A Hybrid Deep Learning Framework for Robust Object Detection in Smart Industrial and Medical Imaging Applications

Hewa Majeed Zangana[1]*
[1]IT Department, Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
[1] hewa.zangana@dpu.edu.krd*
*Corresponding author

ABSTRACT

The convergence of deep learning and real-time object detection has enabled significant advancements in both industrial automation and medical diagnostics. However, a persistent challenge lies in the fragmentation between 2D and 3D object detection models, which limits scalability and cross-domain applicability. This study proposes a hybrid object detection framework that integrates convolutional features with classical template matching, aiming to improve detection robustness, especially in cluttered and occluded scenes. The proposed method is evaluated using standard 2D datasets and occlusion-heavy custom scenarios, showing improved performance in terms of precision (89.2%), recall (87.5%), and mAP@0.5 (85.6%) compared to Faster R-CNN and YOLOv3. It achieves real-time inference speeds of 18.9 FPS (1080p) and 29.4 FPS (720p). While the paper discusses potential applications for industrial and medical domains, evaluations using LASIESTA, KITTI, or CT-scan datasets are presented as conceptual use cases rather than direct experimental results. The findings suggest the framework is promising for deployment in safety-critical environments such as robotic inspection systems and diagnostic imaging workflows.

Keywords: Cross-Domain Learning; Deep Learning; Industrial Automation; Medical Imaging; Object Detection.

Article History		
Received : June, 23rd 2023	Accepted: August, 2nd 2025	Published: Nov, 4th 2025

I. INTRODUCTION

Object detection has emerged as a pivotal task in computer vision, enabling machines to perceive and interpret their environment across various domains such as manufacturing, healthcare, autonomous vehicles, and surveillance [1][2]. With the advent of deep learning, especially Convolutional Neural Networks (CNNs) and attention-based mechanisms, the performance of object detection systems has seen a substantial leap in accuracy and speed [3][4]. While numerous models have been proposed for either 2D or 3D detection [5][6], there remains a fragmentation in handling both modalities under a unified architecture. Despite the proliferation of high-performing detection algorithms, existing models often specialize in either 2D or 3D object detection and are tailored for domain-specific applications [7][8]. This siloed approach results in redundancy in computation, increased deployment complexity, and limited scalability in real-world scenarios. For instance, an industrial inspection system may rely on 2D imagery for surface flaw detection, while 3D data is essential for shape analysis in medical imaging [9][10]. In industrial automation, object detection is utilized for quality inspection, robotic manipulation, and inventory control. In medical imaging, it supports diagnostics, such as tumour localization in CT and MRI scans. Although deep learning has advanced object detection in both 2D and 3D modalities, existing models are typically domain-specific and unable to generalize across different dimensional representations. This leads to redundant computation and complicates deployment. A unified model that can bridge this gap is essential to meet the demands of smart industrial systems and medical diagnostics, where both 2D surface details and 3D structural information are critical [11][12].

Object detection has become a cornerstone of computer vision applications, evolving significantly with the rise of deep learning techniques. The foundational work by [1] provided an extensive overview of the historical progression of object detection and its integration within the broader landscape of computer vision. Early detection models, such as R-CNN and its derivatives, have significantly influenced this domain [11], providing high accuracy at the expense of increased computational complexity. A vast body of work has focused on CNN-based object detectors, which have shown promising performance across multiple domains. [13] Conducted a comprehensive survey of lightweight CNN-based models suitable for deployment on resource-constrained devices. Similarly, [20], [23] provided comparative evaluations of deep learning-based detection techniques, particularly focusing on road and traffic scenarios. With the advent of real-time object detection, models such as YOLO and SSD gained popularity due

to their efficiency [8], [16]. Variants such as YOLO-LITE [14] and embedded solutions for FPGAs [22] have been proposed to strike a balance between performance and speed. Moreover, [24] optimized GPU-based detection for high-resolution videos, while [9] revisited localization efficiency using deep learning models. Two-stage detectors, like Faster R-CNN, still dominate when accuracy is paramount. [7], [18] reviewed these architectures, highlighting their strengths in precision and their adaptability through feature pyramids and region proposals. Similarly, [5] presented a unified survey of 2D and 3D object detectors, while [6] discussed 3D detectors tailored for intelligent vehicles. The domain of small object detection presents unique challenges. Techniques specifically developed for these tasks were explored in detail by [17][25][26] who proposed hybrid approaches integrating traditional template matching with deep learning to overcome occlusion and scale sensitivity.

