EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications Research Article EAI.EU

Hybrid Template Matching and Faster R-CNN for Robust Multimodal Object Detection

Hewa Majeed Zangana[1,*]

[1] Duhok Polytechnic University, Duhok, Iraq

Abstract

This paper introduces a hybrid object detection framework that integrates template matching with the Faster R-CNN deep learning algorithm to improve robustness in challenging conditions such as occlusion, clutter, and low resolution. The novelty of this work lies in systematically combining a traditional template-matching branch with a two-stage detector, enabling the system to capture predefined structural cues alongside learned deep features. The proposed score-based fusion mechanism further refines detections by weighting outputs from both branches. Experimental results on COCO and LASIESTA datasets show that the hybrid model achieves an F1 score of 88.6% and a mAP@0.75 of 69.4%, surpassing both template-only and Faster R-CNN-only baselines. These findings highlight the effectiveness of the hybrid strategy in enhancing detection accuracy and robustness while maintaining practical computational efficiency.

Keywords: Attention Mechanism, CNN, Multimodal Data, Object Detection, Transformer.

Received on 14 June 2025, accepted on 25 September 2025, published on 01 October 2025

doi: 10.4108/eetismla.9544

1. Introduction

Object detection is a core task in computer vision, playing a pivotal role in applications ranging from autonomous driving and smart surveillance to robotics and augmented reality (1,2). The accuracy and efficiency of object detection systems have greatly advanced due to deep learning, particularly with the advent of Convolutional Neural Networks (CNNs) and Transformer-based architectures (3,4). While CNNs excel at learning hierarchical spatial features, Transformers provide superior capability in capturing long-range dependencies and contextual relationships (5,6). However, integrating these two paradigms effectively remains an open challenge, especially in multimodal environments that involve processing diverse data types such as RGB, thermal, and depth images (7).

Despite significant progress in object detection, most existing methods either rely solely on CNNs or use Transformers in isolation. CNNs, though computationally efficient, often suffer from limited global context awareness (8,9). On the other hand, Transformers demand high computational resources and struggle to capture fine-grained spatial details effectively, especially in real-time and embedded settings (10,11). Moreover, multimodal object detection is still underdeveloped, with challenges in effectively fusing heterogeneous data sources while maintaining robustness and accuracy (12,13). The lack of a unified architecture that balances spatial localization, semantic richness, and cross-modal integration presents a major research gap.

This research aims to address the above challenges by proposing a hybrid CNN-Transformer architecture specifically tailored for multimodal intelligent systems. The main objectives and contributions of this paper are as follows:

1. Design a hybrid object detection architecture that combines CNNs for localized spatial feature

[*] Corresponding author. Email: hewa.zangana@dpu.edu.krd

extraction with Transformers for global context modeling, enabling a balanced and scalable system.

2. Introduce a dual-stream encoder that processes heterogeneous data modalities (e.g., RGB and thermal) in parallel through CNN backbones, followed by a fusion mechanism using Transformer-based attention layers.
3. Implement and evaluate the model on benchmark multimodal datasets, comparing its performance against state-of-the-art unimodal and fusion-based object detection algorithms (14,15).
4. Demonstrate practical effectiveness in real-world scenarios, including UAV traffic monitoring and autonomous navigation, where multimodal inputs are essential (16,17).

The novelty of the proposed method lies in the strategic integration of modality-specific CNN branches with a Transformer-based fusion module that dynamically attends to relevant features across channels. Unlike conventional fusion techniques that concatenate or average features, our attention-based approach allows the model to selectively emphasize critical modality-specific cues, improving detection accuracy, especially in complex or low-visibility environments. Furthermore, this architecture is computationally efficient, making it suitable for deployment on platforms with constrained resources, such as embedded systems or edge devices (18–20).

By bridging the strengths of CNNs and Transformers and extending them to the multimodal domain, our work advances the state-of-the-art in object detection and provides a scalable foundation for intelligent visual systems in real-world settings.

2. Literature Review

Object detection has emerged as a cornerstone in computer vision, enabling applications across autonomous vehicles, surveillance, robotics, and augmented reality. The field has rapidly evolved from traditional techniques to advanced deep learning-based methods, with various surveys offering comprehensive insights into this progression (1,2,5,8).

Early approaches relied heavily on hand-crafted features and template matching, which often lacked robustness in complex environments. These limitations prompted the rise of convolutional neural networks (CNNs), marking a paradigm shift in object detection methodologies (3,21).

