLOv11	2024	C3k2 bottleneck + C2PSA attention
YOLOv12	2025	Area Attention (A2) + R-ELAN
YOLOv26	2025	NMS-free + DFL-free + MuSGD + STAL

2.3. Training Configuration

To ensure a fair comparison, all six models were trained under strictly identical experimental conditions. No model-specific hyperparameter tuning was applied beyond the default configurations provided by each model's official implementation. The training configuration is specified in Table 3.

Table 3. Unified Training Configuration for All Models

Parameter	Value
Input Image	640 × 640 pixels
Epochs	50
Batch Size	16
Number of Classes	3

All models were initialized from their respective official pretrained weights on MS COCO before fine-tuning on the pallet detection dataset. The best-performing model checkpoint at each run selected based on the highest validation mAP@0.5 and lowest validation loss was saved as best.pt and used for all subsequent evaluation. The total combined training loss for each model follows the general formulation:

$$L_{total} = \lambda_1 L_{box} + \lambda_2 L_{obj} + \lambda_3 L_{cls} \quad (1)$$

where L_{box} is the bounding box regression loss, L_{obj} is the objectness confidence loss, L_{cls} is the classification loss, and $\lambda_1, \lambda_2, \lambda_3$ are the corresponding loss balancing weights. The specific loss computation method differs per model according to their respective architectural formulations

2.4. Evaluation Metrics

All six models were evaluated on the held-out 426-image test set using five standard performance metrics. Detection accuracy is assessed through Precision (P), Recall (R), F1-Score, and mean Average Precision at IoU thresholds of 0.5 (mAP@0.5) and 0.5:0.95 (mAP@0.5:0.95):

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

$$mAP@0.5 = \frac{1}{C} \sum_{c=1}^C AP_c \quad (5)$$

where TP , FP , and FN denote true positive, false positive, and false negative detections respectively; C is the total number of classes; and AP_c is the Average Precision for class c , computed as the area under the Precision-Recall curve at an Intersection over Union (IoU) threshold of 0.5. mAP@0.5:0.95 extends this evaluation across IoU thresholds from 0.50 to 0.95 in steps of 0.05, providing a stricter measure of bounding box localization quality.

2.5. Experimental Procedure

The complete experimental procedure for each of the six models follows these sequential steps:

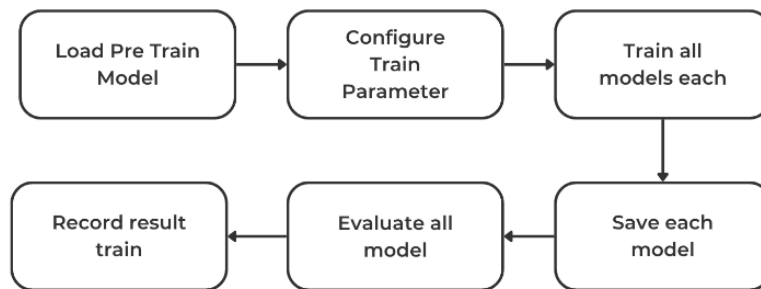


Figure 2. Experimental Procedure

This procedure is repeated independently for all six models, ensuring that no information leakage or cross-model dependency occurs between training runs. The collected results are then compiled into a unified comparative analysis as presented in result.

3. RESULTS AND DISCUSSION

This section presents and discusses the comparative performance of six YOLO model variants trained and evaluated under identical conditions on the pallet detection dataset. Results are organized across four analytical dimensions: training loss convergence, detection accuracy metrics, confidence curve analysis, and overall comparative discussion including training efficiency.

3.1. Training and Validation Loss Convergence

The training box loss, classification loss, and DFL loss curves for all six models across 50 epochs are presented in Figure 3. All models demonstrated consistent downward convergence throughout training with no evidence of overfitting, as confirmed by the parallel decline of validation losses alongside training losses in each case.