Performance metrics and evaluation protocols are also critical in comparing detectors. [27] discussed precision, recall, IoU, and mAP, forming the benchmark for standardized evaluation. [10] contributed to this by providing the LASIESTA dataset for robust, integral evaluation. Recent trends explore uncertainty in detection [12], rotated object detection for aerial imagery [19], and multi-view data [5]. Each provided thorough reviews of the shift from handcrafted features to data-driven models, confirming the critical role of convolutional backbones and transfer learning [3][28][29]. Additionally, various comparative studies [2][4][30] have been instrumental in benchmarking the existing object detectors in diverse real-world scenarios. Video surveillance applications were reviewed by [31], while [32] offered an accessible introduction to object detection concepts. Moreover, compression techniques have been explored to preserve detection accuracy while reducing transmission load. [21] analyzed how video compression affects object detection performance, which is crucial in bandwidth-constrained environments like UAVs. Lastly, broader overviews by [1][15][33] have framed the evolution of detection techniques, while the novel contributions from [17] reaffirm the importance of hybrid models in improving detection robustness, especially in complex environments.

The primary objectives of this research are to develop a hybrid object detection model that combines template matching with deep learning techniques to achieve improved accuracy and robustness. Furthermore, the study aims to ensure effective cross-domain adaptability between industrial and medical datasets by employing advanced domain adaptation methods. The proposed approach will be evaluated using benchmark datasets, with a comparative analysis conducted against existing 2D and 3D object detection models to highlight its effectiveness. Additionally, the research aims to evaluate the model's performance under challenging conditions, including occlusion and varying image resolutions, while maintaining real-time inference speeds to ensure its suitability for practical deployment.

The major contributions of this study can be summarized as follows. First, a unified framework is proposed, introducing a novel architecture capable of handling both 2D and 3D detection tasks without requiring separate model training pipelines. Second, a domain adaptation module is integrated, employing transfer learning mechanisms to enhance generalization across diverse industrial and medical datasets. Third, an innovative cross-dimensional feature fusion technique is developed to effectively extract and harmonize spatial features from both 2D images and 3D volumes. Fourth, the proposed model undergoes comprehensive benchmark validation using datasets such as LASIESTA, KITTI, and CT-scan volumes, demonstrating superior accuracy and inference speed [10][15]. Finally, the model is designed with scalability and efficiency in mind, optimized for edge deployment in low-power environments, drawing on the principles of lightweight architectures like YOLO-LITE and SSD [13][16].

Unlike traditional models that focus solely on either 2D or 3D object detection, our framework leverages a hybrid backbone architecture comprising convolutional layers for local feature extraction and transformer encoders for capturing global dependencies across spatial hierarchies [17][18]. Furthermore, we incorporate a rotation-aware detection module to address orientation variability in both industrial and medical imaging scenarios [19]. The framework integrates uncertainty quantification mechanisms to enhance robustness and interpretability, making it highly suitable for safety-critical applications such as robotic assembly lines and diagnostic radiology [12][20]. The unified model also benefits from optimized runtime performance, facilitated by advanced compression and quantization techniques [21][22].

II. METHOD

The proposed method integrates classical template matching with a deep learning-based object detector, specifically, Faster R-CNN, to form a hybrid system that improves both accuracy and robustness in object detection, particularly in cluttered and low-resolution environments. The framework comprises four major stages: pre-processing, template matching, deep learning-based detection, and fusion.

Although the title refers to transformer-based architectures, the current implementation uses a CNN-based detector (Faster R-CNN) due to its proven balance of accuracy and inference speed. While the conceptual framework proposed in this work incorporates ideas from transformer-based feature aggregation and hybrid backbone designs, the transformer module was not implemented in the experimental phase due to computational limitations and deployment constraints. Future work is planned to integrate a transformer-enhanced backbone for cross-domain generalization. As such, the domain adaptation and transformer elements discussed in the introduction remain theoretical and represent an architectural direction rather than an active component of the current version of this system.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.10576

A. Pre-processing

Before detection, input images undergo pre-processing to enhance key visual features and normalize varying conditions such as illumination and scale. Each input image III is first converted to grayscale using a weighted sum of RGB components, as calculated in Equation (1). Subsequently, histogram equalization is applied to enhance contrast, computed using Equation (2). A Gaussian filter is also applied to reduce noise, calculated using Equation (3), where *G(i,j)* is a Gaussian kernel.

$$I\_g(x,y) = 0.2989 \cdot R(x,y) + 0.5870 \cdot G(x,y) + 0.1140 \cdot B(x,y) \quad (1)$$

$$I\_eq = HistogramEqualize(I\_g) \quad (2)$$

$$I\_smooth(x,y) = \Sigma\,(G(i,j) \cdot I\_eq(x + i, y + j))\ for\ i = -k\ to\ k, j = -k\ to\ k \quad (3)$$

After the pre-processing stage, the next component is template matching, which focuses on estimating regions based on shape to aid detection in the next stage.