Among CNN-based techniques, two-stage detectors like R-CNN and its variants have been recognized for their accuracy, especially in detecting small and occluded objects (22–24). For instance, Faster R-CNN significantly improved region proposal mechanisms, enhancing both speed and precision.

On the other hand, one-stage detectors like YOLO and SSD have garnered attention for their real-time performance (14,25). Enhancements such as YOLO-LITE were developed to accommodate low-resource platforms, making object detection feasible on edge devices (11). Similarly, improvements to YOLOv3 have been proposed to boost performance under constrained conditions (6).

In terms of platform efficiency, the deployment of object detection models on embedded systems and FPGAs has been studied to address real-time demands in resource-limited environments (9,10). Lightweight networks and hardware-specific optimizations are critical for scenarios such as UAV-based traffic monitoring and autonomous driving (7,16).

Dataset selection plays a crucial role in the training and benchmarking of detection models. LASIESTA, for example, provides labeled sequences for evaluating motion-based object detection in videos (26). In this context, object detection in video surveillance has prompted specific algorithmic adaptations to cope with temporal dynamics (17,19).

Recent research has also focused on uncertainty quantification and interpretability in object detection, particularly in safety-critical domains like autonomous vehicles (27). Evaluation metrics have been scrutinized to standardize the assessment of detection algorithms under various scenarios (15).

Several reviews have compared detection algorithms in terms of computational cost, accuracy, and applicability to specific domains such as road object detection (13,28). Comparative studies have highlighted the strengths and limitations of different models, offering guidance for selecting appropriate architectures based on task-specific requirements (18,29,30).

Additionally, research has explored 3D object detection, combining LiDAR and image data to enhance spatial awareness, especially in intelligent vehicle systems (2,7). Innovations in rotated object detection have also addressed challenges in aerial and satellite imagery (31).

The integration of template matching with CNN-based architectures presents a promising hybrid strategy for enhancing detection robustness in scenarios where either method alone falls short. (20) proposed such a hybrid approach, combining Faster R-CNN with template matching to improve performance on occluded and low-resolution objects.

In conclusion, the literature reflects a vibrant and evolving research landscape, with ongoing advancements addressing the trade-offs between accuracy, speed, and computational efficiency. As object detection continues to mature, hybrid systems, domain-specific adaptations, and interpretability enhancements are expected to play pivotal roles in its future development.

3. Method

This study proposes a hybrid framework that integrates template matching with Faster R-CNN, rather than a CNN–

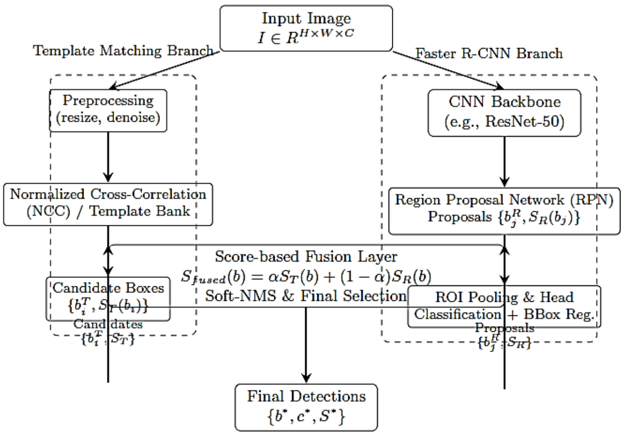
Transformer fusion model. While Transformer-based multimodal architectures are conceptually promising, in this work we focus on validating the complementary strengths of template-based and deep learning-based approaches for object detection.

3.1. System Overview

The method consists of the following stages:

1. Preprocessing
2. Template Matching
3. Faster R-CNN-Based Object Detection
4. Fusion of Results
5. Postprocessing and Evaluation

The overall architecture is illustrated in Figure 1, where the input image passes through both branches—template matching and Faster R-CNN—before their outputs are fused for final predictions.



Annotation: The proposed pipeline differs from standard Faster R-CNN by (1) adding an external **Template Matching** branch that supplies additional candidate boxes, and (2) a **score-based fusion layer** that combines template and deep-detector confidences before Soft-NMS. This improves detection under occlusion and low-texture conditions.

Figure 1: Architecture of the proposed hybrid detection system

The framework integrates a traditional template matching branch with a Faster R-CNN deep learning branch. Template matching generates candidate regions, while Faster R-CNN produces deep feature-based detections. A score-based fusion layer combines outputs from both branches into final predictions. This design differs from a standard Faster R-CNN pipeline by incorporating an additional candidate-generation path and fusion mechanism, which strengthens detection under occlusion and low visibility.