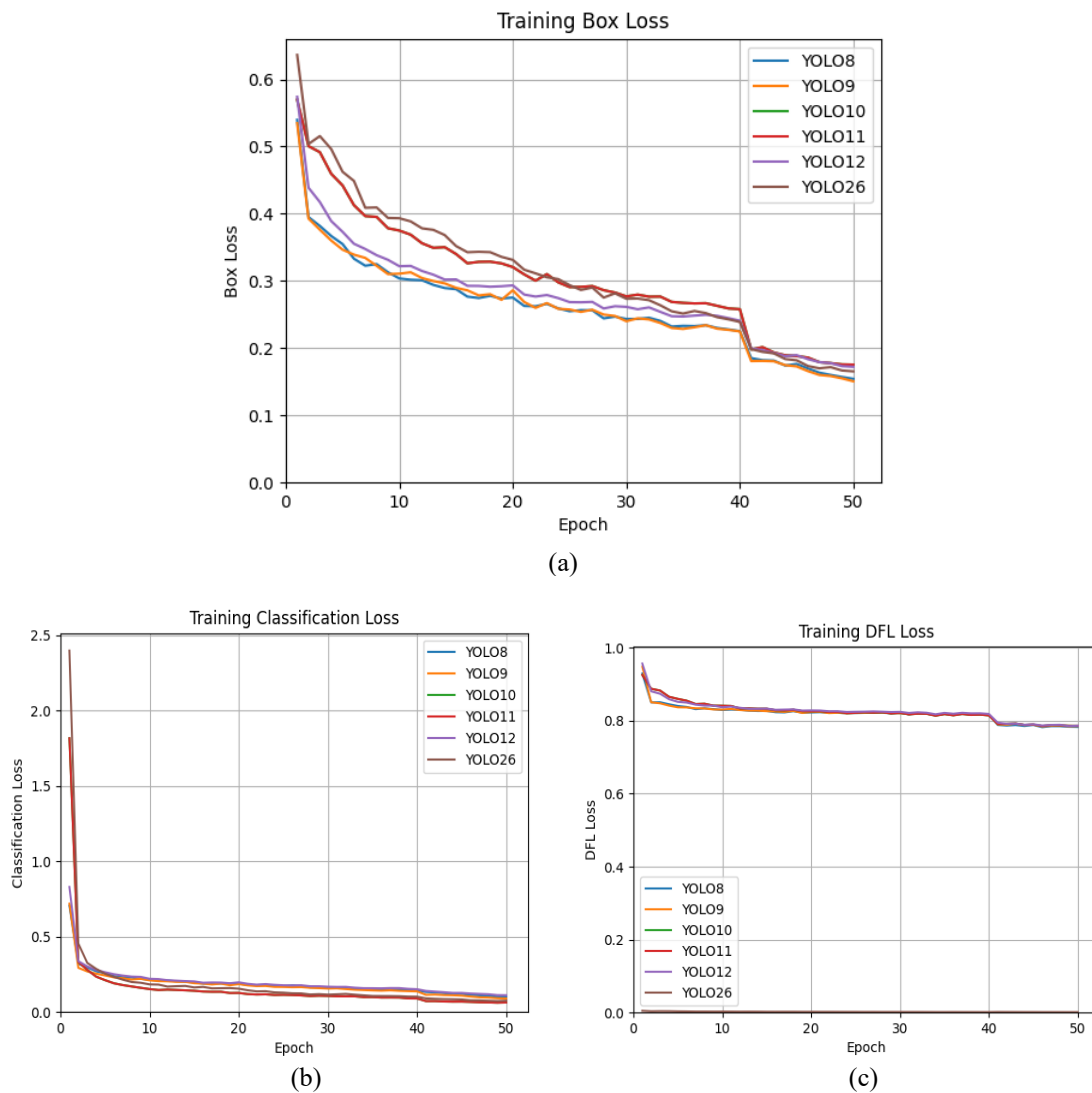


Figure 3. Model train result (a) training box loss (b) classification loss (c) DFL loss

At epoch 1, the training box losses of all models were comparable, ranging from 0.5349 (YOLOv9) to 0.6364 (YOLO26), reflecting their similar initialization conditions from pretrained COCO weights. By epoch 50, box losses converged to a narrow range of 0.1507–0.1755, indicating that all models reached a similar level of bounding box regression quality. Classification losses exhibited more pronounced differences at initialization: YOLO26 started with the highest classification loss (2.3985) due to its distinct MuSGD optimizer dynamics, while YOLOv8, YOLOv9, and YOLOv12 started in the range of 0.71–0.83. By epoch 50, YOLO26 achieved the lowest final classification loss (0.0748), followed by YOLOv9 (0.0896) and YOLO11/YOLOv10 (0.0632).

