Towards Trustworthy Defense AI: Real-Time Military Asset Detection with On-Demand Explainable YOLOv8

Omaia Al-Omari
Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
oalomari@psu.edu.sa

Awad Alyousef
Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
aalyousef@psu.edu.sa (corresponding author)

Suliman Mohamed Fati
College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
sgaber@psu.edu.sa

Esam Othman
Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia
eothman@psu.edu.sa

Muhammad Rashid Naeem
Department of Software Engineering, Prince Sultan University, Riyadh, Saudi Arabia
mnaeem@psu.edu.sa

Received: 30 December 2025 | Revised: 22 January 2026, 8 February 2026, 18 February 2026, and 22 February 2026 | Accepted: 25 February 2026

 DOI: https://doi.org/10.48084/etasr.17259

ABSTRACT

This paper introduces Digital Shield, an interpretable deep learning framework designed for real-time military asset detection in Saudi aerial imagery. Digital Shield, a typical answer to interpretable DL models for the detection of military assets—merging YOLOv8l and SHAP to maximize high-level and transparent support for decision-making. Digital Shield is trained on the Military Assets Dataset (12 classes) after an augmentation scheme to address environmental disturbances such as desert glare. Experimental results indicate that Digital Shield outperforms both baseline models (YOLOv5 and Faster R-CNN), boasting mAP@0.5 of 92.3% and a real-time inference speed at full blast, equaling 35 fps. The proposed framework operates in two complementary modes: a continuous real-time detection pipeline and an on-demand (triggered) explainability module that is activated only during high-risk detections or upon operator request. With on-demand visual explanations, the proposed framework enhances transparency and supports operator verification, contributing to the accountable deployment of automated defense detection systems.

Keywords-SHAP; YOLOv8; Explainable AI (XAI); UAV surveillance; real-time object detection; military asset detection

I. INTRODUCTION

In modern defense systems, the need for rapid and accurate detection of military assets from aerial imagery has become the norm in strategic planning, threat assessment, and situation awareness [1]. Such timely detection is critical in providing early warnings, efficient resource allocation, and mission-

critical decision support for border surveillance and counter-intrusion operations, especially in geographically large regions such as the Kingdom of Saudi Arabia. Traditional approaches in image processing, such as statistical models, have shown poor generalization performance in varying environmental conditions, complex object morphology, and varied backgrounds, as commonly found in military scenes [2]. Deep learning methods have been shown to outperform traditional ones in terms of accuracy in various computer vision tasks [3]. In particular, the YOLO series has proven high accuracy in object detection with the additional benefit of real-time processing, which is critical in defense surveillance and UAV-based reconnaissance operations [4]. YOLOv8 has shown significant improvements in the reduction of detection time, which is critical in real-time processing, as required in defense operations [5]. Recent works have proven the robustness of the YOLOv8l model in autonomous detection under adverse weather conditions, highlighting its superiority in feature extraction under such conditions [6]. YOLOv8 introduces an improved backbone and detection head that outperform prior YOLO variants on multiple benchmarks [7]. Highly advanced object detection models have been engineered, including Transformer-based detectors, e.g., DETR, and the newer versions of the YOLO family. These State-Of-The-Art (SOTA) methods have taken accuracy to new levels on test datasets such as COCO. Most SOTA methods are largely geared towards improving performance, sometimes with trade-offs that leave an Achilles heel for faithful and effective deployment arrangements. Although modern military infrastructures can leverage superior hardware acceleration (e.g., FPGAs and high-performance clusters) to deploy computation-heavy architectures such as Transformers (e.g., DETR) and YOLO with minimal latency, a critical paradox remains. Although these models achieve superior accuracy benchmarks, they function as 'black boxes' whose internal reasoning is opaque. Consequently, the availability of raw computational power cannot bridge the trust gap in high-stakes environments, where explainability is a prerequisite. Since these models are black boxes, military organizations cannot use them in operations that require transparent decision-making processes [8].

Explainable Artificial Intelligence (XAI) elucidates the logic behind how models generate predictions, explaining the reasoning behind a result. Both real-time explainability and real-time detection are essential components of operational defense systems, and it is essential to differentiate between the two. In this study, the object detection pipeline continues to function continuously at a rate of 35 frames per second (fps), which defines its real-time performance. Explainability is implemented in a triggered (on-demand) mode so that it does not interfere with the surveillance carried out. This dual-mode design emphasizes continuous situational awareness while selectively providing explanations to aid essential decision-making and human verification when it is required, embedding practical defense needs. Organizations use XAI frameworks to establish transparency, improve accountability, and build trust in their decision-making processes [9]. The Shapley Additive Explanations (SHAP) method has become an established model-agnostic technique that allows users to measure the impact of features in different situations throughout a system [10]. Although current research investigates explainable deep learning applications in safety-critical fields, there is an insufficient study of XAI usage for military object detection in aerial and satellite imaging. Current research evaluates models using general datasets that do not capture the specific requirements of military operational environments [11].

