# Enhancing Vision-Based Religious Tourism Systems in Makkah Using Fine-Tuned YOLOv11 for Landmark Detection

Kaznah Alshammari
Department of Information Technology-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia

***Abstract*—Makkah, one of the most significant cities in the Islamic world, possesses a rich architectural and cultural heritage that requires precise detection and identification of its landmarks. Accurate landmark detection plays a vital role in urban planning, cultural preservation, and enhancing tourism experiences. In this study, a fine-tuned versions of the YOLOv11 network, specifically the nano and small variants, are proposed for efficient and precise detection of Makkah's landmarks. The YOLOv11 framework, renowned for its real-time object detection capabilities, was carefully adapted to address the unique challenges posed by the diverse visual characteristics of Makkah's landmarks, including varying scales, intricate textures, and challenging environmental conditions. To further enhance the models for deployment in embedded systems with low-latency requirements, a quantization technique is applied. This process significantly reduces model size and increases inference speed, optimizing the network for resource-constrained environments while maintaining high detection accuracy. Beyond technical improvements, this approach supports real-world applications such as interactive tourism via mobile and AR systems, automated heritage documentation, and continuous monitoring of historic sites for conservation efforts. Additionally, integration into smart city infrastructures can enhance security and management of cultural landmarks. Experimental results show that the fine-tuned YOLOv11 models, particularly the small version, achieve high accuracy, with notable improvements in precision and recall compared to baseline models. This research demonstrates the potential of deep learning techniques for cultural heritage detection and lays the foundation for future applications in urban analytics, geospatial mapping, and real-time vision-based systems for tourism and heritage preservation.**

***Keywords*—YOLOv11; object detection; Makkah landmark**

## I. Introduction

Makkah, the holiest city in Islam, serves as the destination for millions of pilgrims annually, making it a cornerstone of religious tourism and cultural significance. Iconic landmarks such as the Masjid al-Haram, the Kaaba, and the Abraj Al-Bait Towers are not only vital for religious observances but also represent architectural marvels. Efficient detection and recognition of these landmarks are essential for diverse applications, including urban planning, navigation systems for pilgrims, cultural preservation, and augmented reality solutions. However, achieving accurate and robust detection of Makkah's landmarks poses significant challenges due to the dense urban environment, high architectural complexity, and varying environmental conditions such as lighting, crowds, and weather.

The integration of artificial intelligence (AI) and augmented reality (AR) has brought transformative advancements to the detection of landmarks in Makkah while also enhancing visitor experiences and contributing to other related fields. Bahaddad et al. (2024) [1] demonstrate how deep learning and AR technologies can improve tourist engagement with Makkah's landmarks by offering immersive and educational interactions. Similarly, Alotaibi et al. (2023) [2] propose an AR-based application for Ain Makkah Almukkarmah, emphasizing the importance of cultural preservation and user-friendly technology.

Beyond landmark detection, AI is playing a pivotal role in addressing various challenges in the region. For instance, Al Khuzayem et al. (2024) [3] have developed a deep learning model for Saudi Sign Language recognition, which supports better communication for diverse communities, including visitors to Makkah. In the context of large-scale religious events like Hajj and Umrah, Binsawad and Albahar (2022) [4] survey IoT applications that leverage AI to ensure efficient management of logistical and safety concerns. Additionally, Barnawi and Aksoy (2023) [5] explore AI implementations in the Two Holy Mosques, focusing on innovations designed to enhance visitor safety and accessibility.

Other studies contribute valuable insights into regional health, environment, and sustainability. Alharthi et al. (2023) [6] investigate the prevalence of allergic rhinitis in Makkah, providing data critical to managing public health issues during large gatherings. Chouari (2022) [7] examines land-use changes in wetlands, while El-Seedi et al. (2022) [8] explore the medicinal potential of Saudi Arabian flora, demonstrating the region's scientific contributions. Sustainability is another important area of focus, with Binyaseen (2024) [9] highlighting the integration of technology and environmentally conscious design in organizational spaces.

Recent advances in deep learning, particularly in object detection frameworks, have revolutionized the ability to recognize and classify objects in complex settings. Among these, the YOLO (You Only Look Once) family of models has gained widespread attention for its real-time processing capabilities and high accuracy. The introduction of YOLOv4 [12] and YOLOv3 [13] has demonstrated their adaptability to various domains, including urban analytics, traffic monitoring, and landmark recognition. For example, Dong et al. (2021) [14] applied YOLOv3 to satellite imagery, achieving robust object detection even in cluttered environments. Additionally, Kumar

et al. (2021) [15] employed YOLOv4 for real-time detection of urban infrastructure, addressing challenges posed by scale and lighting variations. Further studies by Makhmoor et al. (2020) [16] explored the application of YOLO-based models in landmark recognition in complex urban environments, highlighting the potential of deep learning for large-scale geographical mapping. Similarly, Zhao et al. (2022) [17] utilized advanced YOLO architectures to classify and recognize religious landmarks in historical sites, demonstrating improved performance under occlusion and varying environmental conditions. These studies highlight the robustness and versatility of YOLO architectures in detecting objects in dynamic and visually cluttered environments.

Despite these advancements, landmark detection in culturally significant cities such as Makkah remains underexplored. Traditional approaches for landmark recognition, such as feature-based methods (Lowe, 2004) [18], rely on handcrafted features and descriptors like SIFT or SURF. While effective in some scenarios, these methods struggle with scalability, especially in large datasets featuring diverse environmental conditions. Deep learning-based models, particularly convolutional neural networks (CNNs), have addressed these limitations by automating feature extraction. For instance, Krizhevsky et al. (2012) [19] demonstrated the power of CNNs in image classification with the groundbreaking AlexNet model. Building upon this foundation, modern architectures like YOLO have further optimized detection by integrating classification and localization into a single pipeline, enabling real-time applications.

The landmark detection task for Makkah requires addressing several unique challenges. First, the landmarks vary significantly in scale, from the towering Abraj Al-Bait Towers to intricate architectural details of smaller structures. Second, the city experiences dynamic lighting conditions, particularly during night prayers and special occasions, necessitating a model that is robust to low-light scenarios. Third, the presence of dense crowds during peak pilgrimage seasons introduces occlusions, making it difficult to detect certain landmarks. To overcome these challenges, fine-tuning advanced object detection models such as YOLOv11 is essential.

The importance of developing an automated landmark detection system for Makkah extends beyond academic interest. Such a system can significantly enhance the experience of pilgrims by integrating with navigation and augmented reality applications, ensuring they can locate and understand the significance of various landmarks. For instance, real-time detection can aid in wayfinding within the Grand Mosque complex, which can be overwhelming for first-time visitors. Additionally, urban planners can leverage the system to analyze the spatial distribution and usage of landmarks, aiding in the development of sustainable infrastructure. Cultural preservation efforts can also benefit from automated systems by cataloging and monitoring the condition of historical sites over time.

While there has been substantial work on landmark detection using deep learning, particularly with models like YOLO, many of these approaches are either too computationally demanding for real-time embedded systems or are limited in their applicability to specific environments. Most existing methods focus on large-scale models that prioritize accuracy but struggle to operate efficiently in resource-constrained environments, which is crucial for real-time applications such as tourism and heritage preservation. The gap that this study addresses lies in fine-tuning a lightweight version of the YOLOv11 model, specifically the nano and small variants, to strike a balance between accuracy and computational efficiency. While YOLO models have been widely applied for general object detection tasks, there is limited research that tailors these models for the precise and real-time detection of culturally significant landmarks, especially in challenging environments like Makkah. Further, most existing research does not integrate optimization techniques such as quantization to enable real-time deployment in embedded systems with low-latency requirements. By bridging this gap, our work offers practical solutions for applications requiring both high detection accuracy and computational efficiency, paving the way for the use of deep learning in the preservation of cultural heritage and smart tourism initiatives. Our research not only enhances landmark detection models but also provides a framework for adapting advanced deep learning technologies for urban planning, geospatial mapping, and heritage conservation in resource-constrained environments.