Table 4. Unified Training Configuration for All Models

Model	Box Epoch 1	Box Epoch 50	Box $\Delta\%$	Cls Epoch 1	Cls Epoch 50	Cls $\Delta\%$
YOLOv8	0.5398	0.1544	71.4%	0.7056	0.1016	85.6%
YOLOv9	0.5349	0.1507	71.8%	0.7189	0.0896	87.5%
YOLOv10	0.5697	0.1755	69.2%	1.8166	0.0632	96.5%
YOLOv11	0.5697	0.1755	69.2%	1.8166	0.0632	96.5%
YOLOv12	0.5741	0.1723	70.0%	0.8298	0.1113	86.6%
YOLOv26	0.6364	0.1655	74.0%	2.3985	0.0748	96.9%

The validation losses for all models followed closely their respective training curves, with no significant divergence observed. YOLO26 achieved the lowest final validation box loss (0.1577) and validation classification loss (0.0534), suggesting the best generalization among all models. YOLOv9 attained the second-lowest validation box loss (0.1546) and validation classification loss (0.0797). These results confirm that all six models successfully learned discriminative representations for the three pallet classes without overfitting to the training data. The final detection performance of all six models evaluated on the 426-image test set at epoch 50 is summarized in Table 5.

Table 5. Final Performance at epoch 50

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95	Best mAP@0.5:0.95	Best Epoch
YOLOv8	1.0000	1.0000	0.9550	0.9738	0.9794	45
YOLOv9	1.0000	1.0000	0.9550	0.9776	0.9782	33
YOLOv10	1.0000	1.0000	0.9550	0.9736	0.9778	28
YOLOv11	1.0000	1.0000	0.9550	0.9736	0.9778	28
YOLOv12	1.0000	1.0000	0.9550	0.9718	0.9766	29
YOLOv26	1.0000	1.0000	0.9550	0.9758	0.985	44

All six models achieved a Precision of 1.0000 and Recall of 1.0000 at the final epoch, indicating zero false positives and zero missed detections on the validation set. The mAP@0.5 was uniformly 0.9950 across all models, reflecting that the pallet detection task with its structured geometric layouts and sufficient dataset size is well within the representational capacity of all modern YOLO variants at this scale.

The primary differentiator among models is the mAP@0.5:0.95, which measures bounding box localization precision across multiple IoU thresholds and therefore captures subtle differences in detection quality that mAP@0.5 alone cannot reveal. At epoch 50, YOLOv9 achieved the highest mAP@0.5:0.95 of 0.9776, followed by YOLO26 (0.9758), YOLOv8 (0.9738), YOLOv10 and YOLO11 (both 0.9736), and YOLOv12 (0.9718). When considering the best mAP@0.5:0.95 achieved at any epoch during training, YOLO26 attained the highest peak value of 0.9805 at epoch 44, surpassing YOLOv8's peak of 0.9794 at epoch 45 and YOLOv9's peak of 0.9782 at epoch 33. This indicates that YOLO26's localization capability continues to improve over more training epochs compared to other models, making it the most suitable choice for extended training regimes.

While detection accuracy metrics are similar across all models for this structured pallet detection task, significant differences emerge in terms of training time efficiency. Table 6 summarizes the total training time required by each model to complete 50 epochs, ranked from fastest to slowest.

Table 6. Training Time

Model	Training Time (s)
YOLOv26	4,845
YOLOv12	5,248
YOLOv11	5,823
YOLOv10	8,925
YOLOv9	6,092
YOLOv8	6,570

YOLO26 demonstrated the fastest training completion at 4,845 seconds (80.8 minutes), representing a 26.3% reduction in training time compared to YOLOv8 (6,570 seconds, 109.5 minutes). This advantage is attributable to YOLO26's architectural simplifications the removal of DFL decoding and NMS post-processing steps reduces computational overhead in both the forward and backward passes during training. YOLOv12 was the second fastest at 5,248 seconds (87.5 minutes), benefiting from its FlashAttention-based efficient computation.

4. CONCLUSION

This study conducted a controlled comparative evaluation of six YOLO model variants YOLOv8, YOLOv9, YOLOv10, YOLO11, YOLOv12, and YOLO26 on an pallet detection dataset comprising 4,366 images across three classes under identical training conditions of 50 epochs, batch size 16, and 640×640 input resolution, confirming that all models achieved perfect Precision (1.0000), Recall (1.0000), and mAP@0.5 (0.9950), while meaningful differentiation emerged at mAP@0.5:0.95 where YOLO26 attained the highest peak localization accuracy of 0.9805 and the fastest training completion at 80.8 minutes (26.3% faster than YOLOv8) establishing YOLO26 as the optimal model for industrial pallet detection in Vision-Guided Robotics applications due to its superior combination of localization precision, training efficiency, and deployment-friendly NMS-free and DFL-free architecture.