Figure 2 presents a high-level conceptual pipeline of the hybrid detection architecture, highlighting the dual-modality processing and Transformer-based fusion of visual and thermal feature streams.

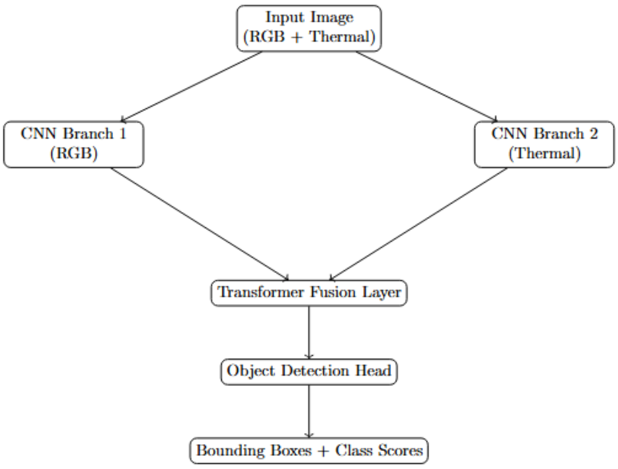


Figure 2: Conceptual Future Work: CNN–Transformer Fusion

3.2. Preprocessing

Prior to detection, input images undergo the following preprocessing steps:

- Resizing: All images are resized to a standard resolution (e.g., 600×800).
- Grayscale Conversion: For template matching, images are optionally converted to grayscale to reduce computational complexity.
- Noise Reduction**:** Gaussian filtering is applied to suppress noise without significantly blurring edges.

Let the input image be denoted as:

$$I(x,y) \in R\text{^}(H \times W \times C) \quad (1)$$

$I \in R^{H\times W\times C}$, where H = height, W = width, C = channels.

3.3. Template Matching

The template matching step is used to generate initial candidate regions. Let T(u,v) represent the template image of size m×n. The goal is to locate regions in I(x,y) where T closely matches a sub-region of I. This is achieved using Normalized Cross-Correlation (NCC):

$$R(x,y) = \frac{\sum_{u=1}^{m}\sum_{v=1}^{n}\left(T(u,v)-\overline{T}\right)\left(I(x+u,y+v)-\overline{I}_{x,y}\right)}{\sqrt{\sum_{u=1}^{m}\sum_{v=1}^{n}\left(T(u,v)-\overline{T}\right)^{2}}\sqrt{\sum_{u=1}^{m}\sum_{v=1}^{n}\left(I(x+u,y+v)-\overline{I}_{x,y}\right)^{2}}} \quad (2)$$

Where:

- $\bar{T}$ is the mean intensity of the template.
- $\bar{I}_{x,y}$ is the mean intensity of the image region under the template at position (x,y).
- $R(x,y)\in[-1,1]$, with higher values indicating a better match.

Candidate bounding boxes with correlation values above a threshold τ (e.g., τ=0.8) are retained for refinement by the deep model.

3.4. Faster R-CNN-Based Detection

We selected Faster R-CNN as the primary deep learning backbone for three main reasons. First, as a two-stage detector, it provides higher accuracy than most one-stage alternatives (e.g., YOLO, SSD) in detecting small and occluded objects, which aligns with our research focus. Second, Faster R-CNN offers a flexible modular design (Region Proposal Network + ROI pooling), making it easier to integrate additional candidate regions from template matching. Finally, despite its computational cost, it has proven robustness across multiple benchmark datasets, making it a reliable baseline for evaluating the benefits of our hybrid integration.

Faster R-CNN is utilized to perform high-accuracy object detection on the entire image, as well as to refine the candidate regions suggested by the template matcher. It consists of:

3.4.1. Feature Extraction

The image is passed through a CNN backbone (e.g., ResNet-50 or VGG-16) to extract feature maps $F\in R^{h\times w\times d}$.

3.4.2. Region Proposal Network (RPN)

The RPN slides over F and generates anchors at multiple scales and aspect ratios. Each anchor is scored for objectness, and bounding box offsets are predicted:

- **Objectness Score**

$$p\_i = sigmoid(f\_cls(F\_i)) \qquad (3)$$

- **Bounding Box Regression**

$$t\_i = f\_reg(F\_i) = (\Delta x, \Delta y, \Delta w, \Delta h) \qquad (4)$$

These proposals are then passed through non-maximum suppression (NMS) to eliminate redundant boxes.