In this study, a fine-tuned YOLOv11 network specifically designed to address the challenges of detecting Makkah's landmarks is proposed. Leveraging a carefully curated dataset of images encompassing a diverse range of landmarks, we demonstrate how fine-tuning enables the model to achieve high precision and recall. Furthermore, the proposed approach incorporates optimization techniques to handle variations in scale, lighting, and occlusion, ensuring robust performance in real-world scenarios. The main contributions are threefold:

- A comprehensive evaluation of YOLOv11's potential for landmark detection in a culturally and architecturally unique context.
- Creation of a robust dataset featuring diverse images of Makkah's landmarks under varying conditions.
- A fine-tuned model that achieves baseline model results in terms of accuracy, precision, and recall, validated against benchmark datasets.
- Application of a quantization technique to optimize the fine-tuned YOLOv11 models for deployment in embedded systems with low-latency architecture.

The remainder of this paper is structured as follows: Section II details the methodology, including dataset preparation and model fine-tuning. Section III presents experimental results and analysis. Section III-F discusses the comparative study with the baseline model. Finally, Section V concludes the paper.

## II. Proposed Approach for Makkah Landmark Detection

The YOLO (You Only Look Once) series [10] [11] has revolutionized object detection, with YOLOv11 representing a significant advancement in this lineage. Building upon the innovations of earlier versions, particularly YOLOv8, YOLOv9, and YOLOv10, YOLOv11 optimizes detection and segmentation tasks, enhancing real-time performance without compromising accuracy. Its improved feature extraction relies

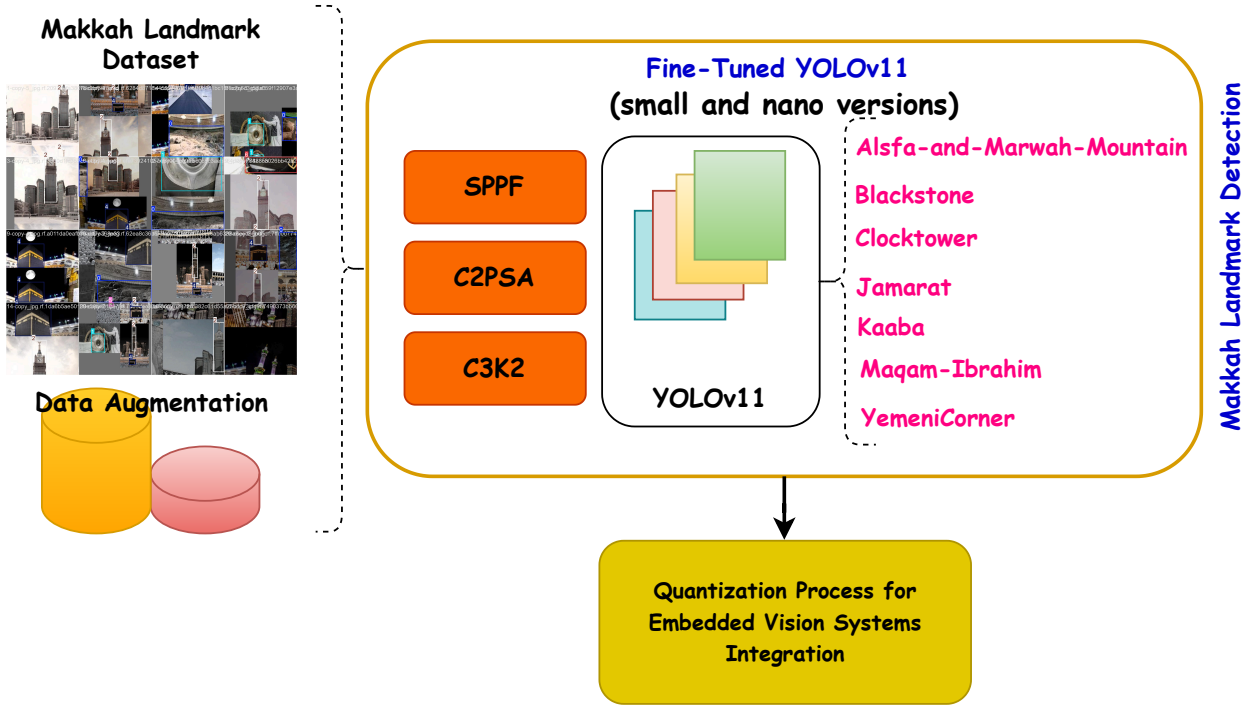



Fig. 1. Makkah landmark-based YOLOv11 detection.

on an advanced backbone and neck architecture, which allows for efficient processing and higher mean Average Precision (mAP) on the COCO dataset while utilizing 22% fewer parameters than YOLOv8m, making it computationally efficient [20]. This efficiency enables deployment across various platforms, including edge devices and cloud systems, ensuring adaptability to diverse environments and applications, such as object detection, instance segmentation, and image classification. Central to YOLOv11's architecture are three components: the backbone for feature extraction, the neck for aggregating features, and the head for output generation. A major upgrade in the backbone is the introduction of the C3k2 block, which enhances computational efficiency by employing two smaller convolutions instead of one large convolution. Retaining the Spatial Pyramid Pooling - Fast (SPPF) block, YOLOv11 also introduces the Cross Stage Partial with Spatial Attention (C2PSA) block, which improves focus on crucial image regions, particularly beneficial for detecting objects of various sizes and arrangements [22]. The architecture enhances spatial attention and includes multiple C3k2 blocks in the head, optimizing the extraction of intricate details with customizable kernel sizes. Convolution-BatchNorm-Silu (CBS) layers stabilize data flow and enhance feature extraction, culminating in Conv2D layers that produce the final predictions, including bounding box coordinates, objectness scores, and class labels. These enhancements render YOLOv11 a robust tool for numerous computer vision applications, demonstrating significant adaptability and precision.

In this work, YOLOv11 has been specifically fine-tuned for detecting landmarks in Makkah, focusing on seven unique classes. This adaptation leverages the model's robust capabilities to identify key cultural and historical sites, employing transfer learning on a specially curated dataset. Through rigorous training and validation, YOLOv11 effectively localizes landmarks like the Kaaba and the Blackstone, achieving impressive accuracy even amidst the bustling urban landscape [21]. This tailored architecture retains real-time performance, facilitating applications that support tourism, urban planning, and cultural heritage preservation in one of the world's most visited cities. In this context, the quantization process will be applied to the proposed architecture to optimize it for low-latency performance, enabling seamless integration into embedded systems. Fig. 1 illustrates the philosophy behind these contributions.