The Saudi Vision 2030 objectives for the development of defense technology through autonomous surveillance systems require organizations to implement interpretable AI systems that use effective detection methods. Recent studies demonstrate how AI helps secure operations using Unmanned Aerial Vehicles (UAVs) to detect landmines [12] or by combining multiple UAV images for reconnaissance missions [13]. Existing research has not yet presented a real-time interpretable decision support system for military asset detection that uses YOLOv8 and SHAP to meet the specific requirements of Saudi Arabian defense operations. The Saudi Arabian defense modernization goals require the implementation of XAI high-performance detection systems because the country needs autonomous aerial surveillance and rapid threat identification capabilities. Advanced YOLO architectures have proven their ability to assist monitoring tasks in multiple resource-management areas, demonstrating their ability to handle extensive surveillance operations [14].

The main contributions of this study are:

- An interpretable deep learning pipeline that integrates YOLOv8 with SHAP to enable accurate and explainable detection of 12 classes of military assets from UAV and satellite imagery.
- A defense-focused evaluation uses the publicly available Military Assets Dataset in YOLOv8 format, augmented and tuned for Saudi operational scenarios.
- Comprehensive experimental benchmarking against YOLOv5 and Faster R-CNN using mAP@0.5, mAP@0.5:0.95, precision-recall curves, and inference speed to validate both accuracy and real-time performance.
- A Saudi defense application perspective, aligning explainable object detection with Vision 2030 objectives for secure, autonomous, and trustworthy AI-driven surveillance systems.

II. PROPOSED METHOD

The proposed Digital Shield framework integrates an optimized YOLOv8 detector with XAI techniques to achieve accurate and interpretable military asset detection from Saudi aerial imagery.

A. Data Acquisition and Preprocessing

The system utilizes the publicly available Military Assets Dataset [15], which is curated for object detection and classification in military-related aerial environments. The dataset comprises a total of 26,315 labeled images distributed across 12 distinct classes, including both military and non-military objects. Defense imagery datasets usually suffer from class imbalance and highly variable operating environments,

such as desert dazzle and the coastal obscurity that characterizes the Saudi landscape [12]. To address the challenges of the Saudi operational environment, several preprocessing steps were applied:

- Normalization of pixel intensity values to the range [0,1].
- Geometric augmentation (horizontal and vertical flips, random rotation).
- Photometric augmentation (color jitter, brightness scaling, contrast adjustment) to address variable weather and lighting [16].

These augmentation strategies were selected due to the environmental and operational characteristics of Saudi aerial imagery. Photometric augmentations, including brightness scaling, contrast adjustment, and color jittering, were primarily employed to mitigate extreme illumination conditions, high contrast, and desert glare caused by intense sunlight and reflective terrain. These transformations enhance robustness against lighting variability commonly observed in arid regions. Geometric augmentations, such as random rotations and horizontal or vertical flips, were applied to improve invariance to changes in viewing angle, sensor orientation, and UAV altitude. This is particularly important in military surveillance scenarios, where targets may be observed from diverse perspectives during dynamic flight paths. Together, these augmentation strategies improve generalization while preserving the semantic structure of military assets.

The dataset was split into 70% training, 15% validation, and 15% testing, ensuring no geographic or scene overlap to avoid data leakage. To combat these issues, a variety of data augmentation methods and stratification splits were used to counter overfitting and ensure sound validation evaluation results in the face of all 12 classes. Table I presents the dataset structure, while Figure 1 shows the distribution of categories.

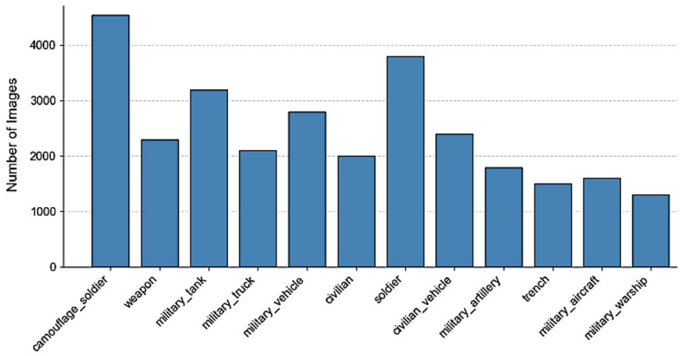

Fig. 1. Class distribution across the 12 military asset categories in the Saudi aerial imagery dataset.