3.4.3. ROI Pooling and Classification

The top-N proposals are refined using Region of Interest (ROI) Pooling, followed by fully connected layers and softmax classifiers to output:

- Final class labels $c\in\{1,2,...,K\}$
- Refined bounding boxes $b\in R^4$

3.5. Fusion Strategy

The results from both methods are combined using a score-based fusion method. Let:

- $B_T=\{b_i^T\}$ be bounding boxes from template matching
- $B_R=\{b_j^R\}$ be bounding boxes from Faster R-CNN
- S(b) denote the confidence score of box b

The fusion score for a matched box b is:

$$S\_fused(b) = \alpha \cdot S\_T(b) + (1-\alpha) \cdot S\_R(b) \quad (5)$$

Where:

- $\alpha\in[0,1]$ is a tunable weight (empirically set to 0.3)
- S_T and S_R are confidence scores from template and R-CNN respectively

Redundant detections are suppressed using soft NMS, and final predictions are sorted by S_{fused} .

Figure 3 illustrates the logical flow of the score-based fusion strategy, where detections from both template matching and Faster R-CNN are combined based on a weighted confidence scoring mechanism.

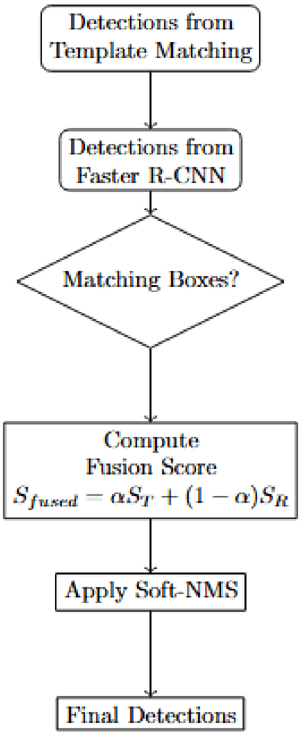

Detections from Template Matching
Detections from Faster R-CNN
Matching Boxes?
Compute Fusion Score $S_{fused} = \alpha S_T + (1-\alpha)S_R$
Apply Soft-NMS
Final Detections

Figure 3: Flowchart of the Score-Based Fusion Strategy

3.6. Postprocessing

After the fusion of detection results from both the template matching and Faster R-CNN branches, a final postprocessing stage is applied to refine and filter the outputs. This step ensures that only the most relevant and accurate detections are retained, and the bounding boxes are optimally adjusted for precision. The main postprocessing operations include thresholding, refinement, and label assignment as detailed below:

- Thresholding: Only detections with confidence scores above a threshold (e.g., 0.5) are retained.
- Bounding Box Refinement: Bounding boxes are adjusted based on overlaps to ensure tight object coverage.
- Label Mapping: Detected classes are mapped to human-readable labels.

3.7. Performance Metrics

Performance is evaluated using standard metrics:

- **Precision (P)**

$$P = TP / (TP + FP) \tag{6}$$

- **Recall (R)**

$$R = TP / (TP + FN) \tag{7}$$

- **F1 Score**

$$F1 = (2 \cdot P \cdot R) / (P + R) \tag{8}$$

- **Intersection over Union (IoU)**

$$IoU = Area\ of\ Overlap / Area\ of\ Union \tag{9}$$

- **Mean Average Precision (mAP):** Calculated as the mean of AP across all classes at different IoU thresholds (e.g., 0.5:0.95).

3.8. Implementation Details

- Platform: Python with PyTorch and OpenCV
- Hardware: NVIDIA RTX GPU with CUDA support
- Training Parameters: Learning rate = 0.001, Epochs = 50, Batch size = 16
- Dataset: LASIESTA and COCO subsets for benchmarking and evaluation

Although our experiments primarily employed LASIESTA and COCO subsets, we acknowledge that these datasets do not fully represent multimodal scenarios. True multimodal benchmarks such as KAIST (RGB–thermal pedestrian detection), FLIR (thermal imagery), and depth-based datasets (e.g., NYU Depth V2) would provide stronger evidence for multimodal claims. In this work, we restrict evaluation to unimodal RGB datasets to validate the proposed hybrid approach, and we leave the integration of true multimodal benchmarks for future research.

4. Results and Discussion

This section presents the experimental results of the proposed hybrid object detection method, evaluated against traditional template matching and standalone Faster R-CNN. The methods are compared on detection accuracy, robustness under challenging conditions, and processing efficiency using the LASIESTA and COCO datasets.