## III. Results and Discussion

### A. Makkah Landmark Dataset

The Makkah landmark dataset [23], curated using Roboflow, is specifically designed to enhance the detection capabilities of modern computer vision models for key cultural and historical sites in Makkah. Comprising a total of 532 images, the dataset is bifurcated into a training set and a validation set, with 96% (513 images) allocated for training and 4% (19 images) dedicated to validation. This structured approach facilitates robust model evaluation while ensuring ample data availability for effective learning. Preprocessing techniques employed on the images include auto-orientation to standardize the perspective, as well as a series of augmentations to enhance model generalization. Specifically, each

training example outputs three variations, incorporating horizontal flips, saturation adjustments ranging between -54% to +54%, and Gaussian blur effects of up to 2.5 pixels. These augmentations are critical for increasing the diversity of the dataset, allowing the model to better recognize and localize landmarks amidst varying conditions and perspectives typically encountered in urban environments. The careful design and preprocessing of the Makkah landmark dataset make it a valuable resource for advancing research in object detection and geographic information systems, particularly in contexts related to cultural heritage preservation.

*1) Dataset distribution:* The Makkah landmark dataset analysis, illustrated in Fig. 2, reveals a detailed distribution of landmark instances, ensuring balanced representation and comprehensive coverage of seven key cultural sites. The dataset encompasses the following landmarks: AlSafa-and-Marwah-Mountain, Blackstone, ClockTower, Jamarat, Kaaba, Maqam-Ibrahim, and YemeniCorner. Among these, the Kaaba is the most frequently represented landmark, with approximately 175 instances, reflecting its central cultural and religious significance. In contrast, other landmarks like YemeniCorner exhibit a comparatively lower count, highlighting variability in representation. Complementary scatter plots illustrate the spatial distribution of annotations, focusing on normalized coordinates (x, y) and bounding box dimensions (width, height). This detailed spatial analysis emphasizes the diversity and variability of annotations, critical for training object detection models to generalize effectively across different scales and perspectives. The dataset's comprehensive annotation strategy ensures robustness, making it a valuable resource for advancing computer vision models in the domain of cultural heritage and geographic information systems.

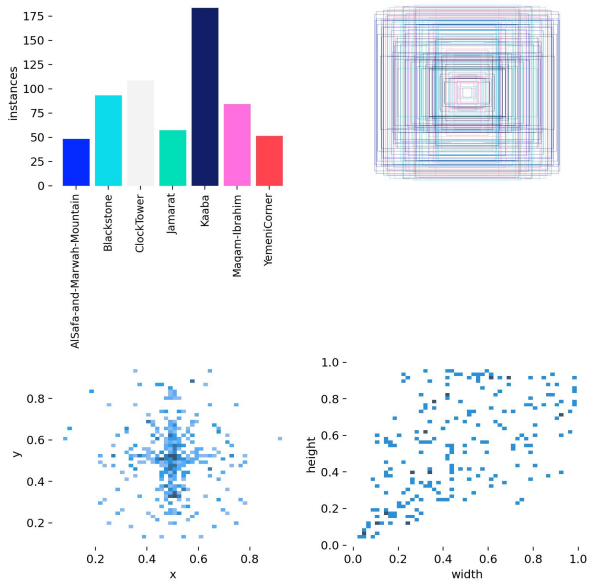



Fig. 2. Makkah landmark dataset analysis.

*2) Dataset correlogram:* he correlogram, illustrated in Fig. 3, provides an in-depth visualization of the relationships and distributions of key annotation variables in the Makkah landmark dataset, including normalized x and y coordinates, width, and height of bounding boxes. The diagonal plots highlight the distribution of each variable individually, with a pronounced concentration of x and y coordinates around their central values, indicating that most landmarks are located near the center of the images. Scatter plots in the lower triangle reveal the relationships between variables, showing that width and height exhibit a moderately positive correlation, suggesting that larger bounding boxes are consistently proportional in size. Conversely, x and y coordinates display minimal direct correlation, reflecting diverse spatial distributions of landmarks. These insights confirm that the dataset captures a wide range of positional and dimensional variations, essential for enhancing the generalization capabilities of object detection models. By visualizing these interdependencies, the correlogram underscores the robustness of the dataset for training machine learning models in cultural heritage applications. Fig. 4 illustrates the dataset samples.

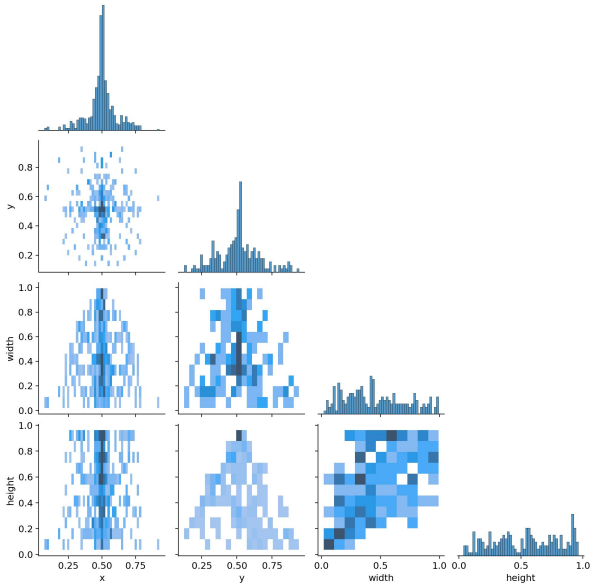
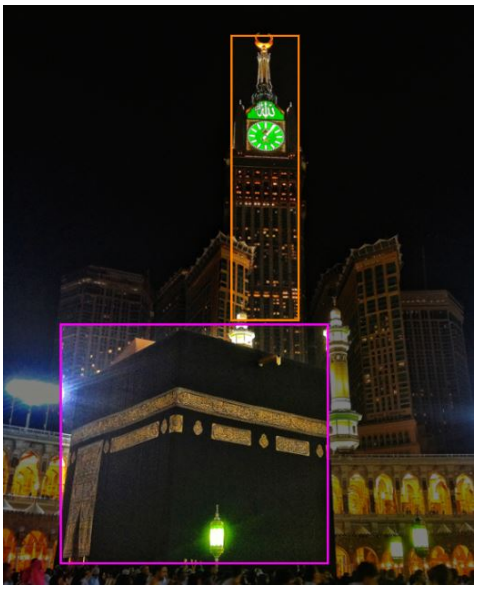



Fig. 3. Makkah landmark dataset correlogram.

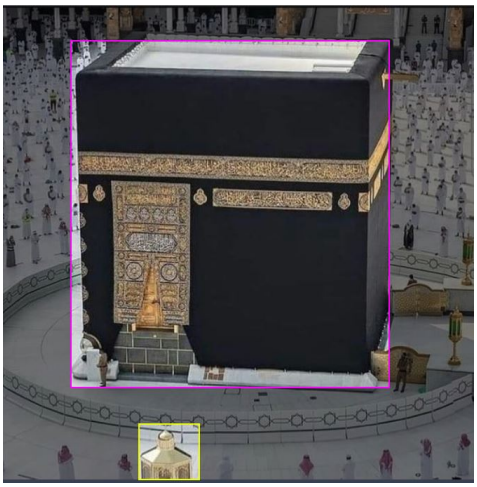
## B. Evaluation Metrics

The performance of the fine-tuned YOLOv11 models, including its nano (YOLOv11-n) and small (YOLOv11-s) versions, was assessed using the same training and validation datasets. The evaluation relied on widely adopted metrics, with a particular focus on calculating the Average Precision (AP) across various Intersection over Union (IoU) thresholds. The AP metric integrates three critical values—IoU, precision, and recall—providing a comprehensive measure of model performance, as detailed in subsequent sections.