TABLE I. DATASET STRUCTURE

Subset	Images	Description
Training	21,978	Images used to train the YOLOv8 detection model
Validation	2,941	Images for fine-tuning hyperparameters and early stopping
Testing	1,396	Images for final evaluation of detection performance
Total	26,315	Labeled images across 12 military asset categories

B. YOLOv8 Detection Model

The proposed Digital Shield system is based on the YOLOv8 framework [5]. The anchor-free detection head in YOLOv8 reduces the complexity of bounding box regression and improves the overall robustness of the model, especially with reference to the detection of small and densely clustered targets, as is the nature of military targets in general. Furthermore, the use of Mosaic data augmentation during training increases the overall generalizability of the model, as it is exposed to various scales of objects, which is important in the context of the heterogeneous environment of the defense sector. YOLOv8 was selected as the backbone model based on its overall balance of speed and detection accuracy, with YOLOv8l (large) being the model of choice over the smaller variants (n/s/m), despite the computational cost, as the targets in the environment, such as soldiers or the landmines, are generally small and may not be distinguishable from the background, such as the sandy terrain in the Middle East.

This model has the required depth of feature extraction to prevent False Negatives, which is critical, and can maintain a frame rate of 35 fps on high-end hardware (Nvidia A6000), which is within the required time constraints for processing video feeds from satellites or UAVs. This platform is used as a high-performance benchmark to ascertain the overall fidelity of the model in the context of explainability, as well as the overall accuracy, before the optimization to improve the overall embedded platform with reference to the required efficiencies. The Ultralytics implementation was used due to its decoupled detection head that is based on CSPDarknet.

The Intersection over Union (IoU) metric assesses the connection that exists between every predicted bounding box B_{pred} and the ground truth box B_{gt} that corresponds to it.

$$IoU = \frac{|B_{pred} \cap B_{gt}|}{|B_{pred} \cup B_{gt}|}$$

The model's total training loss is defined as:

$$\mathcal{L} = \lambda_{box}\mathcal{L}_{box} + \lambda_{obj}\mathcal{L}_{obj} + \lambda_{cls}\mathcal{L}_{cls}$$

which combines localization, classification, and distribution focal loss to refine object boundaries in complex Saudi aerial terrains. $\mathcal{L}_{\text{box}}$ employs CIoU for bounding box regression, while $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{cls}}$ use binary cross-entropy and multiclass cross-entropy, respectively. During inference, Non-Maximum Suppression (NMS) removes redundant detections by suppressing boxes with IoU above a threshold τ:

$$\text{Suppress if } IoU\left(B_i, B_j\right) \geq \tau \text{ and } s_i < s_j$$

where s represents the confidence score of detection.

C. Explainable AI Integration

To alleviate the black-box limitation of deep learning and to ensure transparency in critical military decision-making, XAI is introduced using SHAP [10]. SHAP is a model-agnostic framework derived from cooperative game theory that assigns importance to each feature for some prediction. Within Digital Shield, SHAP is intended to assign importances to certain regions of pixels on the YOLOv8 detection output. The SHAP value for a feature i is calculated using the following formula:

$$\phi_i(f,x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} [f(S \cup \{i\}) - f(S)]$$

where F is the set of all features. This method provides global interpretability (feature importance over the entire dataset) and local interpretability (for each detection). The visual explanations integrated in the Digital Shield framework help military operators verify and trust the model's output, thus adhering to the ethical AI policies stipulated in current defense technologies. Although other XAI methods, such as LIME and Saliency Maps, have been explored for object detection interpretability, they present certain limitations:

- LIME provides localized surrogate models but often incurs high computational cost and lacks stability under large input variation [8].
- Saliency Maps highlight influential pixels but are limited in providing class-specific or global-level understanding [17].

In contrast, SHAP is based on math and is model agnostic, hence supported by both global and local explanatory interpretations, keeping an integral aspect of this approach comparatively lightweight regarding computational resources. This makes SHAP more suitable for real-time defensive applications, where decisions have to be both immediate and transparent. The necessity of such explainable hybrid frameworks is paralleled in other critical fields, such as medical diagnostics, where the interpretation of deep learning decisions is vital to trust and safety [18].