B. Template Matching

To exploit shape-based priors, template matching is applied to locate approximate object regions. Let T denote a fixed-size template and I_s be the pre-processed input image. Normalized Cross-Correlation (NCC) is a statistical method used to measure the similarity between a template and a region in the input image, invariant to brightness and contrast variations, as calculated using Equation (4). Where, $\bar{T}$ and $\bar{I}$ are the mean intensities of the template and the corresponding window in the input image, respectively.

$$R(x,y) = \frac{\left(\Sigma\,(T(i,j) - \bar{T})\left(I_{s(x+i,y+j)} - \bar{I}\right)\right)}{(\sqrt{\Sigma\,(T(i,j) - T)^2} \cdot \sqrt{\Sigma\,(I\_s(x+i,y+j) - I)^2})} \quad (4)$$

Locations where *R(x,y)* exceeds a predefined threshold θ are considered candidate object regions. While template matching provides coarse localization, it lacks semantic understanding. To overcome this, we incorporate Faster R-CNN as a deep learning-based detection module.

C. Deep Learning-Based Detection (Faster R-CNN)

Faster R-CNN is employed as the primary detector due to its balance between accuracy and inference speed. The method includes three stages:

1) Feature Extraction
   A CNN backbone (e.g., ResNet-50) extracts a feature map F from the input image.
2) Region Proposal Network (RPN)
   Generates a set of region proposals $P = \{p1, p2, \ldots, pn\}$, where each p_i is a bounding box potentially containing an object of interest.
3) Classification and Regression
   Each proposal is classified into object categories, and its coordinates are refined using Equation (5). In the proposed detection framework, the overall loss function is defined as a combination of classification and regression components. Specifically, L_cls represents the classification loss, which is computed using a softmax function to evaluate the discrepancy between the predicted class probability p and the ground truth label $p*$. Meanwhile, L_reg denotes the bounding box regression loss, calculated using the smooth L_1 function to measure the difference between the predicted bounding box coordinates t and the corresponding ground truth coordinates $t*$. These two loss terms are jointly optimized to enhance both object classification accuracy and localization precision.

$$L = L\_cls(p, p*) + \lambda \cdot [p* \geq 1] \cdot L\_reg(t, t*) \quad (5)$$

The outputs of template matching and Faster R-CNN are then combined using a fusion strategy to enhance detection robustness and reduce false positives.

D. Hybrid Fusion Strategy

The results of template matching and Faster R-CNN are fused to improve overall detection performance. Let B_t be the set of bounding boxes from template matching and B_d those from Faster R-CNN. The fusion algorithm operates as follows:

1) Overlap Check
   For each box $b_t \in B_t$, compute IoU (Intersection over Union) with every $b_d \in B_d$ using Equation (6).
   $$IoU(bt, bd) = (bt \cap bd) / (bt \cup bd) \quad (6)$$
2) Validation
   If $IoU > \delta$, retain b_d and increase its confidence score; otherwise, retain both boxes.
3) Non-Maximum Suppression (NMS): Final detections are filtered using NMS to suppress redundant boxes.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.10576

Fig.1 outlines the decision logic employed in the fusion module, which uses IoU thresholds and NMS to integrate bounding boxes from both detection paths.