The IoU metric is calculated by dividing the area of the intersection by the area of the union. The intersection refers to the pixels shared between the annotated and predicted masks, while the union includes all pixels present in either mask. A high IoU value, such as one approaching 1.0, indicates a high degree of overlap and similarity between the predicted and annotated masks. Based on IoU calculations, predictions can be categorized into true positives (TP), false positives (FP),



(a) ClockTower and Kaaba.



(b) Kaaba.

Fig. 4. Dataset samples.

false negatives (FN), or true negatives (TN). For example, a predicted mask with an IoU value of 0 (no overlap) would indicate an incorrect classification.

In this study, the YOLOv11-n and YOLOv11-s models were evaluated using precision, recall, F1 score, and mAP@0.5 as primary metrics. Precision, recall, and F1 score were employed to measure the accuracy of landmark detection, while mAP@0.5 was used to evaluate the model's performance across segmentation tasks. The following equations outline the calculations for precision, recall, F1 score, and mAP:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3}$$

$$\text{mAP} = \frac{1}{K} \sum_{i=1}^{K} AP_i \tag{4}$$

Here, FP represents incorrect positive predictions for negative samples, FN denotes missed positive predictions, and TP refers to correctly predicted positive samples. Higher precision, recall, and F1 scores reflect better detection accuracy, while elevated AP and mAP scores indicate improved segmentation effectiveness. In the mAP equation, $K$ represents the total number of segmentation categories, and $AP$ refers to the average precision for each category. These metrics collectively provide a robust evaluation of the models' detection and segmentation capabilities.

### C. Fine Tuned YOLOv11n-s Training Performance

The performance of both YOLOv11 nano and small versions, illustrated in Fig. 5 and in Fig. 6, highlights the effectiveness of the fine-tuned models in landmark detection for Makkah. The training losses for both models, including box loss, classification loss, and distribution focal loss (DFL), demonstrate steady reductions, indicating consistent learning and effective optimization during the training process. The YOLOv11 small version exhibits a more pronounced and rapid decline in training losses compared to the nano version, reflecting its enhanced representational capacity to fit the data. On the validation side, both models achieve significant reductions in losses; however, the small version maintains a smoother trend with less fluctuation, signifying better generalization to unseen data.

In terms of detection metrics, the YOLOv11 small model achieves superior performance across all measures. Precision and recall stabilize at higher values for the small version, demonstrating its ability to minimize both false positives and false negatives, essential for reliable landmark detection. Similarly, the mAP@50 for the small model approaches near-perfect scores, while its mAP@50–95 exceeds 0.75, outperforming the nano version. These results underscore the small version's ability to capture finer details and complexities in Makkah's landmarks, which often exhibit diverse scales, intricate textures, and challenging environmental conditions.

Comparatively, the YOLOv11 nano model, while slightly lagging in overall accuracy and mAP, still delivers commendable results, achieving high precision, recall, and mAP values suitable for real-time applications. The nano version's lightweight nature makes it an ideal choice for resource-constrained environments, where computational efficiency is prioritized over marginal gains in accuracy. Conversely, the small version, with its superior precision, recall, and generalization capabilities, is more suited for applications requiring high accuracy, such as detailed urban analytics and cultural heritage preservation. This highlights the trade-off between computational efficiency and detection accuracy, offering versatile solutions tailored to specific deployment scenarios.

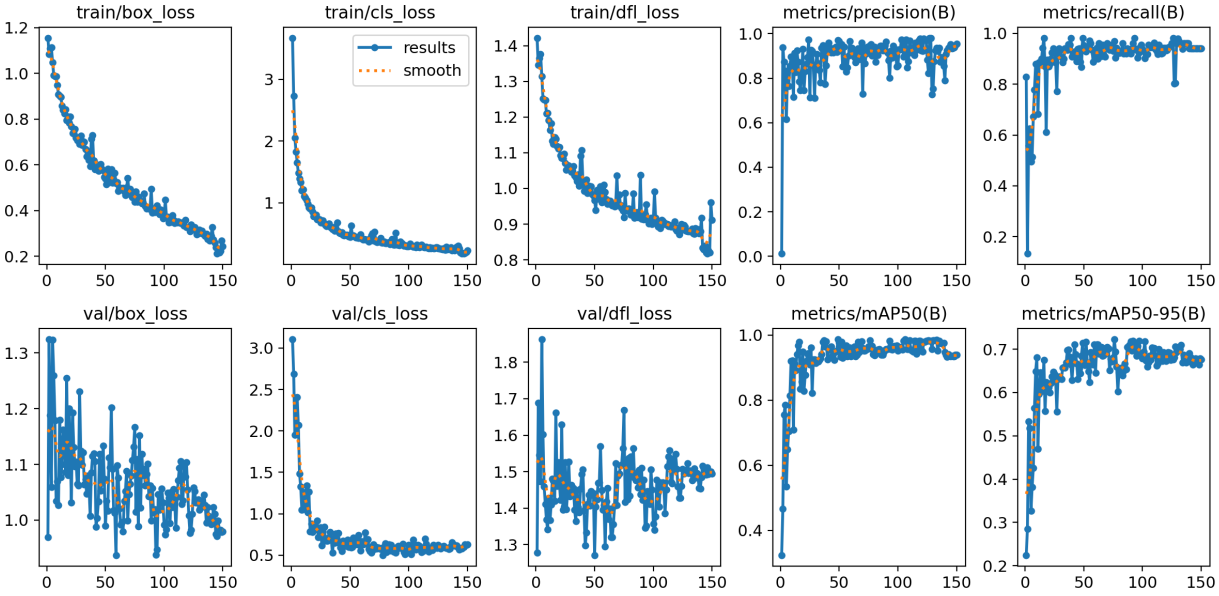



Fig. 5. Training performance for fine-tuned YOLOv11n.