TABLE II. COMPARATIVE ANALYSIS OF XAI TECHNIQUES FOR OBJECT DETECTION

Method	Global+Local explanations	Real-time suitability	Computational cost	Interpretability quality
LIME [8]	Local only	Low	High	Moderate
Saliency Maps [17]	Local only	Medium	Low	Low to moderate
SHAP (proposed)	Global & local	High	Medium	High

D. Triggered (On-Demand) Explainability Policy

To preserve real-time throughput, Digital Shield adopts a triggered (on-demand) explainability mode. The YOLOv8l detection pipeline operates continuously at 35 fps, whereas SHAP explanations are computed only when required. Triggering is activated under three configurable conditions:

1. Detections belonging to predefined high-priority classes (e.g., weapon, military_artillery, military_tank, military_aircraft).
2. Detections with confidence scores within an uncertainty range that requires verification (e.g., 0.4–0.6 or an operator-defined threshold).
3. Explicit operator request in ambiguous operational scenarios.

This selective activation ensures that explainability enhances decision support and human verification without interrupting continuous surveillance performance.

E. System Overview

The Digital Shield framework provides an end-to-end pipeline for real-time and interpretable military asset detection. Aerial images are preprocessed and augmented before being analyzed by an optimized YOLOv8 detector. Model outputs are then explained using SHAP, and the results are visualized in a decision support dashboard that shows detection confidence, class metrics, and key contributing features to support reliable defense decision-making. In operational deployment, the triggering policy is configurable and can be tailored to mission context (e.g., prioritizing weapon-related classes or low-confidence detections requiring verification). These architectural and training choices—specifically the anchor-free detection head and Mosaic augmentation—contribute to the robustness and accuracy of the Digital Shield framework under challenging Saudi aerial surveillance conditions.

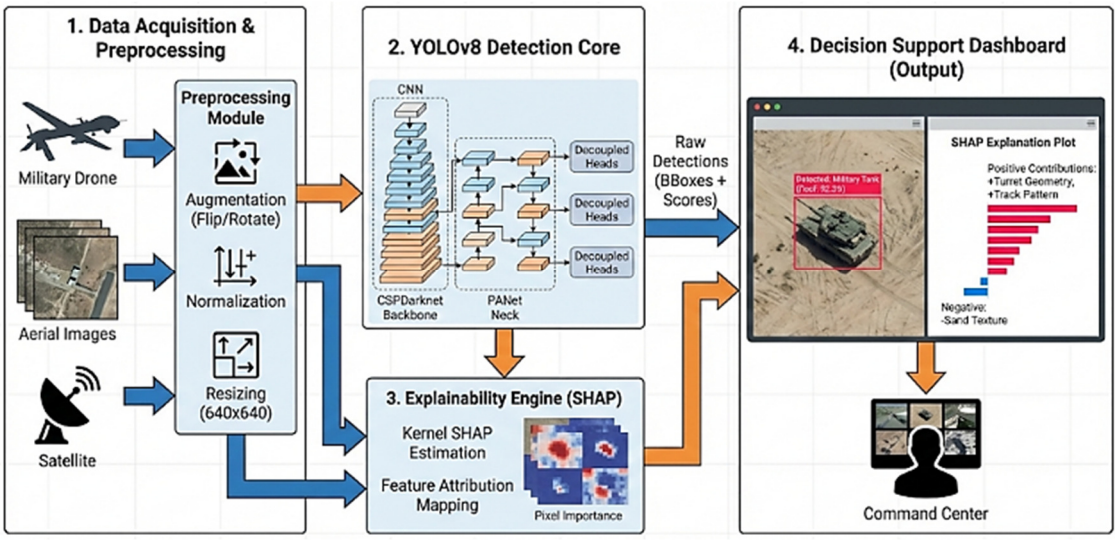

Fig. 2. End-to-end pipeline of the proposed Digital Shield framework showing data preprocessing and augmentation, YOLOv8-based detection, SHAP-based interpretability, and the defense decision support dashboard.

III. RESULTS

A. Dataset and Experimental Setup

Experiments were carried out using the Military Assets Dataset [15], consisting of 26,315 labeled aerial images in 12 classes of military and non-military scenes, such as tanks, armored vehicles, aircraft, missile launchers, naval vessels, and civilian vehicles. The dataset contains images of both UAV and satellite imagery that depict scenes of Saudi Arabia's operational environments, such as desert, coastal, and urban. The images were resized to 640×640 pixels and augmented with geometric transformations, such as horizontal flip and rotation, and photometric transformations, such as color jittering and brightness scaling. The dataset was randomly assigned to 70% for training (21,978 images), 15% for validation (2,941 images), and 15% for testing (1,396 images) without any overlapping scenes.

