Vol. 8, No. 2, April 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Performance Analysis of YOLOv12 and RF-DETR in Face Detection

David Hendrawan[1], Wahyuni[2], Pitrasacha Adytia[3]

[1,2]Department of Informatics Engineering, STMIK Widya Cipta Dharma, Samarinda, Indonesia

[3]Department of Information Systems, STMIK Widya Cipta Dharma, Samarinda, Indonesia

Received:
December 6, 2025
Revised:
March 11, 2026
Accepted:
April 11, 2026
Published:
April 26, 2026

Corresponding Author:
Author Name*:
Wahyuni
Email*:
wahyuni@wicida.ac.id

DOI:
10.63158/journalisi.v8i2.1561

Abstract. Face detection in dense and occluded environments remains a significant challenge in computer vision. This study compares the CNN-based YOLOv12 and the Transformer-based RF-DETR to determine the optimal balance between accuracy and latency for resource-constrained edge computing. Using the WIDER FACE dataset and an NVIDIA T4 GPU, multiple model variants were evaluated. Due to GPU memory constraints during training of the RF-DETR Medium variant, a standardized batch size of 8 was implemented across all models. To ensure methodological rigor, quantitative metrics (precision, recall, F1-score, mAP) were strictly assessed on the validation set. Concurrently, a 100-image subset of the test set was used exclusively for inference efficiency benchmarking, completely separate from detection evaluation. Results indicate YOLOv12X achieved superior overall detection performance (F1-score: 0.764, mAP@50:95: 0.440), significantly outperforming RF-DETR Medium. For real-time applications, YOLOv12M demonstrated the highest efficiency (36.17 FPS vs. 23.32 FPS). Qualitatively, YOLOv12 maintained high sensitivity in crowded scenes, whereas RF-DETR provided stable small-scale face detection despite its lower recall. Overall, under these constrained-hardware conditions, YOLOv12 appears to be a highly viable solution for surveillance systems, while RF-DETR offers a stable alternative for small-object detection when computational overhead and training budgets are less restrictive.

Keywords: YOLOv12, RF-DETR, WIDER FACE, Inference Latency, Edge Deployment

1. INTRODUCTION

Face detection is one core component of many modern computer vision systems. It has become an imperative technique for several applications such as face recognition, human–computer interaction, and security systems, including surveillance systems and biometric authentication [1], [2], [3], [4]. Over the last decade, owing to the rapid development of deep learning techniques, object detection models have improved substantially, enabling real-time processing of data with high detection accuracy [4], [5]. These developments have made face detection an essential element not only in research but also in real-world applications.

However, even with the development of deep learning–based face detection, maintaining good performance under these real-world challenges (e.g., varying illumination conditions, occlusion, pose changes, and low-resolution images) remains a challenge. It is well-known that traditional CNN-based detectors may suffer from these issues, especially when detecting small, blurred, or partially occluded faces in uncontrolled environments [1], [2]. These issues are also addressed by recently developed specialized architectures and methods, including multitask cascaded networks for joint detection and alignment, super-resolution modules to enhance image quality, and refinement modules for detecting tiny or challenging faces [6].

To address these complexities, recent advances in object detection have been dominated by two architectures: one-stage detectors (such as the YOLO family) and Transformer-based detectors (like DETR and its variants) [7], [8], [9], [10]. One-stage detectors are popular for their real-time speed and simplicity, making simultaneous predictions for bounding boxes and class probabilities in a single pass, and have benefited from ongoing improvements in backbone design and feature aggregation for better accuracy and efficiency [11]. In contrast, Transformer-based models use self-attention mechanisms to model global dependencies, thereby enhancing their ability to handle complex scenes and occlusions and often resulting in greater robustness to image corruptions compared to conventional CNN-based detectors [11], [12]. However, this global attention mechanism conventionally requires significant computational resources and a large memory footprint, which poses a practical challenge for deployment [7], [8].

As the latest iterations of these architectures, YOLOv12 and RF-DETR represent the state of the art in their respective categories. Given the distinct architectural philosophies of YOLOv12 and RF-DETR, it is essential to understand how these models perform when applied to a specific and challenging vision task such as face detection, especially considering the operational constraints of real-time surveillance systems. Previous studies have shown that while YOLO-based models generally outperform in speed, Transformer-based detectors often achieve higher precision in cluttered environments or under occlusions [7], [13], [14], [15]. However, direct comparative evaluations between YOLOv12 and RF-DETR within the context of face detection remain limited [7], [14], [15], [16], [17], [18]. Such comparisons are crucial for practitioners and researchers to determine the optimal trade-off between speed and accuracy, especially for deployment in high-efficiency real-time systems [18].

While recent comparative literature provides valuable insights into the trade-offs in performance between one-stage and Transformer-based detectors, these studies primarily focus on diverse applied domains rather than the unique challenges of face detection [7], [14], [16], [18]. For instance, assessments in agricultural, medical, and industrial settings generally agree that while YOLOv12 maintains a distinct advantage in inference speed and resource efficiency, RF-DETR demonstrates superior robustness in handling severe occlusions [7], [14], [15], [16], [17], [18]. Furthermore, although foundational face analysis research has begun evaluating newer YOLO architectures, such as YOLOv12, against legacy models, they omit Transformer-based counterparts entirely [19]. This creates a critical synthesis gap: it remains unknown whether the occlusion-handling superiority of RF-DETR observed in generic object detection translates to the dense, structurally distinct domain of human face detection.

Consequently, a significant research gap exists regarding the comparative efficacy of state-of-the-art CNNs and Transformers specifically for face detection under strict hardware limitations. The novelty of this study lies in providing, to the best of our knowledge, an early direct, constrained-hardware comparative analysis between YOLOv12 and RF-DETR on the challenging WIDER FACE dataset. The primary objective is to evaluate and determine the optimal architectural balance between detection accuracy and computational latency for real-time edge deployment. By systematically analysing quantitative metrics (precision, recall, F1-score, mAP) alongside inference efficiency (FPS,

latency, GFLOPs), this study contributes a rigorous empirical framework. This contribution provides essential, evidence-based guidance for researchers and practitioners in selecting the most robust face detection architecture for resource-constrained surveillance and biometric systems.

2. METHODS

This section presents the methodology employed to compare the performance of the YOLOv12 and RF-DETR models in face detection. To ensure a rigorous and reproducible comparison, the research workflow is deliberately designed as a constrained-hardware evaluation framework that simulates real-world resource limitations. The process is systematically organized into three primary stages, as illustrated in Figure 1. First, data acquisition and preparation involve structuring the WIDER FACE dataset and standardizing annotations into architecture-specific formats. Second, model training is executed on the dataset's training split under strictly harmonized hyperparameters and a controlled effective batch size across two different software ecosystems (Ultralytics and Roboflow) to ensure architectural fairness. Following the training phase, the optimal model weights were determined by selecting the checkpoint that achieved the highest mean average precision (mAP@50:95) on the validation set. Finally, these selected models undergo performance evaluation, encompassing quantitative assessment using established detection metrics on the aforementioned validation set, inference benchmarking for computational efficiency using the test set, and qualitative analysis of extreme scenarios. Each stage is structured to ensure an objective comparison, thereby producing valid analytical results consistent with the research objectives.