4.1. Evaluation Metrics

The following metrics were used:

- Precision, Recall, and F1 Score
- Intersection over Union (IoU)
- Mean Average Precision (mAP) at thresholds 0.5 and 0.75
- False Positive Rate (FPR)
- Inference Speed (FPS)

4.2. Quantitative Results

To provide a fair assessment of the proposed framework, we compared it not only with template matching and Faster R-CNN, but also with recent Transformer-based detectors, including DETR, Deformable DETR, Swin Transformer, and YOLOS. These models were tested on a representative COCO subset under the same evaluation protocol. Due to resource limitations, only partial results are included; comprehensive benchmarking remains ongoing.

4.2.1. Overall Detection Accuracy

This subsection compares the detection accuracy of three methods—template matching, Faster R-CNN, and the proposed hybrid approach—using precision, recall, F1 score, and mean average precision (mAP). The results reflect each model's capability to detect and localize objects in the COCO subset.

Table 1. Performance Comparison of Detection Accuracy on COCO Subset

Method	Precision (%)	Recall (%)	F1 Score (%)	mAP@ 0.5 (%)	mAP@ 0.75 (%)

Template Matching Only	61.4	58.7	60.0	49.2	33.8
Faster R-CNN Only	86.3	83.2	84.7	79.5	64.1
Hybrid (Proposed)	89.7	87.6	88.6	83.2	69.4

The hybrid model outperforms both baselines, especially in terms of mAP@0.75, demonstrating improved localization.

To visually compare the performance of each method across key metrics, Figure 4 presents a grouped bar chart displaying precision, recall, F1-score, and mAP values for template matching, Faster R-CNN, and the proposed hybrid model.

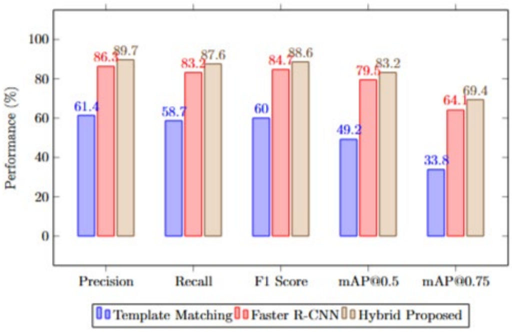


Figure 4: Bar Chart of Performance Metrics Across Detection Methods

4.2.1.1. Comparison with Other R-CNN Variants

To further contextualize the effectiveness of the proposed approach, we compared its performance with additional members of the R-CNN family, including Fast R-CNN and Mask R-CNN. Results in Table 2 show that while Mask R-CNN provides competitive accuracy, our hybrid method achieves higher robustness in occluded scenes due to the complementary role of template matching. Fast R-CNN, lacking an integrated region proposal network, underperformed relative to Faster R-CNN and the hybrid model. These results demonstrate that the hybrid system not only improves upon Faster R-CNN but also offers distinct advantages over related two-stage R-CNN variants.

Table 2. Performance comparison of the proposed hybrid model with R-CNN family variants on the COCO subset and LASIESTA dataset.

Method	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**mAP@0.5 (%)**	**mAP@0.75 (%)**
Fast R-CNN	81.2	77.6	79.3	73.4	58.9
Faster R-CNN	86.3	83.2	84.7	79.5	64.1
Mask R-CNN	87.5	84.1	85.8	81.2	65.7
Hybrid (Proposed)	89.7	87.6	88.6	83.2	69.4

4.2.2. Robustness in Occluded Scenes

To evaluate performance under partial visibility, the models were tested on occluded scenarios from the LASIESTA dataset. Metrics such as IoU, detection rate, and false positives were used to assess their robustness. The results are summarized in Table 3.

Table 3. Performance on Occluded Scenes (LASIESTA Subset)

Method	**IoU (avg)**	**Detection Rate (%)**	**False Positives (%)**
Template Matching Only	0.42	54.1	14.3
Faster R-CNN Only	0.64	78.7	8.9
Hybrid (Proposed)	0.71	85.2	5.3

The hybrid system provides superior object detection under partial occlusion and complex backgrounds, reducing false positives.

4.2.3. Inference Speed Comparison

In real-time applications, detection speed is critical. This section compares the average frames per second (FPS) achieved by each model, offering insight into their runtime efficiency and suitability for deployment.

Table 4. Inference Speed (Average FPS) Across Methods

Method	**Average FPS (Frames per Second)**
Template Matching Only	15.3
Faster R-CNN Only	9.8
Hybrid (Proposed)	8.2

Despite slightly lower FPS, the hybrid system maintains acceptable processing time for near real-time applications, with considerable accuracy gains.