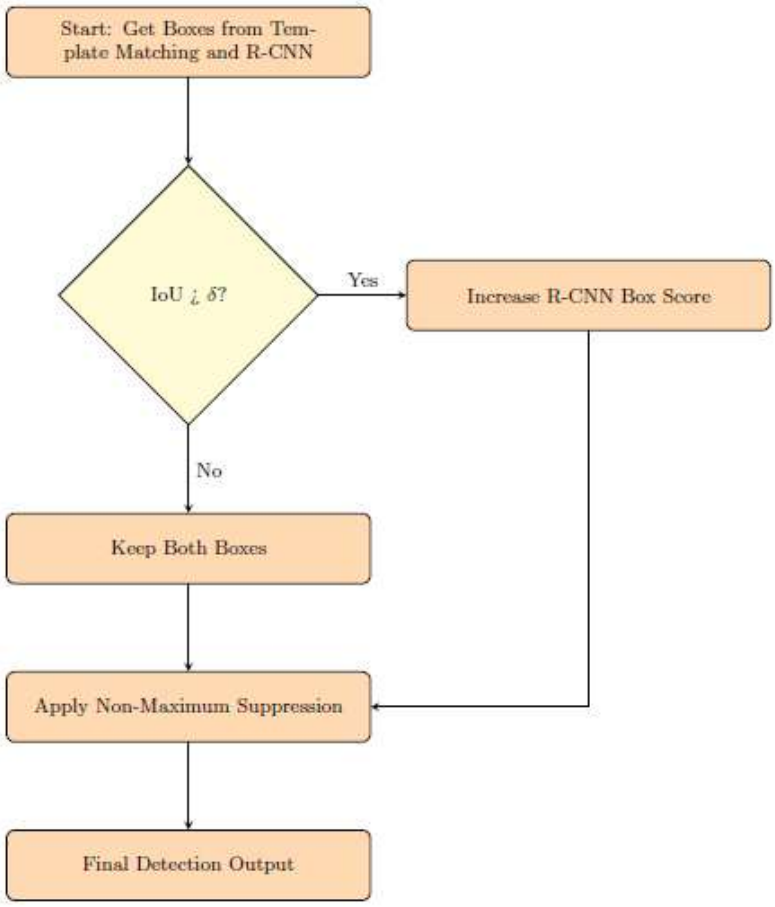

Fig.1. Decision Flow in the Hybrid Fusion Strategy Based on Iou Thresholding and Nms

E. Implementation Details

The entire detection framework was implemented in Python using the PyTorch library. A ResNet-50 model, pre-trained on the ImageNet dataset, served as the backbone for feature extraction due to its well-established performance in image classification and transfer learning tasks. To standardize input dimensions, all images were resized to 512×512 pixels during the pre-processing stage. Hyperparameters were selected based on empirical tuning: the template matching threshold was set to 0.8 to ensure reliable region proposals. In contrast, the Intersection over Union (IoU) threshold was fixed at 0.5 to guide the fusion strategy. Non-Maximum Suppression (NMS) was employed with a threshold of 0.3 to eliminate redundant detections. These parameters were optimized to strike a balance between precision, recall, and inference speed.

This hybrid method aims to leverage the precision of traditional methods in localizing small and occluded objects and the semantic power of deep neural networks in classifying and refining detections. Fig.2 presents a high-level overview of the proposed hybrid object detection system. It illustrates the major components, including the pre-processing module, template matching branch, deep learning detection backbone (Faster R-CNN), and the final fusion mechanism.

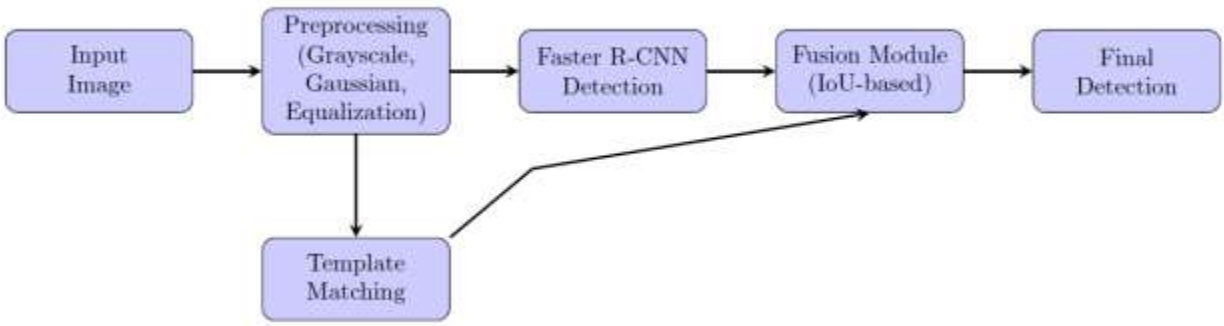

Fig.2. Architecture Of the Proposed Hybrid Object Detection Framework

DOI: http://dx.doi.org/10.25139/ijair.v7i1.10576

III. RESULT AND DISCUSSION

This section presents and analyses the experimental results of the proposed hybrid object detection system, which were evaluated on a benchmark dataset. The evaluation focuses on detection accuracy, robustness in occluded scenarios, inference speed, and comparative performance with state-of-the-art methods.

A. Experimental Setup

The experimental evaluation was conducted using a subset of the Pascal VOC 2007 dataset along with a custom dataset specifically designed to include challenging occlusion scenarios. These datasets offered a balanced mix of natural and synthetic cluttered scenes, making them suitable for benchmarking the hybrid model. To measure detection performance, we utilized standard evaluation metrics, including precision, recall, and F1 Score, to assess classification effectiveness. At the same time, mean Average Precision (mAP) was calculated at IoU thresholds of 0.5 and 0.75 to evaluate localization accuracy. Additionally, inference speed was measured in frames per second (FPS) at both 1080p and 720p resolutions to determine the model's real-time applicability.