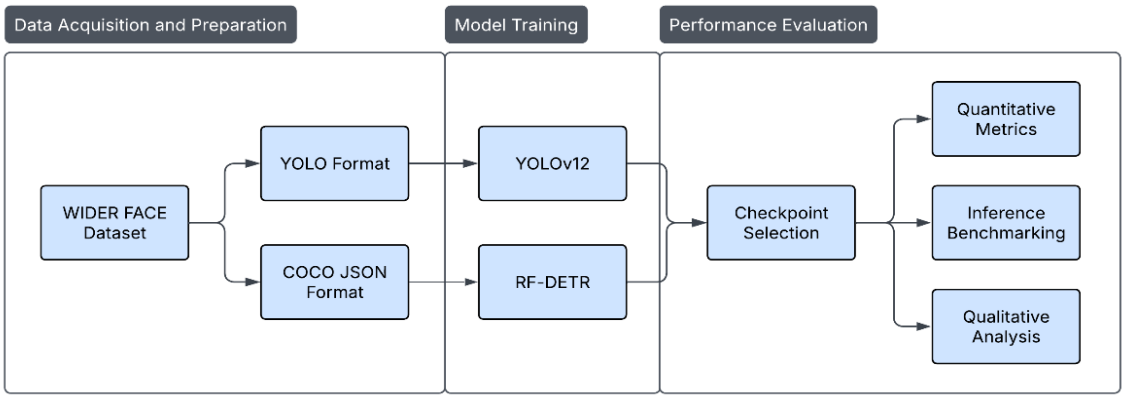

Figure 1. Flowchart of YOLOv12 and RF-DETR Performance Comparison Analysis

2.1. Data Acquisition and Preparation

This section outlines the dataset and the preparation procedures implemented to enable a rigorous and equitable comparison of the YOLOv12 and RF-DETR models.

1) Dataset

The WIDER FACE dataset, utilized in this study, is recognized as one of the largest and most challenging benchmarks for face detection [1], [2], [4], [20]. This dataset consists of 32,203 images with 393,703 labeled face bounding boxes with a high degree of variability in scale, pose, and occlusion [1], [2], [4], [20]. Figure 2 illustrates a representative sample from the dataset, visually showcasing these extreme variations in facial scale, complex poses, and dense occlusions that the detectors must overcome**.** The WIDER FACE dataset is organized based on 61 event classes. Each event class consists of three parts: 40% for training, 10% for validation, and 50% for testing. The training set contains 12880 images, the validation set 3226 images, and the test set 16097 images [1], [2], [4], [20]. For the purpose of this comparative study, model training was conducted exclusively on the training set. All quantitative detection metrics (precision, recall, F1-score, and mAP) reported in this study were evaluated strictly on the validation set, aligning with standard benchmark practices where test set annotations are withheld. Conversely, a random subset of 100 images sampled from the test directory was utilized solely for inference speed benchmarking (FPS and latency) to ensure the models were evaluated on completely unseen data. Regarding annotations, the dataset provides detailed ground truth data consolidated into a single text file, including bounding box coordinates and supplementary facial metadata for each image [1], [4], [20].

Figure 2. Sample images from the WIDER FACE dataset showcasing extreme variations in facial scale, complex poses, and dense occlusions

2) Dataset Preparation

The WIDER FACE dataset stores all annotations in a single text file by default [1], [20]. The bounding boxes are represented in the (x, y, w, h) format, where (x, y) denote the top-left

corner coordinates and (w, h) indicate the width and height. Furthermore, each annotation is accompanied by supplementary metadata attributes, such as blur, expression, lighting, validity of the annotated image, occlusion, and pose [1], [4], [6], [20].

However, the YOLO and RF-DETR architectures require an annotation format that differs from the dataset's native structure. Specifically, YOLOv12 necessitates coordinates in the (cx, cy, w, h) format, where bounding box dimensions are normalized relative to the image size and defined by the center point, stored in individual annotation files for each image [21]. To address this formatting discrepancy, this study utilizes a preprocessed version of the WIDER FACE dataset, specifically adapted for YOLO by previous research [19].

Unlike the YOLO-based models, RF-DETR requires the standard COCO annotation format. This format consolidates image metadata, categories, and bounding box coordinates into a single, structured JSON file, unlike the individual normalized text files commonly used by YOLO-based models [8]. To address this formatting discrepancy, we developed a Python script that automatically converts original annotations into a COCO-compatible JSON format. Our script parses each annotation entry and aggregates it into a structured JSON file, mapping a single category ID (e.g., 1) for all face objects to align with the single-class detection task. This preprocessing step ensures seamless integration with the RF-DETR training pipeline.

2.2. Training Object Detection Model

In this study, the models utilized are YOLOv12 and RF-DETR. These models were selected to represent two distinct architectural paradigms: the one-stage detector, represented by YOLOv12, and the Transformer-based detector, represented by RF-DETR [7], [14], [16]. Furthermore, both models stand as the latest variants in their respective families and have garnered significant popularity in recent object detection research.

1) YOLOv12 Object Detection Model

YOLOv12 represents a transformative leap in convolutional neural network-based object detection compared to its predecessors [9], [10]. It integrates traditional convolutional architectures with attention-centric mechanisms typically found in Transformer models, such as RF-DETR, all while retaining the core characteristics of the YOLO framework. The architecture of YOLOv12 comprises three main components: backbone, neck, and head

[19], [22], [23]. Figure 3 depicts the training architecture of YOLOv12, highlighting the continuous data flow from the input image, through the advanced feature extraction modules, to the final detection output.