B. Baseline Models

To evaluate the proposed YOLOv8l+SHAP approach, it was compared against widely used object detectors:

- Faster R-CNN, a two-stage detector with region proposal networks delivering high-quality detection but slower inference [19].
- YOLOv5s, a single-stage anchor-based detector recognized for its balance between accuracy and speed [20].

Transfer learning was utilized to improve the proposed Digital Shield model, with the initial weights pre-trained on the COCO dataset. For training, a cosine annealing learning rate scheduler (lr_0 = 0.01, lr_f = 0.01) was utilized, along with a batch size of 16, an SGD optimizer (*momentum* = 0.937), a weight decay of 0.0005, and a total of 100 epochs. An early stopping strategy was implemented with a patience of 50 epochs, and validation loss was used as the criterion to prevent overfitting. In this work, optimization refers strictly to hyperparameter tuning within the training process rather than architectural modification. Specifically, genetic evolution was applied to optimize training-related hyperparameters such as learning rate, momentum, weight decay, and loss coefficients. Evolutionary search was conducted for 50 iterations (generations), where each candidate configuration was trained for 30 epochs and evaluated using validation mAP@0.5 as the fitness criterion. The final selected configuration was then retrained for 100 epochs to ensure fair and reproducible comparison with baseline models. The backbone architecture and detection head of YOLOv8l remained unchanged to ensure structural consistency and enable fair comparison. The genetic hyperparameter evolution was conducted for 50 iterations (generations) using the Ultralytics tuning framework, and candidate configurations were selected based on validation mAP@0.5 performance. During tuning, each configuration was evaluated for 30 epochs to balance computational cost and convergence stability. The final Digital Shield model was subsequently trained for 100 epochs using the optimized hyperparameters, while keeping the YOLOv8l architecture unchanged to ensure fair comparison and reproducibility.

C. Evaluation Metrics

Frames Per Second (FPS): iThe nference speed was measured on an NVIDIA RTX A6000 GPU. Performance was evaluated using Mean Average Precision (mAP@0.5 and mAP@0.5:0.95), Precision, Recall, and F1-score (the results are shown in Table III):

$$\text{mAP} = \frac{1}{\text{K}}\sum_{\text{k=1}}^{\text{K}} \text{AP}_\text{k}\,, \text{AP} = \int_0^1 \text{P(R)}\,\text{dR}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}\,, \text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}}$$

D. Quantitative Results

The proposed Digital Shield (YOLOv8+SHAP) framework outperformed baseline detectors in both detection accuracy and inference speed. The baseline models (Faster R-CNN and YOLOv5s) were trained using the same dataset splits (70/15/15) and their default optimal hyperparameters provided in their original repositories to ensure a fair 'out-of-the-box' comparison against the tuned Digital Shield framework.

TABLE III. COMPARATIVE PERFORMANCE OF YOLOV8L+SHAP VERSUS YOLOV5 AND FASTER R-CNN

Model	mAP@0.5	mAP@0.5:0.95	Precision	Recall	FPS
Faster R-CNN	83.4	59.1	84.2	82.7	9
YOLOv5s [20]	88.7	63.4	87.9	85.1	28
Digital Shield (proposed)	92.3	68.9	90.7	88.3	35

TABLE IV. PER-CLASS DETECTION PERFORMANCE ON THE MILITARY ASSETS DATASET

Class	AP@0.5	AP@0.5:0.95
camouflage_soldier	93.5	71.8
Weapon	90.2	66.4
military_tank	95.1	74.2
military_truck	91.8	70.5
military_vehicle	92.4	69.9
Civilian	87.2	63.1
Soldier	94.3	72.0
civilian_vehicle	89.6	64.5
military_artillery	94.8	73.7
Trench	86.9	62.0
military_aircraft	92.0	68.2
military_warship	91.4	67.3

IV. DISCUSSION

The proposed model outperforms other foundational models, including YOLOv5 and Faster R-CNN, achieving higher mAP. However, challenges such as class imbalance and diverse operating environments are highlighted through the propositions in [21]. Although data augmentation can address class imbalance to an extent, it is recommended to use Generative Adversarial Networks (GANs) to generate images for underrepresented classes, such as trenches, to further minimize the gap in terms of recall.