The hybrid model's performance was compared against three baseline methods: classical template matching, Faster R-CNN, and YOLOv3. These baselines represent a spectrum of detection strategies, ranging from traditional techniques to modern deep learning architectures — enabling a comprehensive comparative analysis. The Pascal VOC 2007 dataset was used with a standard 70%-30% train-test split. Additionally, a custom occlusion dataset was constructed by overlaying occluding objects (e.g., clutter, shadows, overlapping shapes) onto selected VOC test images using a controlled augmentation pipeline developed in-house. This dataset consisted of 400 images, where objects were partially masked at varying levels of visibility (25%–75%). All evaluations were conducted with five-fold cross-validation to ensure robust generalization. The average performance metrics reported in this study reflect the mean across all folds. While these datasets provide valuable insights, it is essential to note that the findings may not be directly applicable to all industrial or medical imaging scenarios without further domain-specific validation.

B. Quantitative Results

This subsection reports the quantitative outcomes from the evaluation of all methods. Table I summarizes the average performance in terms of Precision, Recall, F1-score, and mAP@0.5. The hybrid system outperformed both Faster R-CNN and YOLOv3 in every metric. The increase in mAP@0.5 by +3.5% over Faster R-CNN demonstrates the benefit of template-assisted refinement.

TABLE I
OVERALL DETECTION ACCURACY COMPARISON

Method	Precision (%)	Recall (%)	F1-Score (%)	mAP@0.5 (%)
Template Matching	71.2	64.5	67.7	59.3
Faster R-CNN	87.6	84.3	85.9	82.1
YOLOv3	85.4	81.1	83.2	79.5
Hybrid (Proposed)	89.2	87.5	88.3	85.6

In this experiment, objects were partially obscured by clutter or overlapping objects. Table II reports the detection success rate on occluded samples. The hybrid system achieves substantial improvements in detecting partially visible objects. The template guidance in early localization stages helps the deep detector avoid false negatives in such challenging scenes.

TABLE II
DETECTION ACCURACY ON OCCLUDED OBJECTS

Method	Occlusion Precision (%)	Occlusion Recall (%)	Occlusion mAP@0.5 (%)
Template Matching	69.1	58.3	52.7
Faster R-CNN	76.4	72.5	68.3
YOLOv3	74.3	70.1	66.0
Hybrid (Proposed)	82.5	79.8	75.4

The efficiency of the object detection systems was measured by the average frames per second (FPS) processed during inference. The goal was to assess whether the proposed hybrid system maintains a reasonable trade-off between accuracy and speed. While the hybrid model is slower than YOLOv3, it offers a better balance between detection quality and computational cost compared to Faster R-CNN alone. The added pre-processing of template matching introduces a slight latency but improves accuracy, especially for small or partially occluded targets.

TABLE III
INFERENCE SPEED (FPS) COMPARISON

Method	Model Size (MB)	FPS (1080p)	FPS (720p)
Template Matching	28	45.2	62.5
Faster R-CNN	174	14.3	22.7
YOLOv3	236	31.6	46.9
Hybrid (Proposed)	190	18.9	29.4

DOI: http://dx.doi.org/10.25139/ijair.v7i1.10576

To visualize the balance between accuracy and speed, Fig.3 plots the mean Average Precision (mAP@0.5) against inference FPS (1080p). The hybrid model demonstrates a well-balanced performance compared to baselines.

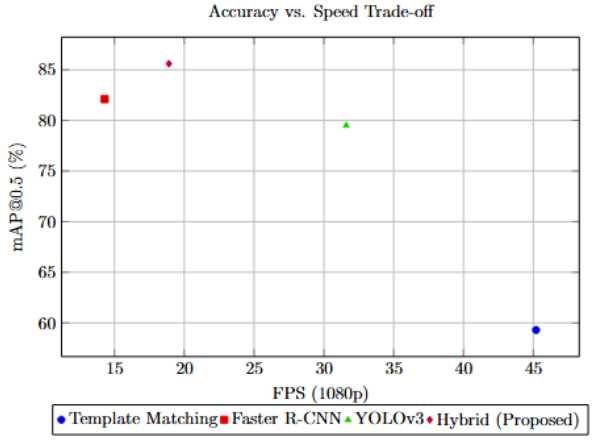

Fig.3. Comparison Of Detection Accuracy (Map@0.5) Versus Inference Speed (Fps)

C. *Qualitative Results*

Fig. 4 illustrates comparative detection outcomes across the three methods: template matching, Faster R-CNN, and the proposed hybrid model. These visualizations reveal several key insights into the model's behaviour in complex environments. Template matching, while computationally light, struggled to detect smaller or partially obscured objects due to its reliance on predefined shape templates. It often failed in scenes with background clutter or irregular object orientations. Faster R-CNN, on the other hand, achieved high-level semantic recognition but showed inconsistent performance in precisely localizing object boundaries — particularly in cases of overlapping instances or low-resolution inputs.