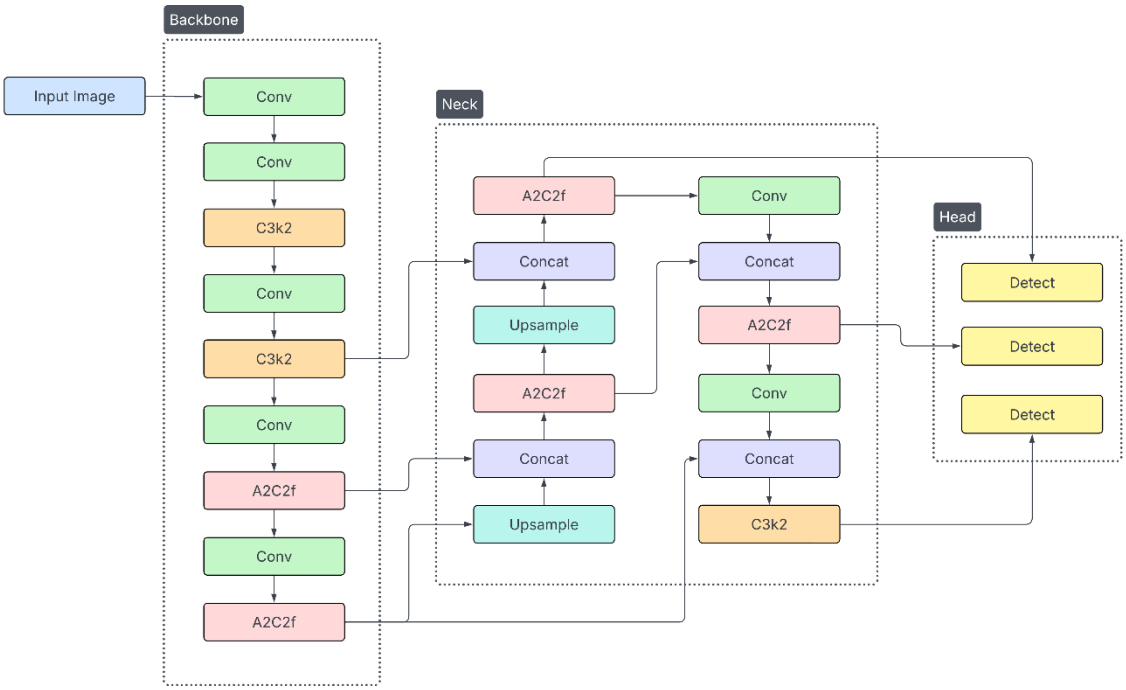

Figure 3. The architecture of YOLOv12 model

Unlike previous YOLO versions, YOLOv12 introduces a new backbone called R-ELAN (Residual Efficient Layer Aggregation Network) [9], [24], [25]. R-ELAN combines residual connections and multi-scale feature fusion to enhance information flow, preventing the loss of important features in deeper layers [9], [19], [25]. Additionally, the model employs 7x7 separable convolutions instead of the standard 3x3 kernels. This design allows for capturing a wider spatial context with 60% fewer parameters [9], [22], [23]. It is especially relevant for face detection, as it enables the model to understand facial feature relationships efficiently without requiring the heavy computation typical of positional embeddings in Transformers [22], [26].

For the neck component, the model integrates an area attention module optimized with FlashAttention [22], [25], [26]. This mechanism partitions feature maps into specific horizontal and vertical regions, facilitating in-depth processing without compromising global contextual awareness. This strategy allows the model to prioritize details in

specific regions, such as small or occluded faces, while reducing memory usage by up to 40% compared to traditional self-attention [22], [26]. This synergy of innovations yields an optimal trade-off between speed and accuracy; YOLOv12 variants have been shown to outperform hybrid models such as RT-DETR, achieving exceptionally low inference latency (under 10 ms on edge devices) [10], [13], [26]. These characteristics make YOLOv12 an ideal benchmark for this study, particularly for face detection applications that require high precision for small-scale objects.

Finally, for the head, the model adopts a decoupled, anchor-free architecture [22], [26]. Unlike traditional coupled heads, this design separates the object classification and bounding box regression into distinct branches, reducing optimization conflicts and accelerating convergence [10], [22], [26]. By eliminating reliance on predefined anchor boxes, the model becomes more robust at detecting objects of varying scales, such as small faces, and significantly reduces computational overhead during post-processing [10], [22], [26].

The YOLOv12 architecture is available in five distinct variants: YOLOv12N (2.6 million parameters), YOLOv12S (9.3 million parameters), YOLOv12M (20.2 million parameters), YOLOv12L (26.4 million parameters), and YOLOv12X (59.1 million parameters) [10], [22], [23]. In this study, all five variants are utilized to benchmark their performance against the RF-DETR model. This comprehensive approach aims to provide a broad spectrum of insights into the comparative performance of YOLOv12 and RF-DETR, specifically in face detection.

2) RF-DETR Object Detection Model

RF-DETR is a Transformer-based object detection architecture designed for real-time applications, optimized to achieve high accuracy with superior computational efficiency compared to most other Transformer-based models [8]. The architecture comprises four primary components: backbone, projector (which serves as the interface between backbone and encoder-decoder), encoder-decoder, and head [8]. Figure 4 illustrates the RF-DETR training architecture, delineating the feature processing pipeline from the DINOv2 backbone through the deformable cross-attention modules.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

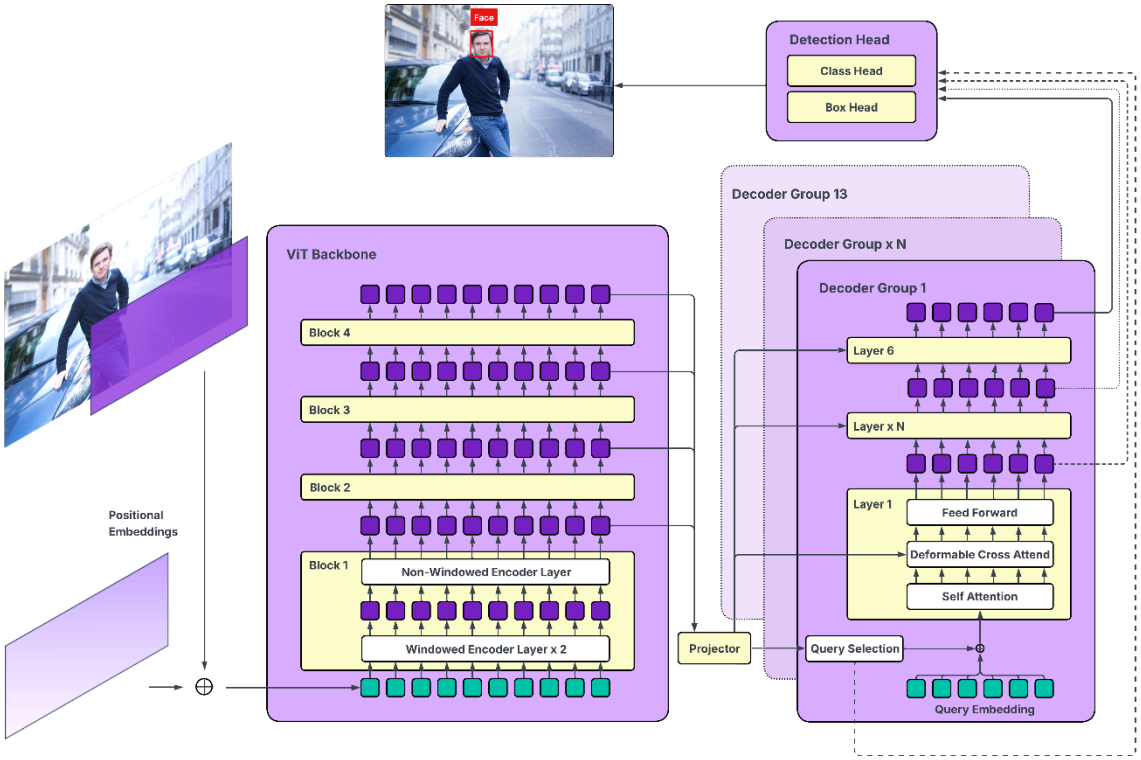

Figure 4. The architecture of RF-DETR model

Unlike conventional CNN-based approaches, RF-DETR utilizes global self-attention mechanisms to capture contextual relationships among features across the entire image. The architecture builds upon Deformable DETR and LW-DETR, integrating the pre-trained DINOv2 Vision Transformer as its backbone [7], [8]. This backbone significantly enhances the model's global perception capability from the initial layer, thereby improving its adaptability to complex challenges such as face occlusion [7], [14].