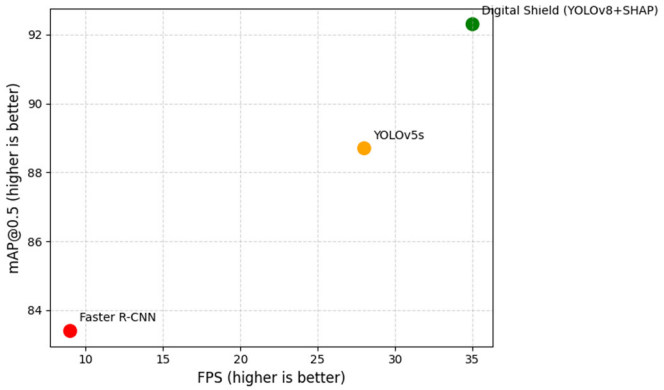

Fig. 3. Accuracy-Speed trade-off comparison of Faster R-CNN, YOLOv5s, and Digital Shield (YOLOv8l+SHAP) on the Military Assets Dataset.

The inclusion of SHAP provides an important capability to understand what factors are most important to each detection, which can be used by defense experts to validate model outputs, thus reducing False Positives. For instance, in bright desert environments, shadows are often incorrectly detected as trucks or other vehicles. The explanations provided by the proposed model are not only important but also actionable, allowing the user to validate that the detection is based on the features of the vehicle chassis, not shadows. Actionable explainability is also important in scenarios where friendly and enemy vehicles are visually similar. SHAP visualizations can also highlight subtle, yet significant characteristics that differentiate hostile assets from neutral or friendly vehicles, such as antenna arrangements, turret shapes, or weapon placements. By ensuring that detections are activated by these particular regions rather than by general vehicle outlines, operators can make decisions with greater confidence and reduce the likelihood of blue-on-blue incidents occurring during time-sensitive operations. Furthermore, simple to understand visual feedback helps identify False Positives that are the result of deliberate camouflage strategies or distractions that are present in the surrounding environment. With the help of SHAP explanations, analysts can determine whether the model focuses on significant structural characteristics or whether it incorrectly identifies irrelevant background patterns in complex landscapes where structures, rocks, or concealed nets may pass for military assets.

If the explanations indicate that the alert is dependent on non-semantic cues, then operators can downgrade or dismiss the alert. This helps prevent unnecessary escalation and preserve operational resources. The example presented in Figure 4 illustrates a representative case of actionable explainability, where SHAP attributions align with semantically meaningful object regions rather than non-semantic background artifacts. This example reflects a realistic and ambiguous scenario in which the features of structural objects are distinguished from environmental clutter, reinforcing the operational principle of preventing false escalation in high-risk surveillance contexts. Although additional visual scenarios could further expand this analysis, this example was selected to demonstrate the operational principle of distinguishing structural object features from environmental distractions under real surveillance conditions. Meanwhile, the exceptional modular structure of the framework also allows for easy integration with defense shoulder equipment, including UAV sets and custom command-and-control systems to directly help advance the country's modernization strategy for autonomous surveillance and border security. Combining the deep learning component with XAI, Digital Shield is a complement to black-box models to foster trust and improve decision-making in the development of secure systems.

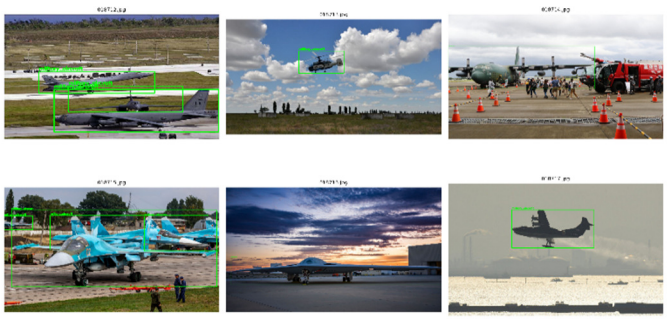

Fig. 4. Qualitative detection example of the proposed Digital Shield (YOLOv8+SHAP) accurately identifying military aircraft in Saudi aerial imagery with bounding boxes and confidence scores.

Table V provides a qualitative comparison of various object detection architectures for military and UAV-based object detection systems. In [11], MHNet was presented for object detection in camouflage-rich images using the MHCD2022 dataset. Although robustness in object detection for high-level concealment was ensured in this framework, there was no consideration of explanation mechanisms. In [21], a comprehensive survey of object detection architectures for UAVs highlighted advances in object detection architectures. However, no real-time explanation and verification were considered. In [14], YOLOv8l was used to detect objects in adverse weather conditions, proving its effectiveness in civilian traffic. However, this framework was limited to object detection in civilian traffic and lacked consideration of explanation mechanisms. Unlike these object detection architectures, the proposed Digital Shield framework is unique in its ability to conduct real-time object detection for military assets (35 fps) and provides explanations for object detection using SHAP and the Military Assets Dataset. To our knowledge, no object detection architecture has integrated YOLOv8l with explanation mechanisms for object detection in military aerial surveillance systems.