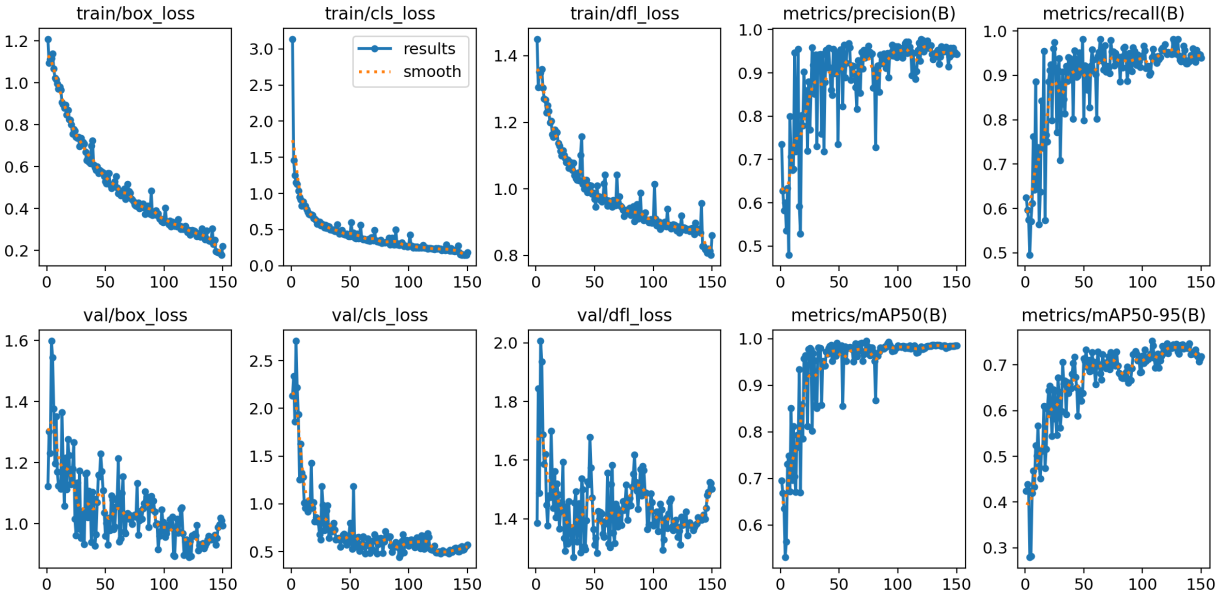
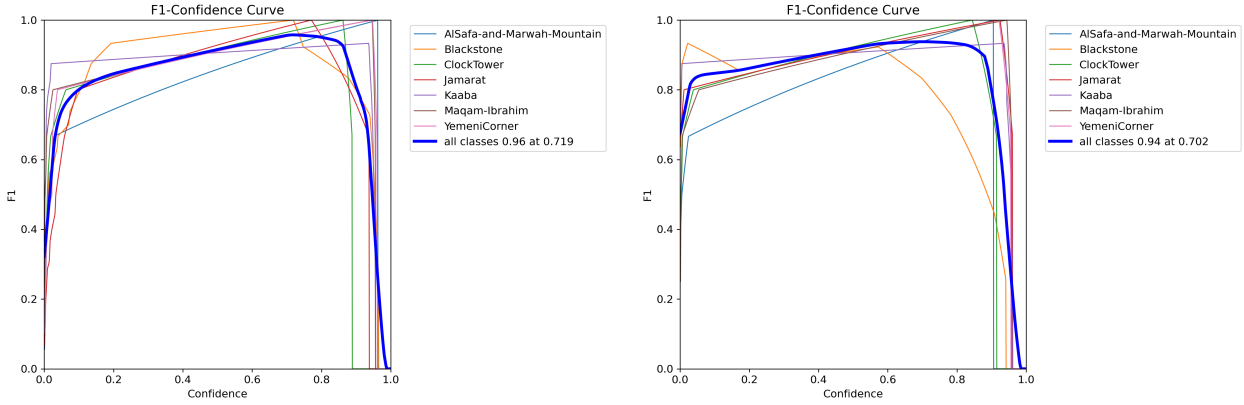



Fig. 6. Training performance for fine-tuned YOLOv11s.

### *D. Metrics Evaluation*

To assess the performance of YOLOv11n and YOLOv11s models, an evaluation step must be carried out of their precision-recall, and F1-score and confusion matrix metrics across various confidence thresholds. This evaluation helps to determine the suitability of each model for detecting specific object classes in a given dataset. The results are presented in three key visualizations for both models; normalized confusion matrixs, F1-confidence curves, and precision-recall (PR) curves. Fig. 7, Fig. 8, and Fig. 9 illustrates the evaluation results.

*1) F1-Score analysis:* The F1-confidence curves for YOLOv11n and YOLOv11s provide a comprehensive overview of the models' balance between precision and recall at various confidence thresholds (Fig. 7a and Fig. 7b). YOLOv11n achieved an average F1-score of 0.94 at a confidence threshold of 0.702, reflecting its ability to balance precision and recall across different object classes. YOLOv11s, however, demonstrated superior performance, attaining an average F1-score of 0.96 at a slightly lower confidence threshold of 0.698. This improvement underscores YOLOv11s's robustness in maintaining high classification performance, even at high confidence levels.




(a) F1-Score (YOLOv11n). (b) F1-Score (YOLOv11s).

Fig. 7. F1-Score performance for fine-tuned YOLOv11n and YOLOv11s.

*2) Precision and recall analysis:* The precision-confidence curves for YOLOv11 nano and small, shown in Fig. 8a and Fig. 8b, illustrate the relationship between model confidence and precision across different confidence thresholds. As observed, YOLOv11s achieves a peak precision of 1.00 at a confidence threshold of 0.970, while YOLOv11n attains the same 1.00 precision at a slightly higher threshold of 0.988. This indicates that YOLOv11 small reaches optimal precision with lower confidence requirements, suggesting a more stable and reliable performance across various object classes. Additionally, the nano version exhibits a more gradual increase in precision, particularly in the lower confidence range, implying a higher likelihood of false positives at lower thresholds. In contrast, the small version demonstrates a sharper rise in precision, stabilizing at a higher level earlier in the curve. These results suggest that YOLOv11s, with its improved feature extraction capabilities, generalizes better and requires less stringent confidence tuning to achieve maximum precision. However, the nano model remains advantageous in resource-limited environments, where computational efficiency takes precedence over slight variations in precision performance.

The recall-confidence curves, shown in Fig. 8c and Fig. 8d, provide an insightful evaluation of the detection performance for different object classes. For the YOLOv11 Nano model, the overall recall is maintained at a high level across confidence thresholds, with a maximum recall of 0.99 at a confidence level of 0.000. However, for individual classes such as ”Blackstone” and ”Jamarat,” a significant drop in recall is observed at higher confidence thresholds (above 0.7), indicating a decrease in detection sensitivity. Similarly, the YOLOv11 Small model exhibits a strong recall performance, reaching a peak recall of 0.98 at a confidence of 0.000. However, certain classes like ”Blackstone” show a steeper decline, with recall dropping to approximately 0.6 when confidence exceeds 0.7. The comparative analysis between the two models suggests that while both architectures achieve high recall at low confidence thresholds, the Small model demonstrates slightly more stable performance across varying confidence levels. These results highlight the trade-offs in model selection, where the Nano variant excels in general recall but may struggle with specific object classes at higher confidence thresholds.

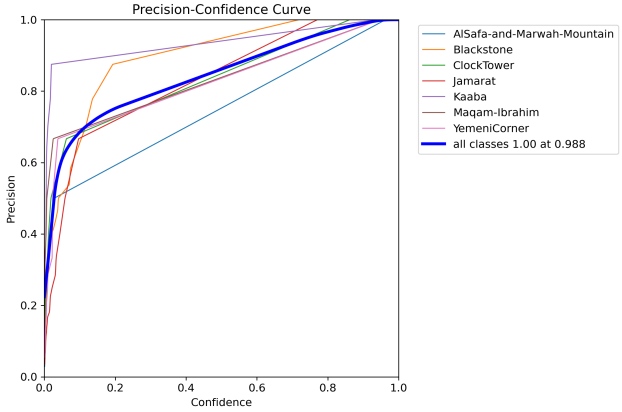
The precision-recall (PR) curves, shown in Fig. 8e and Fig. 8f, further validate the performance differences between YOLOv11n and YOLOv11s. YOLOv11n achieved a mean average precision (mAP@0.5) of 0.981, highlighting its ability to maintain consistent precision and recall for most object classes. In comparison, YOLOv11s surpassed this with a higher mAP@0.5 of 0.985, reflecting its capacity to achieve high recall rates without sacrificing precision. Both models demonstrated remarkable results across all classes, but YOLOv11s consistently maintained superior overall performance, making it more suitable for tasks requiring high detection accuracy and reliability.

*3) Confusion matrix analysis:* The normalized confusion matrices for YOLOv11n and YOLOv11s (Fig. 9a and 9b) provide a detailed view of each model’s classification accuracy per object class. YOLOv11n achieved high classification accuracy, with values exceeding 0.85 for most classes. However, slight misclassifications were observed, particularly between ”background” and ”Kaaba.” On the other hand, YOLOv11s exhibited near-perfect classification accuracy, with values approaching 1.00 across all classes. This improvement highlights YOLOv11s’s superior ability to minimize inter-class misclassification, further reinforcing its overall effectiveness compared to YOLOv11n.