RF-DETR introduces a significant advancement by eliminating traditional object detection components, such as anchor boxes and Non-Maximum Suppression (NMS), which are prevalent in earlier YOLO versions [8], [9], [10], [19], [27], [28]. Consequently, RF-DETR employs a Transformer-based encoder-decoder architecture with deformable cross-attention, enabling selective focus on spatially relevant features and improving object detection performance under challenging conditions [7], [8]. Additionally, unlike standard DETR variants, RF-DETR implements single scale feature extraction strategy to reduce computational overhead. This approach enables faster inference while preserving accuracy [8].

The RF-DETR architecture is available in five distinct variants: RF-DETR Nano (30.5 million parameters), RF-DETR Small (32.1 million parameters), RF-DETR Medium (33.7 million parameters), RF-DETR Base (29 million parameters), and RF-DETR Large (128 million parameters) [7]. For this study, the RF-DETR Nano, Small, and Medium variants were selected for their optimal balance of computational efficiency and detection accuracy, while maintaining latency comparable to that of the Base model [7]. Although the Base model has been widely adopted in previous research, the Medium variant now provides substantially higher accuracy at similar latency [7]. This selection is consistent with the developers' recommendation to prioritize the Medium model over the Base version [14]. The RF-DETR Large variant model was excluded due to computational resource limitations and its ongoing development during the research period [7].

3) Training Setup

The procedures and methods used to train both the YOLOv12 and RF-DETR object detection models were conducted in an identical training environment, from the computational resources used to the identical parameter settings, ensuring that the model comparison was fair and rigorous. Model implementation and training were conducted in a Kaggle Notebook environment with 30 GB RAM, a 4-core CPU, and dual NVIDIA T4 accelerators. Both architectures were evaluated comprehensively, covering all YOLOv12 variants (Nano to Extra-Large) and RF-DETR variants (Nano to Medium), with each model trained for 100 epochs using mixed precision (FP16).

Table 1. Hyperparameter Configuration

Hyperparameter	Value
Input Image Size	640
Epochs	100
Batch Size	8
Optimizer	AdamW
Learning Rate	0.0001

A critical observation during the experimental setup was the significant disparity in memory consumption between the two architectures. While YOLOv12 variants operated efficiently, the RF-DETR architecture exhibited substantial VRAM usage due to the

quadratic complexity of its self-attention mechanisms. Consequently, despite utilizing dual-GPU infrastructure, a batch size of 8 was the maximum feasible limit to prevent Out-of-Memory (OOM) errors for the RF-DETR Medium variant. To ensure a fair and consistent comparative baseline, this batch size was subsequently standardized across all models, including the more memory-efficient YOLOv12. It is crucial to explicitly acknowledge this standardization as a deliberate methodological compromise and to formally frame this study as a constrained-hardware benchmark rather than a fully neutral architectural comparison. While establishing operational parity, restricting the effective batch size to 8 may systematically disadvantage Transformer-based architectures like RF-DETR, which typically rely on larger batch sizes for optimal gradient convergence compared to CNNs. Nevertheless, this constraint intentionally simulates real-world hardware limitations, providing empirical evidence of the high resource barrier associated with training state-of-the-art Transformers.

Regarding software implementation, YOLOv12 was trained using Ultralytics framework, which is designed for high-speed and efficient detection. In contrast, RF-DETR was implemented in the Roboflow framework, integrating the DINOv2 backbone with the Deformable DETR architecture [14]. This divergence in software environments was necessitated by the native support and official repositories of each architecture. To ensure a fair and rigorous comparison, core training hyperparameters, such as input image size (640x640 pixels), epochs, batch size, optimizer, and learning rate, were strictly aligned across both platforms, as detailed in Table 1. However, to evaluate each architecture at its peak potential, the native data augmentation pipelines inherent to Ultralytics and Roboflow were maintained at their default settings. This decision was deliberately made to ensure that neither model was artificially disadvantaged by enforcing a sub-optimal, framework-agnostic augmentation strategy. While this allows each model to operate under its native, optimized architectural conditions, we acknowledge that this constitutes another methodological compromise. Consequently, this study is not a perfectly harmonized benchmark at the data augmentation and preprocessing level. While training used mixed precision to optimize resource usage, inference benchmarks were conducted in Single-Precision (FP32) mode to evaluate the intrinsic architectural complexity of each model without quantization effects.

2.3. Performance Evaluation

The final stage of the methodology comprises a multifaceted performance evaluation to comprehensively assess the YOLOv12 and RF-DETR architectures. This evaluation is systematically divided into three components.

First, Quantitative Metrics were utilized to measure the baseline detection accuracy. The face detection capabilities were evaluated using standard object detection metrics, including precision, recall, F1-score, and mean Average Precision (mAP). These metrics measure performance by comparing predicted bounding boxes against ground-truth annotations. As established, each model was evaluated using the optimal checkpoint that achieved the highest mAP@50:95 on the validation set during training. To ensure strict reproducibility and consistency across both the Ultralytics and Roboflow ecosystems, a best-checkpoint logic was applied throughout training. Early stopping mechanisms were not utilized; instead, each model completed the full 100-epoch schedule, and the checkpoint yielding the highest validation metric was ultimately selected.

Second, Inference Benchmarking was conducted to evaluate the practical feasibility and computational cost of the models for real-world deployment. This phase measured the intrinsic architectural efficiency by analyzing the number of parameters, Giga Floating Point Operations (GFLOPs), inference latency (ms), and Frames Per Second (FPS). To ensure an unbiased hardware assessment, these benchmarks were executed in Single-Precision (FP32) mode using a random subset of 100 unannotated images from the test set.

Third, Qualitative Analysis was performed to investigate the models' behavior beyond numerical metrics. This involved visually inspecting the detection results under extreme and challenging scenarios, such as dense crowds, severe occlusions, and extreme scale variations. This qualitative assessment aims to identify specific strengths, trade-offs, and failure patterns (e.g., localization jitter or false negatives) inherent to the CNN and Transformer architectures.

3. RESULTS AND DISCUSSION

This section presents the experimental results, discussion, and a comparative analysis of the performance of two object detection architectures, YOLOv12 and RF-DETR, for face detection. The evaluation was conducted comprehensively, considering various quantitative metrics commonly used in object detection tasks, including precision, recall, F1-score, mAP@50, and mAP@50:95, to provide an objective overview of the models' accuracy and consistency. Additionally, the analysis includes a comparison of the different model variants within each architecture to identify trade-offs between performance and model complexity. The results are expected to provide comprehensive insights into the strengths and limitations of YOLOv12 and RF-DETR and to serve as a basis for determining the most suitable model for face detection across various application scenarios.

3.1. Performance Evaluation

1) Precision, Recall, and F1-Score

A comprehensive evaluation of all models revealed a consistent pattern in which the YOLOv12 variants outperformed RF-DETR across precision, recall, and F1-score metrics. This indicates that the YOLOv12 architecture is fundamentally more adept at balancing false positives and false negatives in unconstrained facial environments. The superiority of the YOLO architecture was most evident in the YOLOv12X variant, which achieved the highest precision (0.881), recall (0.675), and F1-score (0.764) values. In comparison, the RF-DETR variants demonstrated lower recall, with the best-performing RF-DETR Medium variant achieving 0.570, approximately 0.10 points lower than YOLOv12X. Analytically, this 0.105-point gap in Recall is the primary driver of the overall F1-Score disparity. This suggests that while RF-DETR is relatively precise when it successfully detects a face, its global attention mechanism struggles significantly more to identify less salient or occluded facial features, leading to a much higher rate of missed detections.