REFERENCES

- [1] C. Gabellieri, A. Paleschi, L. Pallottino and M. Garabini, "Autonomous Unwrapping of General Pallets: A Novel Robot for Logistics Exploiting Contact-Based Planning," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 2, pp. 1194-1211, April 2023
- [2] C. D. Santana, "Learning from humans how to grasp: A data-driven architecture for autonomous grasping with anthropomorphic soft hands," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1533–1540, Apr. 2019.
- [3] A. Singh, V. Kalaichelvi, and R. Karthikeyan, "A survey on vision guided robotic systems with intelligent control strategies for autonomous tasks," *Cogent Eng.*, vol. 9, no. 1, p. 2050020, 2022, doi: 10.1080/23311916.2022.2050020.
- [4] J. -S. Kang, J. K. Kang, S. Son, S. H. Park, E. -J. Jung and G. R. Cho, "Image Enhancement Method for Detection in Logistics Loading and Unloading Robots," *IEEE Access*, vol. 13, pp. 143678-143687, 2025, doi: 10.1109/ACCESS.2024.3434599.
- [5] R. R. Arakerimath, A. H. Badve, D. P. Deshmukh, M. N. Chandan and A. Badadhe, "Performance Modelling and Simulation of Cobot for Pick and Place Application Using RoboDK," *2025 International Conference on Emerging Technologies in Electronics and Green Energy (ICETEG)*, MYSORE, India, 2025, pp. 1-6, doi: 10.1109/ICETEG66194.2025.11473126.
- [7] R. Sapkota et al., "YOLO advances to its genesis: a decadal and comprehensive review of the You Only Look Once (YOLO) series," *Artif. Intell. Rev.*, vol. 58, no. 9, 2025, doi: 10.1007/s10462-025-11253-3.
- [8] S. Kang, Z. Hu, L. Liu, K. Zhang, and Z. Cao, "Object Detection YOLO Algorithms and Their Industrial Applications: Overview and Comparative Analysis," *Electronics*, vol. 14, no. 6, pp. 1–36, 2025, doi: 10.3390/electronics14061104.
- [9] R. Sapkota and M. Karkee, "Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8, and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition," *arXiv preprint, arXiv:2510.09653*, 2025.
- [10] G. Jocher et al., "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [11] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *arXiv preprint, arXiv:2402.13616*, 2024.
- [12] A. Wang et al., "YOLOv10: Real-Time End-to-End Object Detection," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 107984–108011, 2024.
- [13] X. Ren, Z. Liu, H. Cheng and L. Chen, "DDR-YOLO: An Efficient Detection Model Based on YOLO11s for Steel Surface Defect," in *IEEE Access*, vol. 13, pp. 147202-147217, 2025, doi: 10.1109/ACCESS.2025.3599978
- [14] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," *arXiv preprint, arXiv:2502.12524*, 2025.
- [15] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "YOLO26: Key Architectural Enhancements and Performance Benchmarking for Real-Time Object Detection," *arXiv preprint, arXiv:2509.25164*, 2025.
- [16] Y. Gu, "CRP-YOLO26: A Lightweight Improved YOLO26 Algorithm for Small Object Detection in UAV Images," *2026 5th International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AloTC)*, Nanchang, China, 2026, pp. 683-689, doi: 10.1109/AloTC71250.2026.11689104.
- [17] R. Sapkota et al., "Comprehensive performance evaluation of YOLOv12, YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments," *Smart Agric. Technol.*, 2026, doi: 10.1016/j.atech.2026.
- [18] P. Nizieniec, M. Iwanowski et al., "YOLO Object Detectors for Robotics a Comparative Study," *arXiv preprint, arXiv:2603.27029*, 2026.
- [19] A. Minsandi et al., "Design and Implementation of Robot Abu Robocon Using Joystik Wireless Based on Extrasensory Perception," *J. Ind. Autom. Electr. Eng.*, vol. 01, no. 01, pp. 39–45, 2024..
- [20] T. P. Kapusi, T. I. Erdei, G. Husi, and A. Hajdu, "Application of Deep Learning in the Deployment of an Industrial SCARA Machine for Real-Time Object Detection," *Robotics*, vol. 11, no. 4, 2022, doi: 10.3390/robotics11040069.
- [21] V. A. Putra, "Cycle time analysis of automation system using PLC and HMI based Scara Robot," *J. Ind. Autom. Electr. Eng.*, vol. 02, no. 01, pp. 112–118, 2025.
- [22] P. Nizieniec and M. Iwanowski, "YOLO Object Detectors for Robotics a Comparative Study," *arXiv preprint, arXiv:2603.27029*, 2026.