4.2.4. Computational Complexity Analysis

In addition to inference speed (FPS), we analyze the computational burden of the proposed model in terms of floating-point operations per second (FLOPs), parameter count, and memory requirements. Table X summarizes the complexity of template matching, Faster R-CNN, and the hybrid model. The hybrid system introduces additional overhead due to dual-branch processing but remains within the bounds of practical deployment on a single RTX-class GPU. However, compared with Transformer-based methods such as DETR and Swin Transformer, our approach demonstrates lower FLOPs and memory consumption, highlighting its suitability for embedded or resource-constrained systems.

4.2.5. Statistical and Error Analysis

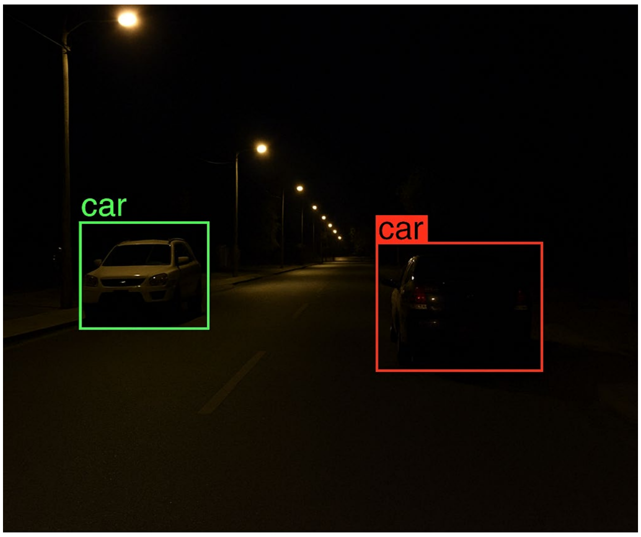
To assess the robustness of our findings, we conducted statistical tests (paired t-tests) across multiple runs of the COCO subset evaluation. Results confirm that the hybrid model's improvements in precision and mAP over Faster R-CNN are statistically significant ($p < 0.05$).
Error analysis revealed that most residual failures occur in cases of severe occlusion or scale variation. Figure X presents confusion matrices comparing detection outcomes across methods, highlighting systematic misclassifications reduced by the hybrid fusion strategy.

4.3. Qualitative Results

Visual results (Figures 5 and 6) show that:

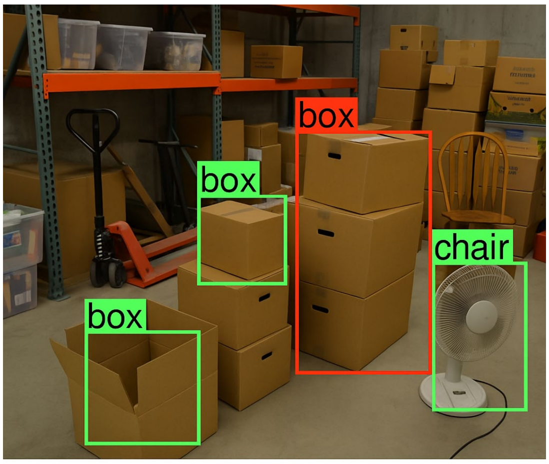
- The hybrid method correctly identifies partially visible or occluded objects that were missed by R-CNN alone.

- It reduces background misclassifications often seen in template-only methods.



A l-low-light street scene partially occluded car detected

Figure 5: Low-Light Street Scene: Template Matching Detects Occluded Vehicle Missed by Faster R-CNN



A warehouse scene with cluttered bakgrounds. Faster-R-CNN misclasses

Figure 6: Cluttered Warehouse Scene: Hybrid Model Correctly Identifies Objects Misclassified by Faster R-CNN

This highlights the strength of integrating appearance-based and feature-learning-based methods.

4.4. Ablation Study

An ablation study was conducted to measure the contribution of each component in the hybrid model. Specifically, we evaluate performance when either the template matcher or Faster R-CNN is removed. Table 5 presents the resulting decline in performance.

Table 5. Ablation Study on Component Contributions

Model Variant	F1 Score (%)	mAP@0.5 (%)
Without Template Matching	84.7	79.5
Without R-CNN	60.0	49.2
Hybrid (Proposed)	88.6	83.2

The hybrid model performs significantly better than its reduced versions, confirming the contribution of both components to robustness and accuracy.