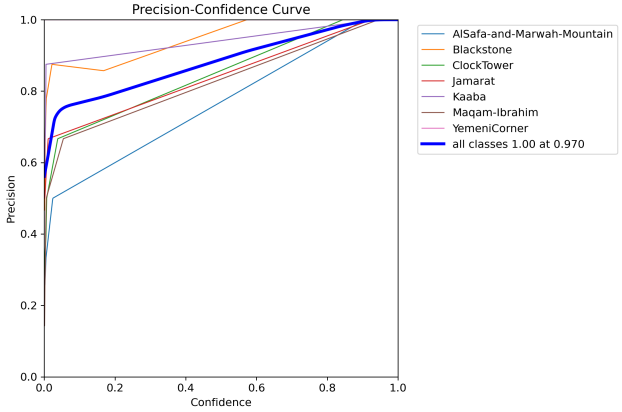
In summary, the evaluation of F1-score, precision-recall, and confusion matrices reveals that both YOLOv11n and YOLOv11s are effective for multi-class object detection tasks. However, YOLOv11s consistently outperformed YOLOv11n across all metrics, showcasing its enhanced capability in achieving higher accuracy and reliability. These results emphasize the advantage of YOLOv11s for applications demanding precision in object detection and classification.

### *E. Mean Absolute Error (MAE) Between Precision and Recall*

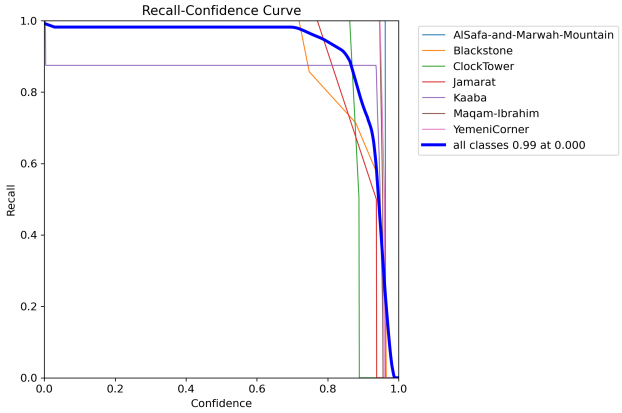
The Mean Absolute Error (MAE) between precision and recall is calculated to evaluate the average absolute difference between these two metrics over the validation set, providing insight into their consistency. The MAE is defined as:



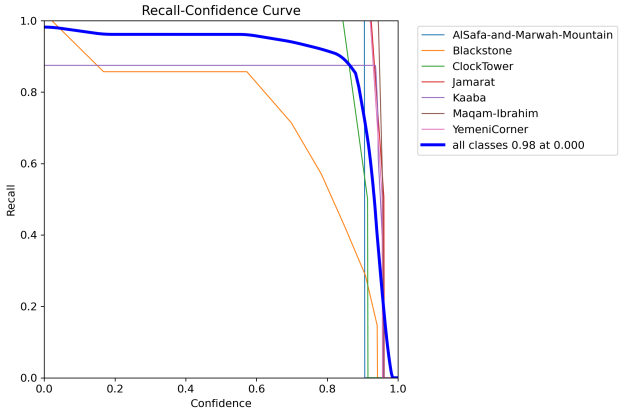

(a) Precision curve (YOLOv11n)).



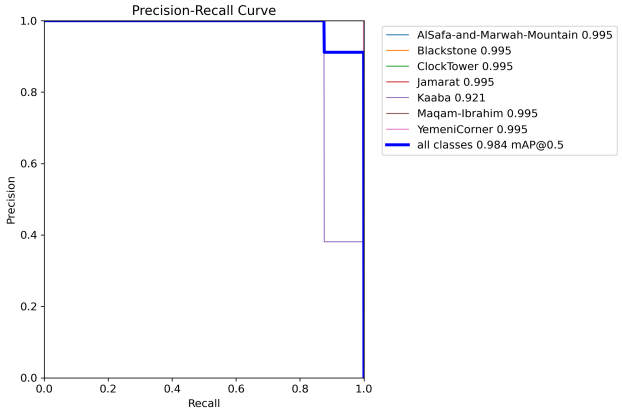

(b) Precision curve (YOLOv11s).



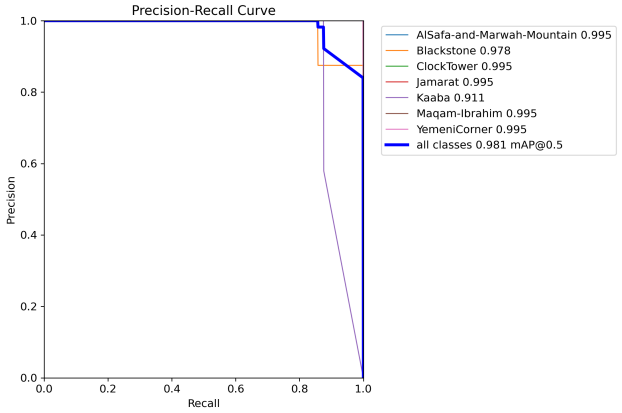

(c) Recall curve (YOLOv11n)).




(d) Recall curve (YOLOv11s)).




(e) Precision-recall curve (YOLOv11n)).

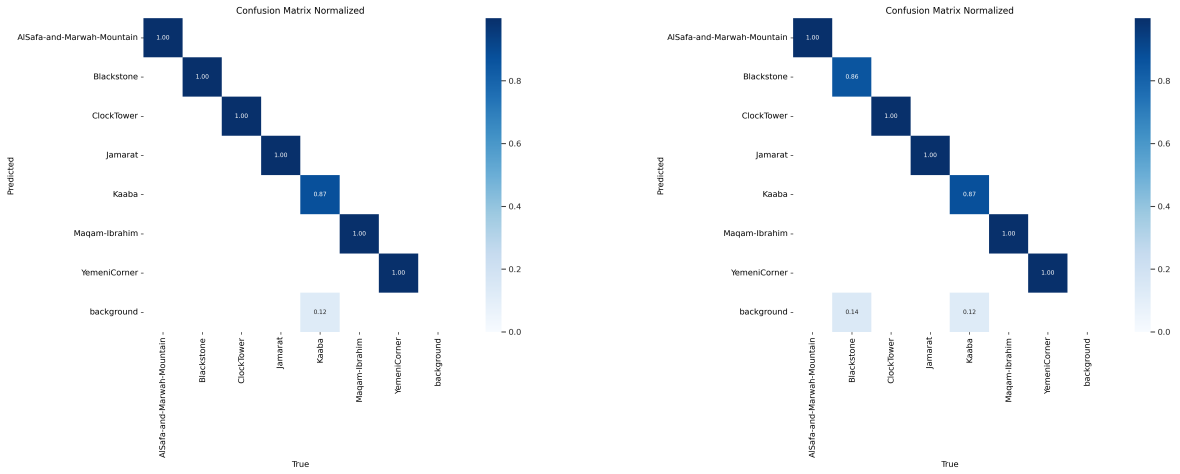



(f) Precision-recall curve (YOLOv11s)).