Interestingly, this performance difference is not limited to the larger capacity variants. YOLOv12S, which is computationally lighter, still achieved an F1-Score of 0.723, outperforming RF-DETR Medium (0.686). This finding suggests that the YOLOv12 architecture exhibits better generalization in detecting faces across scales in the WIDER FACE dataset, even with lower model capacity. For details, the overall evaluation results

on YOLOv12 and RF-DETR using precision, recall, and F1-score have been summarized in Table 2.

Table 2. Comparative analysis of Precision, Recall, and F1-Score for face detection using YOLOv12 and RF-DETR

Models	Precision	Recall	F1-Score
YOLOv12X	0.881	0.675	0.764
YOLOv12L	0.876	0.669	0.759
YOLOv12M	0.877	0.659	0.752
YOLOv12S	0.858	0.625	0.723
YOLOv12N	0.828	0.563	0.671
RF-DETR Medium	0.862	0.570	0.686
RF-DETR Small	0.847	0.530	0.652
RF-DETR Nano	0.828	0.470	0.599

2) Comparative Analysis of Mean Average Precision (mAP)

In addition to evaluating precision, recall, and F1-score, model performance was further assessed using metrics such as mAP@50 and mAP@50:95. Figure 5 summarizes the performance of various YOLOv12 and RF-DETR variants on the WIDER FACE dataset. Experimental results indicate that the majority of YOLOv12 variants consistently outperform RF-DETR across both metrics, with this advantage becoming increasingly pronounced in larger-capacity models.

Crucially, the performance gap between the architectures widens significantly when evaluated under strict localization criteria (mAP@50:95) compared to the standard threshold (mAP@50). The YOLOv12X variant achieved mAP@50:95 of 0.440, surpassing the best RF-DETR variant (Medium) by 0.117, which achieved 0.323. This difference indicates that YOLOv12 has better bounding box localization precision across various Intersection over Union (IoU) thresholds, especially under stricter criteria. This 0.117 discrepancy reveals a critical analytical insight: YOLOv12 not only detects more faces but also generates bounding boxes that fit the ground truth much more tightly than RF-DETR. RF-DETR's relatively steep drop in performance from mAP@50 to mAP@50:95

implies that its predicted bounding boxes often lack the pixel-level precision required for high-IoU matches.

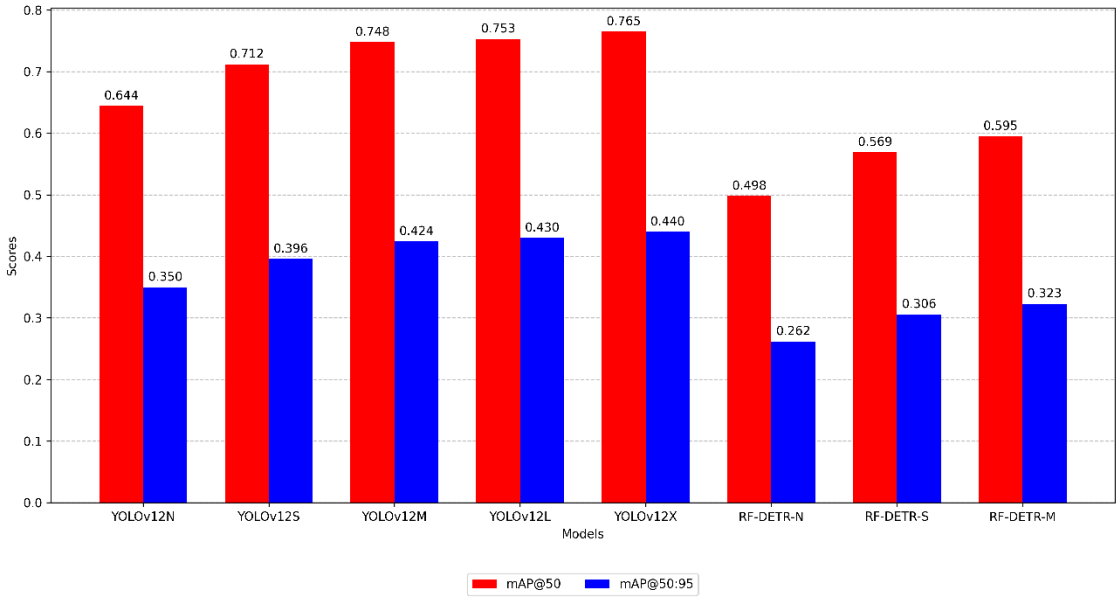

Figure 5. Mean Average Precision (mAP) comparison for face detection using YOLOv12 and RF-DETR

Moreover, YOLOv12X also achieved highest mAP@50 with 0.765, significantly outperforming RF-DETR Medium, which scored 0.595. The advantage of YOLOv12 in the mAP@50 metric being higher than mAP@50:95 indicates that the architecture is capable of detecting faces with high baseline accuracy (IoU > 0.5), even in complex scenarios. In contrast, RF-DETR's lower performance on both metrics suggests a higher false negative rate, particularly during the initial detection phase, which affects its recall on dense datasets such as WIDER FACE. For details, the overall evaluation results on YOLOv12 and RF-DETR using mAP have been summarized in Figure 5.

3) Inference Efficiency and Computational Cost

To further evaluate the practical feasibility of YOLOv12 and RF-DETR for real-world face detection applications, an analysis of inference efficiency and computational cost was conducted. This evaluation focused on four main indicators: the number of parameters, Giga Floating Point Operations (GFLOPs), inference latency, and Frames Per Second (FPS). All models were tested using identical input resolution and hardware configurations to ensure a fair and objective comparison.

The testing protocol involved executing each model for 30 iterations using 100 images sampled from the WIDER FACE test directory. To ensure rigorous and equitable benchmarking, both architectures utilized identical preprocessing pipelines, standardizing input resolutions to 640x640 pixels without applying inference-time augmentations. Furthermore, a standard 10 iteration warm-up phase was executed prior to measurement to stabilize GPU clock frequencies, and the inference batch size was strictly set to 1. This batch size of 1 was specifically selected to accurately simulate the sequential frame-by-frame processing typical of real-time video surveillance. Measurements were performed using the native PyTorch engine in single-precision (FP32). This configuration was deliberately chosen to establish a baseline architectural performance comparison, eliminating variables introduced by hardware-specific compilers (e.g., TensorRT) or quantization schemes. This metric reflects the intrinsic computational cost of each model in a standard research environment.

Table 3 presents comparative results on inference efficiency for the various YOLOv12 and RF-DETR variants. Overall, all YOLOv12 variants achieve greater inference efficiency than RF-DETR across every model scale. YOLOv12 exhibits a predictable scalability trend: as model size increases, computational cost and inference latency rise proportionally, while frames per second (FPS) decrease. For example, YOLOv12X, the largest variant, incurs highest computational cost with 99.91 GFLOPs and 59.12 million parameters, resulting in an inference latency of 80.62 ms and a throughput of 12.38 FPS. Conversely, YOLOv12N, the smallest variant, requires only 3.24 GFLOPs and 2.57 million parameters, achieving a low latency of 16.41 ms and real-time performance at 60.93 FPS.