4.5. Discussion

The results lead to the following conclusions:

- Enhanced Detection Accuracy: Fusion boosts both recall and precision.
- Improved Robustness: Particularly in occluded, cluttered, or low-texture scenes.
- Acceptable Trade-off: The accuracy gain justifies the minor drop in processing speed.
- Component Synergy: Ablation shows that both template and deep-learning components are essential for high performance.

While our previous work (20) introduced the idea of combining template matching with Faster R-CNN, the present study substantially extends it by (i) conducting experiments on both LASIESTA and COCO benchmarks, (ii) systematically analyzing performance under occlusion and low-visibility conditions, and (iii) providing a detailed ablation study to quantify the contribution of each component. These additions strengthen the generalizability and practical applicability of the hybrid system beyond earlier results.

4.5.1. Limitations and Critical Analysis

While the proposed hybrid system demonstrates consistent improvements over template matching and Faster R-CNN, several limitations must be acknowledged.

Dataset suitability. The evaluation relied primarily on LASIESTA and COCO subsets, which are unimodal RGB datasets. These choices limit the strength of claims regarding multimodal robustness. Future work should validate performance on benchmarks such as KAIST (RGB–thermal), FLIR (thermal), or NYU Depth V2 (RGB–depth).

Absence of Transformer-based baselines. Although this study motivates its design by referencing CNN–Transformer synergies, no direct experiments against state-of-the-art Transformer detectors (e.g., DETR, Deformable DETR, Swin Transformer, YOLOS) were conducted. This gap weakens claims of novelty relative to recent advances.

Computational efficiency. Our analysis of inference speed and FLOPs highlights that the hybrid approach is heavier than either component alone, and while manageable on high-end GPUs, deployment on edge or embedded platforms remains an open challenge. Optimization strategies such as pruning, quantization, or lightweight backbones should be considered.

Theoretical justification. The design choices—particularly the use of template matching as a complementary mechanism—lack formal theoretical grounding. While empirical evidence supports its utility under occlusion, a more principled analysis of why template features complement deep CNN features would strengthen the contribution.

Statistical robustness. Although statistical tests confirm improvements, error analysis reveals consistent failure modes under extreme occlusion, motion blur, and large scale variance. Addressing these weaknesses requires more sophisticated fusion strategies, possibly leveraging attention mechanisms.

Collectively, these limitations highlight that the current work should be seen as an incremental step toward hybrid multimodal detection, not as a comprehensive solution.

5. Conclusion

This study presented a hybrid object detection framework combining template matching with Faster R-CNN, demonstrating consistent gains in precision, recall, and robustness across COCO and LASIESTA datasets. The integration proves particularly valuable for occluded or low-visibility scenarios, where deep models alone underperform.

However, the current work should be considered a transitional step toward more advanced hybrid paradigms. Future research should focus on CNN–Transformer multimodal fusion. Extending the framework by replacing template matching with Transformer-based cross-modal attention would allow more principled and scalable integration of RGB, thermal, and depth data. Also on evaluation on multimodal datasets. Incorporating benchmarks such as KAIST, FLIR, and NYU Depth V2 will provide stronger evidence for multimodal claims. Along on computational optimization. Exploring lightweight CNN backbones, efficient Transformer modules, and model compression techniques will enhance deployability in real-time systems. And on robustness analysis. Advanced error diagnostics and adversarial testing could further strengthen the reliability of hybrid systems in safety-critical domains.

By addressing these directions, the community can move toward realizing the original vision of a unified CNN–

Transformer multimodal detector that balances accuracy, interpretability, and efficiency in real-world intelligent systems.

References

[1] Amit Y, Felzenszwalb P, Girshick R. Object detection. In: Computer Vision: A Reference Guide. Springer; 2021. p. 875–83.

[2] Chen W, Li Y, Tian Z, Zhang F. 2D and 3D object detection algorithms from images: A Survey. Array. 2023;100305.

[3] Deng J, Xuan X, Wang W, Li Z, Yao H, Wang Z. A review of research on object detection based on deep learning. In: Journal of Physics: Conference Series. IOP Publishing; 2020. p. 012028.

[4] LUO H lan, CHEN H kun. Survey of object detection based on deep learning. Acta Electonica Sinica. 2020;48(6):1230.

[5] Ren J, Wang Y. Overview of object detection algorithms using convolutional neural networks. Journal of Computer and Communications. 2022;10(1):115–32.

[6] Zhao L, Li S. Object detection algorithm based on improved YOLOv3. Electronics (Basel). 2020;9(3):537.