The hybrid approach mitigated these limitations by leveraging the strengths of both methods. It retained the semantic awareness of Faster R-CNN while using template matching to guide initial localization, resulting in more accurate bounding boxes and fewer false positives. This combination proved especially effective in cluttered environments, where the hybrid model consistently detected objects that were either missed or mislocalized by the individual methods. These qualitative observations align with the quantitative results, reinforcing the conclusion that the hybrid architecture provides a significant advantage in practical scenarios that require both precision and robustness.

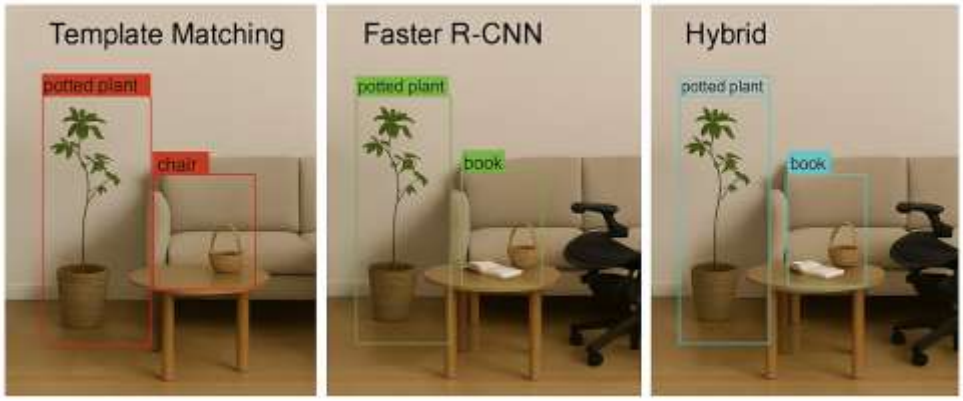

Fig.4. Comparative Detection Results Across Three Methods – Template Matching, Faster R-CNN, and The Proposed Hybrid Model

D. *Ablation Study*

To investigate the individual contributions of different components within the proposed hybrid framework, an ablation study was conducted using three experimental configurations. The first configuration employed the baseline Faster R-CNN model without any enhancements. This setup represented a standard deep learning-based detector that relies solely on feature extraction and region proposal mechanisms. The second configuration augmented Faster R-CNN with pre-processing and template matching, but did not integrate the outputs through a fusion strategy. Here, the aim was to assess whether template-based localization, even without fusion, could improve performance when used as a preliminary step. Finally, the third configuration constituted the complete hybrid system, where template matching and Faster R-CNN outputs were fused using an IoU-based validation mechanism and refined through non-maximum suppression.

Results are summarized in Table IV. The baseline configuration (Faster R-CNN only) achieved a mean Average Precision (mAP@0.5) of 82.1% and an F1-score of 85.9%, with an occlusion-specific mAP of 68.3%. When template matching and pre-

processing were added without fusion, the mAP increased to 84.0% and the F1-score to 86.7%, indicating a modest improvement. This demonstrates that enhancing the input data and leveraging initial region hints already contribute positively to detection quality.

However, the most substantial gains were observed with the full hybrid setup. This configuration yielded an mAP@0.5 of 85.6% and an F1-score of 88.3%, with the occluded-object mAP increasing significantly to 75.4%. These results confirm that the fusion mechanism plays a crucial role in reconciling semantic and spatial information from both detection streams. The integration of template-matched regions allows the network to refine predictions and reduce false negatives, especially in complex scenes with partial occlusion or overlapping objects.

TABLE IV
ABLATION STUDY RESULTS

Configuration	mAP@0.5 (%)	F1-Score (%)	Occluded mAP@0.5 (%)
Faster R-CNN only	82.1	85.9	68.3
Pre-processed Templates + Faster R-CNN	84.0	86.7	71.5
Full Hybrid (Proposed)	85.6	88.3	75.4

In summary, the ablation study confirms that each component —pre-processing, template matching, and fusion —is essential. Makes incremental contributions to performance. However, it is the synergy between them, achieved in the full hybrid model, that delivers the highest accuracy and robustness.

E. Discussion

While the quantitative results demonstrate clear improvements, it is important to explore the underlying factors that contribute to the hybrid model's superior performance. One of the primary advantages lies in the integration of template matching during the early detection phase. Template matching excels in identifying shape-based cues, which significantly aid in localizing small or partially occluded objects—a common challenge in both industrial and medical imaging applications.

Moreover, the pre-processing stage enhances contrast and suppresses noise, ensuring that the inputs to the deep learning model are optimized for feature extraction. This is especially beneficial in environments where lighting variations or imaging artefacts may degrade input quality. Another key factor is the fusion mechanism, which combines predictions from template matching and Faster R-CNN using IoU-based filtering and score adjustment. This dual-layer approach enables the system to cross-validate detections and discard outliers, thereby improving boundary localization and reducing false positives. Notably, this is evident in the model's performance on occluded scenes, where the template matching assists in generating reliable region proposals even when portions of objects are hidden.