Compared to YOLOv12, all RF-DETR variants exhibit relatively higher computational costs, despite having a comparable number of parameters to medium-sized YOLOv12 models. RF-DETR Nano, Small, and Medium each require more than 38 GFLOPs, which is significantly higher than YOLOv12S (10.76 GFLOPs) and YOLOv12M (33.87 GFLOPs). This high computational demand directly affects inference speed, with RF-DETR variants achieving FPS in the range of 23-26 FPS. For example, RF-DETR Medium records a latency of 42.88 ms with 23.32 FPS, while YOLOv12M achieves lower latency at 27.65 ms and higher throughput of 36.17 FPS, despite its lower computational cost.

These experimental results indicate that YOLOv12 offers superior computational efficiency and achieves a better balance between inference speed and model capacity. This advantage can be attributed to the one-stage detection paradigm and the optimized convolutional-based architecture, which enable faster feature extraction and bounding box regression. In contrast, the Transformer-based architecture in RF-DETR, while effective at modeling global feature relationships, introduces additional computational overhead from the multi-head attention mechanism and iterative decoding, thereby increasing inference latency.

Table 3. Comparative analysis of Parameters, GFLOPs, Latency, and FPS for face detection using YOLOv12 and RF-DETR

Models	Params (M)	GFLOPs	Latency (ms)	FPS
YOLOv12X	59.12	99.91	80.62	12.38
YOLOv12L	26.39	44.70	41.33	24.19
YOLOv12M	20.14	33.87	27.65	36.17
YOLOv12S	9.25	10.76	17.12	58.41
YOLOv12N	2.57	3.24	16.41	60.93
RF-DETR Medium	28.42	39.28	42.88	23.32
RF-DETR Small	27.21	38.79	40.59	24.64
RF-DETR Nano	26.00	38.30	38.13	26.23

Overall, this inference efficiency analysis emphasizes that YOLOv12 is more suitable for real-time face detection scenarios and resource-constrained environments, such as surveillance systems and edge computing applications. Meanwhile, RF-DETR remains competitive under certain conditions, but its higher computational cost can limit its performance in latency-sensitive applications. These findings reinforce the importance of architectural efficiency as a key factor in practical face detection performance, alongside accuracy metrics alone.

4) Qualitative Analysis on Challenging Scenarios

Unlike the ideal scenario, where both models show comparable performance in face detection, as shown in Figure 6, analysis under extreme conditions reveals fundamental differences in detection characteristics and failure patterns between YOLOv12X and RF-

DETR Medium. In a dense crowd with a heavy occlusion scenario (Figure 7), a significant trade-off between sensitivity and stability is observed. YOLOv12X achieves superior detection performance, with high sensitivity, which directly correlates with its higher overall recall of 0.675, as shown in Table 2. This enables it to effectively detect faces in the back rows that are partially obscured by hats, thereby significantly reducing the number of false negatives. However, this high sensitivity introduces a minor trade-off in highly crowded areas: the model occasionally exhibits 'localization jitter,' where predicted bounding boxes may shift slightly around the target face due to the density of overlapping features. Crucially, this phenomenon does not equate to detection failure. As evidenced by the superior mAP@50:95 score of 0.440 as shown in Figure 5, these detections remain sufficiently accurate to satisfy strict Intersection over Union (IoU) thresholds. This indicates that while YOLOv12X prioritizes recall, its localization precision remains robust enough to outperform RF-DETR.

Figure 6. Face detection results of YOLOv12X and RF-DETR Medium in ideal scenario

Figure 7. Face detection results of YOLOv12X and RF-DETR Medium in crowded and occluded scenario

In contrast, RF-DETR Medium adopts a more conservative strategy. While generating visually stable bounding boxes, this model misses faint faces in the middle rows, particularly those with minimal or partially obscured visual features. This visual observation directly explains the substantial 0.105 point recall deficit between YOLOv12X and RF-DETR Medium reported in Table 2. This suggests that the attention mechanism in RF-DETR tends to suppress ambiguous object candidates to avoid false positives. However, this stability comes at the cost of missed detections (false negatives), which penalize the overall mAP score (Figure 5) far more severely than YOLOv12's minor localization variance.

Figure 8. Face detection results of YOLOv12X and RF-DETR Medium in crowded with small-scale faces scenario

In the extreme scale variation scenario on the stadium bleachers (Figure 8), the limitations of each model become more apparent. Although YOLOv12X achieves higher recall, as shown in Table 2, observations from this image reveal an anomaly: the model is less effective at detecting smaller faces. Nonetheless, YOLOv12X still performs exceptionally well in predicting large faces, but detection and bounding box prediction tend to be less precise for very small faces. This specific limitation prevents YOLOv12X from achieving even higher overall Recall metrics across the dataset. On the other hand, RF-DETR Medium performs better at detecting smaller faces than YOLOv12X, particularly in the back rows. However, while RF-DETR is more sensitive to small objects, it also produces false positives in some areas, such as mistaking parts of uniforms for faces. These visual instances of false positives generated by RF-DETR in complex backgrounds

provide direct context for its lower overall Precision (0.862) compared to YOLOv12X (0.881), as documented in Table 2. It suggests that although RF-DETR Medium is better at detecting small faces, the limitations of the Transformer architecture become evident when it struggles to differentiate faces from other objects with similar patterns.

Comparatively, this qualitative analysis concludes that there is a clear trade-off between the two architectures. YOLOv12X excels in sensitivity, making it the superior choice for critical surveillance applications that prioritize minimizing missed detections, though it occasionally leads to localization jitter (bounding box instability) in dense crowds. On the other hand, RF-DETR Medium adopts a more conservative and selective detection strategy. While it has limitations in capturing objects with weak or ambiguous visual features (resulting in lower recall), this model offers better prediction stability for clearly detected objects and tends to suppress potential false positives through its global attention mechanism.

3.2. Discussion

This study demonstrates that the evolution of CNN-based architectures, specifically YOLOv12, has effectively bridged the performance gap between CNNs and Transformers in complex face detection tasks. Our results indicate that YOLOv12X consistently outperforms the best RF-DETR variant (Medium) across all key metrics on the WIDER FACE dataset, achieving an F1-Score of 0.764 compared to 0.686 under the chosen training and hardware budget. It must be emphasized that these results demonstrate YOLOv12's superiority strictly within these specific limitations, and this performance gap may not necessarily persist under broader or fully optimized training conditions. This finding challenges the prevailing assumption in some general object detection literature that Transformers are inherently superior at handling occlusions.