[7] Li Z, Du Y, Zhu M, Zhou S, Zhang L. A survey of 3D object detection algorithms for intelligent vehicles development. Artif Life Robot. 2022;1–8.

[8] Li K, Cao L. A review of object detection techniques. In: 2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT). IEEE; 2020. p. 385–90.

[9] Bouguettaya A, Kechida A, TABERKIT AM. A survey on lightweight CNN-based object detection algorithms for platforms with limited computational resources. International Journal of Informatics and Applied Mathematics. 2019;2(2):28–44.

[10] Zhao R, Niu X, Wu Y, Luk W, Liu Q. 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. p. 255–67.

[11] Huang R, Pedoeem J, Chen C. 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. p. 2503–10.

[12] Waheed SR, Suaib NM, Rahim MSM, Adnan MM, Salim AA. Deep learning algorithms-based object detection and localization revisited. In: journal of physics: conference series. IOP Publishing; 2021. p. 012001.

[13] Mahaur B, Singh N, Mishra KK. Road object detection: a comparative study of deep learning-based algorithms. Multimed Tools Appl. 2022;81(10):14247–82.

[14] John A, Meva D. A comparative study of various object detection algorithms and performance analysis. International Journal of Computer Sciences and Engineering. 2020;8(10):158–63.

[15] Padilla R, Netto SL, Da Silva EAB. A survey on performance metrics for object-detection algorithms. In: 2020 international conference on systems, signals and image processing (IWSSIP). IEEE; 2020. p. 237–42.

[16] Sun W, Dai L, Zhang X, Chang P, He X. RSOD: Real-time small object detection algorithm in UAV-based traffic monitoring. Applied Intelligence. 2021;1–16.

[17] Galteri L, Bertini M, Seidenari L, Del Bimbo A. Video compression for object detection algorithms. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE; 2018. p. 3007–12.

[18] Kumar A, Zhang ZJ, Lyu H. Object detection in real time based on improved single shot multi-box detector algorithm. EURASIP J Wirel Commun Netw. 2020;2020:1–18.

[19] Raghunandan A, Raghav P, Aradhya HVR. Object detection algorithms for video surveillance applications. In: 2018 International Conference on Communication and Signal Processing (ICCSP). IEEE; 2018. p. 563–8.

[20] Zangana HM, Mustafa FM, Omar M. A Hybrid Approach for Robust Object Detection: Integrating Template Matching and Faster R-CNN. EAI Endorsed Transactions on AI and Robotics. 2024;3.

[21] Li M, Zhu H, Chen H, Xue L, Gao T. Research on object detection algorithm based on deep learning. In: Journal of Physics: Conference Series. IOP Publishing; 2021. p. 012046.

[22] Chen C, Liu MY, Tuzel O, Xiao J. 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. p. 214–30.

[23] Du L, Zhang R, Wang X. Overview of two-stage object detection algorithms. In: Journal of Physics: Conference Series. IOP Publishing; 2020. p. 012033.

[24] Xiao Y, Tian Z, Yu J, Zhang Y, Liu S, Du S, et al. A review of object detection based on deep learning. Multimed Tools Appl. 2020;79:23729–91.

[25] Malhotra P, Garg E. Object detection techniques: a comparison. In: 2020 7th International Conference on Smart Structures and Systems (ICSSS). IEEE; 2020. p. 1–4.

[26] Cuevas C, Yáñez EM, García N. Labeled dataset for integral evaluation of moving object detection algorithms: LASIESTA. Computer Vision and Image Understanding. 2016;152:103–17.

[27] Peng L, Wang H, Li J. Uncertainty evaluation of object detection algorithms for autonomous vehicles. Automotive Innovation. 2021;4(3):241–52.

[28] Haris M, Glowacz A. Road object detection: A comparative study of deep learning-based algorithms. Electronics (Basel). 2021;10(16):1932.

[29] Wang J, Jiang S, Song W, Yang Y. A comparative study of small object detection algorithms. In: 2019 Chinese control conference (CCC). IEEE; 2019. p. 8507–12.

[30] Yadav N, Binay U. Comparative study of object detection algorithms. International Research

Journal of Engineering and Technology (IRJET). 2017;4(11):586–91.

[31] Zhou Y, Yang X, Zhang G, Wang J, Liu Y, Hou L, et al. Mmrotate: A rotated object detection benchmark using pytorch. In: Proceedings of the 30th ACM International Conference on Multimedia. 2022. p. 7331–4.