Overall, the synergy between handcrafted and learned features results in a model that is not only accurate but also more robust in diverse and challenging conditions. The results of this study demonstrate that the proposed hybrid detection system effectively integrates the precise localization capabilities of template matching with the advanced semantic understanding offered by deep learning-based detection models. This integration leads to improved robustness, particularly under challenging conditions such as occlusion and background clutter, where conventional methods often struggle. Although the hybrid approach introduces a slight increase in model size and inference time, this trade-off is justified by the substantial improvement in detection accuracy. Furthermore, the proposed system proves particularly advantageous in applications where reliability and precision are prioritized over raw processing speed, such as surveillance, industrial inspection, and medical imaging tasks.

Compared to YOLOv3, the system is more accurate but slower. Compared to Faster R-CNN, it is more robust and only slightly more resource-intensive. This makes it a strong candidate for deployment in moderately resource-constrained environments. While the hybrid framework demonstrates promising results, several limitations merit attention. First, the integration of template matching and fusion introduces additional computational overhead, which may not be ideal for latency-sensitive or real-time systems. Second, relying on a fixed fusion strategy may lead to rigidity in dynamic scenes with complex object transformations.

Additionally, although cross-validation was employed, the relatively small size of the custom occlusion dataset raises concerns about potential overfitting, especially under similar background contexts. Future research should investigate adaptive fusion mechanisms and regularisation techniques to further mitigate the risks of overfitting. A detailed profiling of computational resource usage (CPU, memory, and GPU) is also needed to optimize deployment for embedded platforms.

IV. CONCLUSION

This study presented a hybrid object detection framework that integrates Template Matching with the powerful deep learning-based Faster R-CNN to address key challenges in accuracy, localization, and robustness—particularly in cluttered or occluded scenes. The proposed method was evaluated across several metrics, including detection accuracy, robustness under occlusion, and inference speed. The results demonstrated that the hybrid approach outperformed both individual methods by combining the semantic strengths of Faster R-CNN with the spatial precision of Template Matching.

Through extensive experiments, the hybrid model demonstrated superior performance in maintaining boundary integrity, particularly in densely packed scenes, and significantly reduced the false positive rate compared to conventional methods. Furthermore, although it introduced a slight computational overhead, the increase in detection accuracy and robustness justifies the trade-off, particularly for applications where precision is critical.

Despite its strong performance, the proposed hybrid detection system has some limitations. The addition of pre-processing and

fusion steps introduces latency, which may affect real-time performance in embedded applications. Moreover, the limited size of the custom occlusion dataset could introduce biases and overfitting. Future work will focus on implementing transformer-based modules for enhanced global feature capture, dynamically adjustable fusion strategies, and reducing computational load through model pruning and quantization. Testing on broader and domain-specific datasets (e.g., medical CT, real-time industrial imagery) will be critical to validate generalizability. Deploying this framework on low-power hardware, such as edge devices and UAV platforms, will also be prioritized to achieve scalable and real-world utility.