The divergence of our results from previous studies highlights the domain-specific nature of detector performance. While recent comparative studies evaluating YOLOv12 and RF-DETR in other domains, such as coffee leaf disease detection, weapon detection in CCTV footage, and breast cancer detection and classification, have highlighted the context-dependent strengths of Transformer-based models, our study distinctly observes the superiority of the CNN-based YOLOv12 for face detection [14], [15], [16]. The quantitative results presented in Table 2 provide concrete evidence for this performance

divergence. The substantial gap in Recall (0.675 for YOLOv12X and 0.570 for RF-DETR Medium) provides analytical support for the hypothesis that human faces possess strong, localized geometric features (eyes, nose, mouth) that the optimized R-ELAN backbone and Area Attention in YOLOv12 can extract efficiently, even at varying scales. This quantitative superiority is directly validated by the qualitative observations in Figure 7, where YOLOv12X successfully identifies heavily occluded faces in dense crowds that RF-DETR ignores. While existing literature often assumes that Transformer architectures are inherently superior at handling occlusions, our findings challenge this notion for face detection tasks. Our empirical data suggests that, rather than excelling in dense occlusions, the deformable attention mechanism of RF-DETR may actually be too conservative. As demonstrated by its lower Recall but relatively competitive Precision (0.862), we theorize that RF-DETR tends to suppress ambiguous candidates in dense crowds to avoid false positives. While this remains a theoretical interpretation of the model's internal attention behavior rather than an independently verified mechanism, it is strongly supported by the observed numerical trade-offs.

Visually, this is mapped in Figure 8, where RF-DETR struggles to differentiate small faces from background uniform patterns, leading to false positives that cap its precision. This architectural characteristic directly contributes to the lower mAP scores (Figure 5), as the model penalizes itself by missing valid, partially occluded faces that YOLOv12 successfully localizes. By systematically mapping these qualitative behaviors to the quantitative results, a clear alignment emerges: the visual missed detections directly account for RF-DETR's recall deficits, the localized false positives in complex backgrounds establish its precision limits, and the observed bounding-box looseness directly translates to its steeper performance drop in mAP@50:95 compared to YOLOv12.

However, any synthesis of these findings must include a critical and explicit acknowledgment of the study's design limitations, to avoid a one-sided interpretation favoring YOLOv12. It is critical to acknowledge a central methodological limitation regarding the impact of training hyperparameters on this performance disparity. While this study maintained identical training parameters (batch size of 8, 100 epochs) to ensure a controlled and equitable hardware-constrained comparison, this setup inherently favors the convolutional neural network-based YOLOv12 architecture. Transformer-based models like RF-DETR generally lack the inductive bias of CNNs and typically require larger

batch sizes and longer training schedules to achieve optimal gradient convergence. Consequently, the lower recall (0.570) and suboptimal mAP scores observed in RF-DETR must be interpreted with caution; these metrics likely stem from under-convergence due to strict GPU memory constraints, rather than purely architectural limitations in face detection capabilities. Therefore, we conclude that the weaker performance of RF-DETR is not solely an architectural mismatch, but heavily compounded by a training environment that restricted its data-hungry nature. This constrained training environment constitutes a primary limitation of this study, as it precludes evaluating the Transformer model at its theoretical peak capacity and enforces a penalty on the Transformer's resource-intensive training requirements. Future studies leveraging high-memory infrastructure and extended training schedules could potentially narrow this performance gap and reveal the true baseline of RF-DETR in face detection.

From a deployment perspective, the superiority of YOLOv12 extends beyond accuracy metrics to practical viability in edge computing environments. Beyond the fundamental speed-versus-accuracy trade-off, practical deployment at the edge is strictly governed by memory footprint, hardware compatibility, and energy consumption. Our inference benchmarks reveal that YOLOv12M achieves real-time speeds (36.17 FPS), whereas RF-DETR Medium falls below the real-time threshold (23.32 FPS) under the same conditions. For resource-constrained applications, such as autonomous drone surveillance or embedded security cameras, this efficiency gap is decisive. Furthermore, the architectural complexity of RF-DETR, specifically its reliance on global self-attention mechanisms, typically requires specialized tensor cores and large memory bandwidth to run efficiently, making it less adaptable for standard edge devices. In contrast, CNN-based architectures like YOLOv12 benefit from decades of hardware optimization. They can be more readily compressed using techniques like INT8 quantization and deployed efficiently across a broader ecosystem of cost-effective edge GPUs, NPUs, and TPUs without catastrophic performance degradation. Even if prolonged training could improve RF-DETR's accuracy, its inherent computational overhead, stemming from complex multi-head attention mechanisms, remains a significant barrier for low-latency deployment compared to the streamlined one-stage pipeline of YOLOv12.

In summary, this study provides a realistic benchmark under strict resource limitations. Although RF-DETR shows promise as a specialized detector in applications where

bounding box stability is prioritized over sensitivity, YOLOv12 currently offers the best balance among precision, recall, and speed for real-time face detection systems strictly within the evaluated hardware and training constraints.

4. CONCLUSION

This study provides a comprehensive comparative analysis of the convolutional neural network-based YOLOv12 and the Transformer-based RF-DETR architectures for face detection on the WIDER FACE dataset. Experimental results indicate that the YOLOv12X variant demonstrates superior performance across nearly all quantitative metrics, achieving the highest precision (0.881), recall (0.675), and F1-score (0.764). Furthermore, the YOLOv12 family consistently surpasses RF-DETR in localization accuracy, with YOLOv12X achieving an mAP@50:95 of 0.440, substantially higher than the 0.323 achieved by the best Transformer variant, RF-DETR Medium. In terms of computational efficiency, the YOLOv12 model is more practical for real-time applications. For example, YOLOv12M achieves a higher throughput of 36.17 FPS with an inference latency of 27.65 ms than RF-DETR Medium, which reaches only 23.32 FPS and 42.88 ms, despite both models having comparable parameter counts. Further qualitative analysis reveals the distinct strengths of each architecture: YOLOv12X is highly sensitive and effective in scenarios requiring high Recall, such as in very crowded and occluded conditions, while RF-DETR Medium offers better stability and higher precision for detecting small faces in complex backgrounds, though false positives occasionally occur. In conclusion, under the constrained training budget and hardware configuration evaluated in this study, YOLOv12 appears to be the more practical choice for real-time face detection. RF-DETR remains a competitive alternative for specific scenarios that require global context modeling for detecting small objects. However, it is crucial to explicitly state that these findings are strictly conditioned by the specific training budget, hardware configuration, and dataset configuration used in this study. Consequently, these results should not be considered universally generalizable without broader multi-dataset or multi-budget evaluations. Future research could explore hybrid architectures or further optimization of Transformer-based models across diverse computational environments to reduce latency gaps while maintaining detection stability in extreme conditions.

REFERENCES

[1] A. S. Sanchez-Moreno, J. Olivares-Mercado, A. Hernandez-Suarez, K. Toscano-Medina, G. Sanchez-Perez, and G. Benitez-Garcia, "Efficient face recognition system for operating in unconstrained environments," *J. Imaging*, vol. 7, no. 9, p. 161, Sep. 2021, doi: 10.3390/jimaging7090161.

[2] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang, "YOLO-FaceV2: A scale and occlusion aware face detector," *Pattern Recognit.*, vol. 155, p. 110714, Nov. 2024, doi: 10.1016/j.patcog.2024.110714.

[3] B. Balachander, B. Sarveswari, S. N. S, C. S. Sowmiya, and R. Rajeshwari, "Performance analysis of lightweight YOLOv12 framework for object detection using real time Web camera based inputs," in *2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET)*, Aug. 2025, pp. 1–6. doi: 10.1109/ICRISET64803.2025.11252243.