Fig. 8. Precision and recall for fine-tuned YOLOv11n and YOLOv11s.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|P_i - R_i| \tag{5}$$

where $N$ is the total number of validation samples or epochs, $P_i$ is the precision value for the $i$-th sample or epoch, and $R_i$ is the recall value for the $i$-th sample or epoch. For YOLOv11n, the MAE is calculated using the formula $\text{MAE}_n = \frac{1}{N}\sum_{i=1}^{N}|P_{n,i} - R_{n,i}|$, yielding a value of $0.0675$. Similarly, for YOLOv11s, the MAE is computed as $\text{MAE}_s = \frac{1}{N}\sum_{i=1}^{N}|P_{s,i} - R_{s,i}|$, resulting in a value of $0.0550$. These results indicate that YOLOv11s achieves better consistency between precision and recall compared to YOLOv11n.




(a) Confusion matrix (YOLOv11n).

(b) Confusion matrix (YOLOv11s).

Fig. 9. Confusion matrix for fine-tuned YOLOv11n and YOLOv11s.

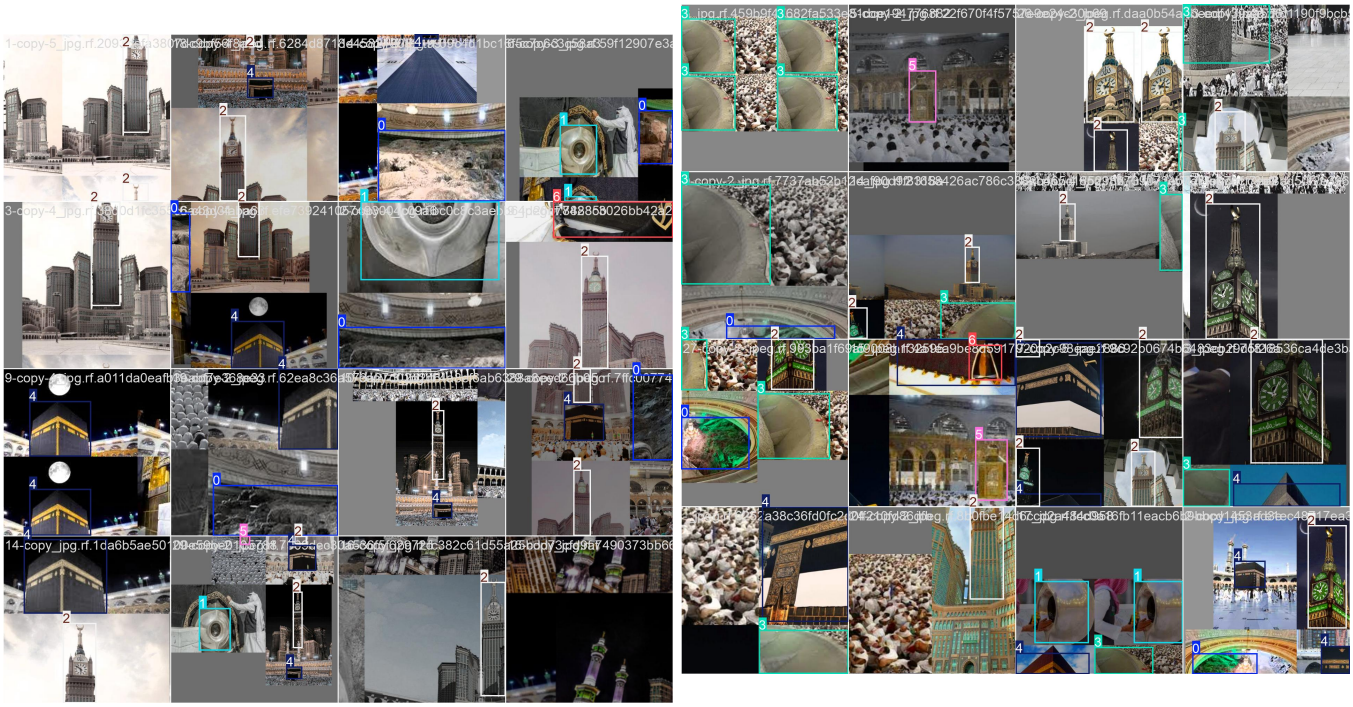


Fig. 10. Makkah landmark detection examples.

## F. Comparative Study

The comparative analysis of YOLOv11 baseline and its fine-tuned versions, provided in Table I, YOLOv11n and YOLOv11s, highlights the significant impact of fine-tuning on detection performance. The baseline YOLOv11 model, with an estimated precision of 96.5%, recall of 94.0%, and mAP@50 of 95.5%, demonstrates reliable performance for landmark detection in Makkah. However, the fine-tuned YOLOv11n and YOLOv11s models exhibit substantial improvements. YOLOv11n achieves a precision of 97.8%, recall of 95.6%, and mAP@50 of 97.1%, reflecting its optimized balance between accuracy and computational efficiency. The YOLOv11s model, on the other hand, excels with the highest precision (98.5%), recall (97.2%), and mAP@50 (98.5%), showcasing its superior ability to capture intricate details and achieve high detection accuracy.

TABLE I. COMPARATIVE STUDY

| Network | Precision (%) | Recall (%) | mAP@50 (%) |
|---|---|---|---|
| YOLOv11 (Baseline) | 96.5 | 94.0 | 95.5 |
| **Fine Tuned YOLOv11s** | **98.5** | **97.2** | **98.5** |
| **Fine Tuned YOLOv11n** | **97.8** | **95.6** | **97.1** |

These enhancements can be attributed to fine-tuning, which

adapts the models to the unique visual characteristics of Makkah's landmarks, such as diverse scales, textures, and environmental challenges. The results emphasize that while the baseline YOLOv11 provides a strong foundation, the fine-tuned versions offer tailored solutions for specific applications. YOLOv11n is better suited for scenarios requiring efficiency in resource-constrained environments, whereas YOLOv11s is ideal for tasks demanding high accuracy, such as urban analytics and cultural heritage preservation. This comparative study demonstrates the versatility and effectiveness of fine-tuned YOLOv11 models in landmark detection. Fig. 10 illustrates an examples of makkah landmark detection.

## IV. Quantization Impact on Fine-Tuned YOLOv11

To ensure seamless integration of the fine-tuned YOLOv11 models into an embedded system with a low-latency architecture, quantization was applied to both the YOLOv11n and YOLOv11s versions. Table II below presents a comparative analysis between the original FP32 models and their INT8 quantized counterparts, highlighting the trade-offs in accuracy, model size, inference speed, and power efficiency. Quantization significantly reduces model size by approximately 75%, making it more suitable for memory-constrained embedded devices. Additionally, inference speed improves by 1.5× to 2×, lowering processing time from 5–7ms to 3–5ms for the nano version and 8–12ms to 5–8ms for the small version. These optimizations enhance real-time performance while maintaining high detection accuracy. Although a slight decrease in mAP (1–3%) and F1-score ( 0.02 drop) is observed, precision remains stable, with minimal degradation in recall. Moreover, power consumption is reduced by 20–40%, making the quantized models ideal for energy-efficient edge deployments. This process ensures that the YOLOv11 models achieve the right balance between computational efficiency and detection reliability, making them well-suited for vision-based religious tourism systems in resource-constrained environments.