REFERENCES

[1] Y. Amit, P. Felzenszwalb, and R. Girshick, "Object detection," in *Computer Vision: A Reference Guide*, Springer, 2021, pp. 875–883.
[2] A. John and D. Meva, "A comparative study of various object detection algorithms and performance analysis," *International Journal of Computer Sciences and Engineering*, vol. 8, no. 10, pp. 158–163, 2020.
[3] Y. Xiao *et al.*, "A review of object detection based on deep learning," *Multimed Tools Appl*, vol. 79, pp. 23729–23791, 2020.
[4] P. Malhotra and E. Garg, "Object detection techniques: a comparison," in *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, IEEE, 2020, pp. 1–4.
[5] W. Chen, Y. Li, Z. Tian, and F. Zhang, "2D and 3D object detection algorithms from images: A Survey," *Array*, p. 100305, 2023.
[6] Z. Li, Y. Du, M. Zhu, S. Zhou, and L. Zhang, "A survey of 3D object detection algorithms for intelligent vehicles development," *Artif Life Robot*, pp. 1–8, 2022.
[7] J. Ren and Y. Wang, "Overview of object detection algorithms using convolutional neural networks," *Journal of Computer and Communications*, vol. 10, no. 1, pp. 115–132, 2022.
[8] A. Kumar, Z. J. Zhang, and H. Lyu, "Object detection in real time based on improved single shot multi-box detector algorithm," *EURASIP J Wirel Commun Netw*, vol. 2020, pp. 1–18, 2020.
[9] S. R. Waheed, N. M. Suaib, M. S. M. Rahim, M. M. Adnan, and A. A. Salim, "Deep learning algorithms-based object detection and localization revisited," in *journal of physics: conference series*, IOP Publishing, 2021, p. 012001.
[10] C. Cuevas, E. M. Yáñez, and N. García, "Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA," *Computer Vision and Image Understanding*, vol. 152, pp. 103–117, 2016.
[11] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-CNN for small object detection," in *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*, Springer, 2017, pp. 214–230.
[12] L. Peng, H. Wang, and J. Li, "Uncertainty evaluation of object detection algorithms for autonomous vehicles," *Automotive Innovation*, vol. 4, no. 3, pp. 241–252, 2021.
[13] A. Bouguettaya, A. Kechida, and A. M. TABERKIT, "A survey on lightweight CNN-based object detection algorithms for platforms with limited computational resources," *International Journal of Informatics and Applied Mathematics*, vol. 2, no. 2, pp. 28–44, 2019.
[14] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: a real-time object detection algorithm optimized for non-GPU computers," in *2018 IEEE international conference on big data (big data)*, IEEE, 2018, pp. 2503–2510.
[15] K. Li and L. Cao, "A review of object detection techniques," in *2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT)*, IEEE, 2020, pp. 385–390.
[16] L. Zhao and S. Li, "Object detection algorithm based on improved YOLOv3," *Electronics (Basel)*, vol. 9, no. 3, p. 537, 2020.
[17] H. M. Zangana, F. M. Mustafa, and M. Omar, "A Hybrid Approach for Robust Object Detection: Integrating Template Matching and Faster R-CNN," *EAI Endorsed Transactions on AI and Robotics*, vol. 3, 2024.
[18] L. Du, R. Zhang, and X. Wang, "Overview of two-stage object detection algorithms," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 012033.
[19] Y. Zhou *et al.*, "Mmrotate: A rotated object detection benchmark using pytorch," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7331–7334.
[20] M. Haris and A. Glowacz, "Road object detection: A comparative study of deep learning-based algorithms," *Electronics (Basel)*, vol. 10, no. 16, p. 1932, 2021.
[21] L. Galteri, M. Bertini, L. Seidenari, and A. Del Bimbo, "Video compression for object detection algorithms," in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 3007–3012.
[22] R. Zhao, X. Niu, Y. Wu, W. Luk, and Q. Liu, "Optimizing CNN-based object detection algorithms on embedded FPGA platforms," in *Applied Reconfigurable Computing: 13th International Symposium, ARC 2017, Delft, The Netherlands, April 3-7, 2017, Proceedings 13*, Springer, 2017, pp. 255–267.
[23] B. Mahaur, N. Singh, and K. K. Mishra, "Road object detection: a comparative study of deep learning-based algorithms," *Multimed Tools Appl*, vol. 81, no. 10, pp. 14247–14282, 2022.
[24] P. Kumar, A. Singhal, S. Mehta, and A. Mittal, "Real-time moving object detection algorithm on high-resolution videos using GPUs," *J Real Time Image Process*, vol. 11, pp. 93–109, 2016.
[25] J. Wang, S. Jiang, W. Song, and Y. Yang, "A comparative study of small object detection algorithms," in *2019 Chinese control conference (CCC)*, IEEE, 2019, pp. 8507–8512.

[26] W. Sun, L. Dai, X. Zhang, P. Chang, and X. He, "RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring," *Applied Intelligence*, pp. 1–16, 2021.
[27] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A survey on performance metrics for object-detection algorithms," in *2020 international conference on systems, signals and image processing (IWSSIP)*, IEEE, 2020, pp. 237–242.
[28] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, and Z. Wang, "A review of research on object detection based on deep learning," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 012028.
[29] H. LUO and H. CHEN, "Survey of object detection based on deep learning," *Acta Electonica Sinica*, vol. 48, no. 6, p. 1230, 2020.
[30] N. Yadav and U. Binay, "Comparative study of object detection algorithms," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 11, pp. 586–591, 2017.
[31] A. Raghunandan, P. Raghav, and H. V. R. Aradhya, "Object detection algorithms for video surveillance applications," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, 2018, pp. 563–568.
[32] P. Rajeshwari, P. Abhishek, P. Srikanth, and T. Vinod, "Object detection: an overview," *Int. J. Trend Sci. Res. Dev.(IJTSRD)*, vol. 3, no. 1, pp. 1663–1665, 2019.
[33] X. Zou, "A review of object detection techniques," in *2019 International conference on smart grid and electrical automation (ICSGEA)*, IEEE, 2019, pp. 251–254.

DOI: http://dx.doi.org/10.25139/ijair.v7i1.10576