A. Explainability Evaluation

SHAP attributions provide intuitive local explanations, which are conducive to operator verification, although no scalar explainability score is discussed in this work. This is because, in the context of object detection, the quality of the explanation is inherently multi-dimensional and context-dependent, incorporating dimensions such as spatial alignment with detected objects, perturbation stability, semantic consistency, and usefulness, among others. Such an assessment, if distilled into a single scalar score, would require significant assumptions, which may not generalize to the

heterogeneous set of military targets and complex scenes. Instead, the work emphasizes the inherent multi-dimensional nature of the explanation, which is deliberately conceptualized as an operational means of enabling trust, opposed to an algorithmic measure thereof. This is because the visual attributions derived from SHAP are designed to support operator verification, allowing the operator to verify whether the detection is based on semantically meaningful structural features (e.g., parts of the chassis, turret, weapons, etc.), as opposed to spurious correlations (e.g., shadows, background, etc.).

TABLE V. QUALITATIVE COMPARISON WITH RECENT MILITARY AND UAV-BASED OBJECT DETECTION APPROACHES

Work	Domain	Dataset	Model	Real-Time	Explainability	Key Focus
[11]	Military camouflage	MHCD2022	MHNet	No	No	Camouflage robustness
[21]	UAV detection survey	Multiple UAV datasets	Various DL models	Partial	No	Architectural review
[14]	Adverse weather traffic	Road scenes	YOLOv8l	Yes	No	Weather robustness
Digital Shield (Proposed)	Military aerial assets	Military Assets Dataset	YOLOv8l+SHAP	Yes (35 FPS)	Yes (Triggered SHAP)	Real-time + Trust-aware detection

This verification capability is intended to support the development of operator trust, although the evaluation of the effectiveness of the system in this regard is deemed outside the scope of the current work and is scheduled as part of the validation process in the near future. From a measurable trust perspective, future validation could involve structured human-in-the-loop studies assessing how SHAP-based explanations influence operator decision accuracy, time-to-decision, and calibrated confidence levels in ambiguous scenarios. Such human factors evaluations would provide empirical evidence linking visual explanations to operational reliability, thereby offering a multi-dimensional validation of trustworthiness beyond purely algorithmic metrics.

As such, although the use of a single numerical value for the Explainability Score (ES) may not fully capture the complexity of explanation quality in the context of object detection, further research will explore the development of structured quantitative protocols that assess the operational effectiveness of the explanation process on the performance of the expert in the high-stakes environment of military defense operations. In this manner, the notion of "trustworthy" is grounded in the objective measurement of operational reliability as opposed to interpretive understanding.

B. Limitations

The proposed framework prioritizes throughput on the detection process in real-time and the use of SHAP in a triggered manner to minimize the impact on the surveillance process. In this regard, the explanation process is not continuously calculated for every frame of the surveillance process. In addition, the proposed framework will explore the development of standardizable quantitative proxies for explanation quality in terms of reliability and stability in the presence of efficiency-oriented optimizations, such as quantization for embedded environments. In terms of qualitative performance, the framework is qualitatively evaluated for its ability to detect military aircraft in the complex backgrounds of Saudi Arabia's aerial images and its effectiveness in the presence of diverse illumination conditions. The framework can detect aircraft in complex backgrounds with diverse illumination conditions, with high-confidence predictions in line with the requirements of real-time defense surveillance operations in military combat scenarios.

Although the current implementation is grounded in the centralized processing environment of a ground-station deployment on the NVIDIA RTX A6000, the 35 fps is an upper-bound benchmark in the context of real-time processing. In terms of prospective deployment in the embedded environment of UAV/UAS, the framework will need to be compressed, with FP16/INT8 quantization and other optimizations necessary in the energy-constrained environment of military UAV systems.