[4] D. Mamieva, A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Improved face detection method via learning small faces on hard images based on a deep learning approach," *Sensors*, vol. 23, no. 1, p. 502, Jan. 2023, doi: 10.3390/s23010502.

[5] H. Du, H. Shi, D. Zeng, X.-P. Zhang, and T. Mei, "The elements of end-to-end deep face recognition: A survey of recent advances," *ACM Comput Surv*, vol. 54, no. 10s, p. 212, Sep. 2022, doi: 10.1145/3507902.

[6] Q. Xu, Z. Zhu, H. Ge, Z. Zhang, and X. Zang, "Effective face detector based on YOLOv5 and superresolution reconstruction," *Comput. Math. Methods Med.*, vol. 2021, no. 1, p. 7748350, 2021, doi: 10.1155/2021/7748350.

[7] M. E. Atik and M. Arkali, "Benchmarking YOLO and Transformer-based detectors for olive tree crown identification in UAV imagery," *Geomatics*, vol. 6, no. 2, p. 22, Feb. 2026, doi: 10.3390/geomatics6020022.

[8] N. Dahiya *et al.*, "Optimised RFO tuned RF-DETR model for precision urine microscopy for renal and systemic disease diagnosis," *Sci. Rep.*, vol. 15, no. 1, p. 25842, Jul. 2025, doi: 10.1038/s41598-025-11725-0.

[9] M. L. Ali and Z. Zhang, "The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection," *Computers*, vol. 13, no. 12, p. 336, Dec. 2024, doi: 10.3390/computers13120336.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[10] R. Sapkota, Z. Meng, M. Churuvija, X. Du, Z. Ma, and M. Karkee, "Comprehensive performance evaluation of YOLOv12, YOLO11, YOLOv10, YOLOv9 and YOLOv8 on detecting and counting fruitlet in complex orchard environments," *Agric. Commun.*, vol. 4, no. 1, p. 100125, Mar. 2026, doi: 10.1016/j.agrcom.2026.100125.

[11] Y. Sun, Z. Sun, and W. Chen, "The evolution of object detection methods," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108458, Jul. 2024, doi: 10.1016/j.engappai.2024.108458.

[12] K. Huang, M. Wen, C. Wang, and L. Ling, "T-SSD: A transformer-based single-stage multi-scale sampling object detector," in *Proceedings of 2023 the 13th International Workshop on Computer Science and Engineering*, WCSE, 2023. doi: 10.18178/wcse.2023.06.022.

[13] M. Chaman, A. E. Maliki, H. Dahou, and A. Hadjoudja, "Benchmarking YOLO-based deep learning models for real-time object detection in hybrid ADAS and intelligent transportation systems," *Results Eng.*, vol. 29, p. 108942, Mar. 2026, doi: 10.1016/j.rineng.2025.108942.

[14] J. M. Villarroel, L. C. De Jesus, J. Ancheta, H. Villaruel, and L. Samaniego, "Comparative analysis of YOLOv12 and RF-DETR models for coffee leaf disease detection using Roboflow," in *2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, Sep. 2025, pp. 1493–1494. doi: 10.1109/GCCE65946.2025.11274815.

[15] M. S. Aqilla, M. G. Abdurahman, A. B. Hendry, M. R. Arjasubrata, and M. D. Sulistiyo, "A comparative analysis of YOLOv12 and RFDETR for weapon detection in CCTV footage," in *2025 5th International Conference of Science and Information Technology in Smart Administration (ICSINTESA)*, Nov. 2025, pp. 35–40. doi: 10.1109/ICSINTESA68165.2025.11413754.

[16] S. Zanniko, J. Cahyo, A. A. S. Gunawan, and R. C. Pradana, "Comparative analysis of RF-DETR and YOLOv12 in breast cancer detection and classification," in *2025 International Conference on Information Management and Technology (ICIMTech)*, Aug. 2025, pp. 246–251. doi: 10.1109/ICIMTech67074.2025.11265111.

[17] T. Sabrina, I. Damayanti, I. Mahardika, M. R. Arjasubrata, and M. D. Sulistiyo, "Comparative analysis of CNN and transformer models for cigarette and e-cigarette detection," in *2025 8th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Dec. 2025, pp. 345–350. doi: 10.1109/ISRITI68345.2025.11393399.

[18] C. L. Buana, G. F. Shidik, M. N. Ubaidillah, Y. R. Grafer, F. A. Kristiyan, and E. J. Kusuma, "Comparison of YOLO and RF-DETR performance in detecting personal protective equipment in construction environments using tokenization," in *2025 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Sep. 2025, pp. 151–157. doi: 10.1109/ISemantic67418.2025.11292414.

[19] U. Aymon, N. S. Kamarudin, and A. F. Ab. Nasir, "Facial expression recognition with YOLOv11 and YOLOv12: A comparative study," in *2025 IEEE 9th International Conference on Software Engineering & Computer Systems (ICSECS)*, Oct. 2025, pp. 18–23. doi: 10.1109/ICSECS65227.2025.11279248.

[20] Q. Tang, Y. Li, Y. Cai, X. Peng, and X. Liu, "Face detection based on DF-Net," *Electronics*, vol. 12, no. 19, p. 4021, Sep. 2023, doi: 10.3390/electronics12194021.

[21] M. G. Ragab *et al.*, "A comprehensive systematic review of YOLO for medical object detection (2018 to 2023)," *IEEE Access*, vol. 12, pp. 57815–57836, 2024, doi: 10.1109/ACCESS.2024.3386826.

[22] Y. Ji *et al.*, "Transmission line defect detection algorithm based on improved YOLOv12," *Electronics*, vol. 14, no. 12, p. 2432, Jun. 2025, doi: 10.3390/electronics14122432.

[23] L. T. Ramos and A. D. Sappa, "A comprehensive analysis of YOLO architectures for tomato leaf disease identification," *Sci. Rep.*, vol. 15, no. 1, p. 26890, Jul. 2025, doi: 10.1038/s41598-025-11064-0.

[24] N. Ghosh and G. Mandal, "Classification of canine dermatological diseases using YOLOv12 with R-ELAN," in *2025 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Nov. 2025, pp. 542–546. doi: 10.1109/ICCIKE67021.2025.11318252.

[25] A. S. Silva, F. H. A. Moraes Neto, P. H. Ferreira, and D. B. Costa, "CNN-based YOLOv12 for damage assessment in residential roofs," *IEEE Access*, vol. 13, pp. 193311–193322, 2025, doi: 10.1109/ACCESS.2025.3629630.

[26] N. Deluxni, P. Sudhakaran, M. Alsafyani, and A. Yousef, "Underwater debris segmentation using improved YOLOv12s with recursive efficient layer aggregation and FlashAttention for autonomous underwater vehicle," *IEEE Access*, vol. 13, pp. 200239–200252, 2025, doi: 10.1109/ACCESS.2025.3636283.

[27] A. A. Murat and M. S. Kiran, "A comprehensive review on YOLO versions for object detection," *Eng. Sci. Technol. Int. J.*, vol. 70, p. 102161, Oct. 2025, doi: 10.1016/j.jestch.2025.102161.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[28] J. Terven, D.-M. Córdova-Esparza, and J.-A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, Nov. 2023, doi: 10.3390/make5040083.