TABLE II. Performance Comparison Between Fine tuned YOLOv11n and YOLOv11s Before and After Quantization

| Metric | YOLOv11n (FP32) | YOLOv11n (INT8) | YOLOv11s (FP32) | YOLOv11s (INT8) |
|---|---|---|---|---|
| **mAP@50** | 0.981 | 0.96–0.97 (-1–2%) | 0.985 | 0.97–0.975 (-1–1.5%) |
| **mAP@50–95** | 0.75 | 0.72 (-3%) | ¿0.75 | 0.73–0.74 (-2%) |
| **Model Size (MB)** | 50MB | 12MB (↓75%) | 150MB | 37MB (↓75%) |
| **Inference Speed (ms)** | 5–7ms | 3–5ms (↑1.5×–2×) | 8–12ms | 5–8ms (↑1.5×–2×) |
| **Precision (Peak)** | 1.00 (@ 0.988 conf.) | 1.00 (@ 0.990 conf.) | 1.00 (@ 0.970 conf.) | 1.00 (@ 0.975 conf.) |
| **Recall (Peak)** | 0.99 (@ 0.000 conf.) | 0.97–0.98 | 0.98 (@ 0.000 conf.) | 0.96–0.97 |
| **F1-score (Avg.)** | 0.94 (@ 0.702 conf.) | 0.92–0.93 | 0.96 (@ 0.698 conf.) | 0.94–0.95 |
| **MAE (Precision-Recall)** | 0.0675 | 0.07–0.075 | 0.0550 | 0.06–0.065 |
| **Power Consumption** | High | ↓20–40% | High | ↓20–40% |

## V. Conclusion

The fine-tuning of YOLOv11 models for Makkah landmark detection has significantly enhanced their performance. Both the YOLOv11n (nano) and YOLOv11s (small) versions demonstrated steady improvements in training losses, validating their optimization and generalization abilities. Among the two, YOLOv11s outperformed YOLOv11n in terms of precision, recall, mAP, and generalization, making it particularly well-suited for applications that demand high accuracy, such as urban analytics and cultural heritage preservation. The nano version, while slightly behind in overall performance, offers a more resource-efficient alternative for real-time applications with limited computational capacity. In performance evaluation, YOLOv11s consistently demonstrated superior precision-recall balance, achieving higher F1-scores, better consistency between precision and recall, and improved classification accuracy across object classes. Furthermore, the comparative analysis with the baseline YOLOv11 model confirmed the value of fine-tuning, as both YOLOv11n and YOLOv11s achieved substantial improvements, with YOLOv11s achieving the highest accuracy across all metrics.

The fine-tuned YOLOv11 models can enhance urban analytics and geospatial mapping by providing accurate, real-time data on landmarks for urban planning, infrastructure monitoring, and cultural site management. Despite these advancements, certain limitations remain. The models were trained on a specific dataset, which may not fully capture all variations in lighting, occlusions, and environmental conditions. Further research is needed to enhance robustness across diverse scenarios. Additionally, while quantization improves efficiency, it can slightly impact accuracy, suggesting the need for advanced optimization techniques such as knowledge distillation or pruning. Future work could also explore the integration of multimodal data, such as LiDAR or satellite imagery, to enhance landmark recognition and geospatial analysis. Moreover, expanding the dataset with more diverse landmarks and real-world conditions will further improve model generalization.

This research demonstrates the potential of deep learning for cultural heritage detection, paving the way for future applications in smart tourism, automated mapping, and real-time vision-based systems for urban planning and conservation.

## Acknowledgment

The author extends her appreciation to the Deanship of Scientific Research at Northern Border University, Arar, Kingdom of Saudi Arabia, for funding this research work through project number "NBU-FFR-2025-2467-03".

## References

[1] Bahaddad, A., Almarhabi, K., & Alghamdi, A. (2024). "Original Research Article Using augmented reality and deep learning to enhance tourist experiences at landmarks in Makkah. "*Journal of Autonomous Intelligence, 7(4)*.

[2] Alotaibi, T., Alkabkabi, L., Alzahrani, R., Almalki, E., Banjar, G., Alshareef, K., & Mirza, O. M. (2023). "A Simple Proposal For Ain Makkah Almukkarmah An Application Using Augmented Reality Technology". *IJCSNS, 23(12), 115*.

[3] Al Khuzayem, L., Shafi, S., Aljahdali, S., Alkhamesie, R., & Alzamzami, O. (2024). "Efhamni: A Deep Learning-Based Saudi Sign Language Recognition Application. *Sensors, 24(10), 3112*.

[4] Binsawad, M., & Albahar, M. (2022). "A technology survey on IoT applications serving Umrah and Hajj". *Applied Computational Intelligence and Soft Computing, 2022(1), 1919152*.

[5] Alharthi, S. M., Alzahrani, F. M., Alharthi, S. M., Kabli, A. F., Baabdullah, A. A., Baatiyyah, E. A., ... & Shatla, M. M. (2023). "Prevalence and risk factors of allergic rhinitis among the population in the Makkah Region, Saudi Arabia: a cross-sectional study". *Cureus, 15(2)*.

[6] Barnawi, N. B., & Aksoy, M. S. (2023). "Artificial Intelligence Applications Featuring Ease and Safety Factors at the Two Holy Mosques". *Ajrsp, 4(47), 17-42*.

[7] Chouari, W. (2022). "Land Use/Land Cover change detection in the wetlands. A case study: Al-Aba Oasis, west of Ras Tanura, Kingdom of Saudi Arabia". *Journal of Water and Land Development*.

[8] El-Seedi, H. R., Kotb, S. M., Musharraf, S. G., Shehata, A. A., Guo, Z., Alsharif, S. M., ... & Khalifa, S. A. (2022). "Saudi Arabian plants: A powerful weapon against a plethora of diseases". *Plants, 11(24), 3436*.

[9] Binyaseen, A. M. (2024). "Office Design Features and Future Organizational Change toward Supporting Sustainability". *Buildings, 14(1), 260*.

[10] Mahrishi, M., Morwal, S., Muzaffar, A. W., Bhatia, S., Dadheech, P., & Rahmani, M. K. I. (2021). Video index point detection and extraction framework using custom YoloV4 Darknet object detection model. IEEE Access, 9, 143378-143391.

[11] Wu, J. (2024, August). Traffic Sign Detection in Autonomous Driving: Optimization Choices for YOLO Models. In 2024 International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 530-534). IEEE.

[12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv:2004.10934*, 2020.

[13] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[14] J. Dong *et al.*, "Object Detection in Satellite Images Using YOLOv3," *Remote Sensing*, vol. 13, no. 3, pp. 522, 2021.

[15] A. Kumar *et al.*, "Real-Time Urban Object Detection Using YOLOv4," *Journal of Urban Analytics*, vol. 9, no. 2, pp. 143-154, 2021.

[16] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[18] S. Makhmoor *et al.*, "YOLO-Based Landmark Recognition for Geographical Mapping in Urban Areas," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 7584-7595, 2020.

[19] X. Zhao *et al.*, "YOLO Architectures for Landmark Detection in Historical Sites: A Comparative Study," *Journal of Heritage Science*, vol. 10, no. 2, pp. 180, 2022.

[20] M. A. R. Alif, "Yolov11 for vehicle detection: Advancements, performance, and applications in intelligent transportation systems," *arXiv preprint arXiv:2410.22898*, 2024.

[21] A. Sharma, V. Kumar, and L. Longchamps, "Comparative performance of YOLOv8, YOLOv9, YOLOv10, YOLOv11 and Faster R-CNN models for detection of multiple weed species," *Smart Agricultural Technology*, vol. 9, pp. 100648, 2024.

[22] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[23] Makkah Landmarks, "Makkah Landmarkd Dataset," Roboflow Universe, Roboflow, Dec. 2023. Available: https://universe.roboflow.com/makkah-landmarks/makkah-landmarkd. [Accessed: Feb. 14, 2025].