V. CONCLUSION

This paper presented the Digital Shield framework, an interpretable deep learning paradigm designed for the real-time detection of military assets in Saudi Arabia's aerial images. By incorporating YOLOv8 with SHAP, the proposed framework is capable of providing state-of-the-art accuracy in military asset detection (mAP@0.5 = 92.3%), as well as maintaining high speeds in terms of real-time processing (35 fps). The proposed framework provides the advantage of interpretability in terms of feature attribution, unlike traditional black-box methods in military asset detection, which can result in False Positive or Negative outcomes in military surveillance scenarios. The proposed framework is based on the integration of data preprocessing, optimization, and visualization, which provides the advantage of trustworthiness in terms of decision support, meeting the requirements of the Saudi Arabian defense establishment, and is in line with the strategic objectives of the Saudi Vision 2030 initiative for the localization of autonomous technologies in the country's defense capabilities. The proposed framework was also found to outperform the YOLOv5 [5] and Faster R-CNN [19] frameworks in terms of accuracy and computational efficiency in military asset detection scenarios.

Future work will explore multi-sensor fusion, combining thermal, radar, and satellite imagery, as well as the assessment of adversarial robustness to improve the overall resilience against counter-surveillance attacks. Moreover, the team will also conduct real-time analysis of defensive actions for embedded devices, addressing the challenges related to power efficiency and the need for model quantization (FP16/INT8) to maintain real-time performance. Another challenge related to the deployment of the model in an embedded environment is the robustness of the model's explainability in the presence of numerical precision reduction. Quantization methods, such as

FP16 or INT8, although efficient in optimizing computational performance and power consumption, may change internal activation distributions and feature attributions within the model. This, in turn, may have an effect on the overall robustness of the model's explainability, as obtained via the SHAP method. Therefore, in the context of this present work, the results obtained in the context of the RTX A6000 platform can be used as a reference benchmark for the assessment of the model's performance in an embedded environment, after the application of efficiency-oriented optimization methods.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of Prince Sultan University for paying the Article Processing Charges (APC) of this publication and for their support making this publication successful.

REFERENCES

[1] "Saudi Arabia - Defense & Security," *International Trade Administration*, Jan. 03, 2024. https://www.trade.gov/country-commercial-guides/saudi-arabia-defense-security.

[2] G. S. Xia *et al.*, "DOTA: A Large-Scale Dataset for Object Detection in Aerial Images," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 3974–3983, https://doi.org/10.1109/CVPR.2018.00418.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012, vol. 25.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[5] M. Sohan, T. Sai Ram, and Ch. V. Rami Reddy, "A Review on YOLOv8 and Its Advancements," in *Data Intelligence and Cognitive Informatics*, 2024, pp. 529–545, https://doi.org/10.1007/978-981-99-7962-2_39.

[6] I. M. Shaikh, M. N. Akhtar, A. Aabid, and O. S. Ahmed, "Enhancing sustainability in the production of palm oil: creative monitoring methods using YOLOv7 and YOLOv8 for effective plantation management," *Biotechnology Reports*, vol. 44, Dec. 2024, Art. no. e00853, https://doi.org/10.1016/j.btre.2024.e00853.

[7] D. Pahuja, S. Jain, and S. Kumar, "Evidence-based Inference and Quantification of Urban Expansion Using YOLOv8 and High-Resolution Satellite Imagery," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 26626–26631, Oct. 2025, https://doi.org/10.48084/etasr.12235.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, May 2016, pp. 1135–1144, https://doi.org/10.1145/2939672.2939778.

[9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, Sept. 2019, https://doi.org/10.1038/s42256-019-0088-2.

[10] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[11] M. Liu and X. Di, "Extraordinary MHNet: Military high-level camouflage object detection network and dataset," *Neurocomputing*, vol. 549, Sept. 2023, Art. no. 126466, https://doi.org/10.1016/j.neucom.2023.126466.

[12] M. Almutiry, "UAV Tomographic Synthetic Aperture Radar for Landmine Detection," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 5933–5939, Aug. 2020, https://doi.org/10.48084/etasr.3611.

[13] A. El-alami, Y. Nadir, and K. Mansouri, "An Efficient Geometric Transformation-Based Approach for Multi-UAV Image Stitching," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25507–25513, Aug. 2025, https://doi.org/10.48084/etasr.11719.

[14] N. U. A. Tahir, Z. Zhang, M. Asim, S. Iftikhar, and A. A. Abd El-Latif, "PVDM-YOLOv8l: a solution for reliable pedestrian and vehicle detection in autonomous vehicles under adverse weather conditions," *Multimedia Tools and Applications*, vol. 84, no. 23, pp. 27045–27070, Sept. 2024, https://doi.org/10.1007/s11042-024-20219-6.

[15] "Military Assets Dataset (12 Classes -Yolo8 Format)." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/rawsi18/military-assets-dataset-12-classes-yolo8-format.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, July 2019, Art. no. 60, https://doi.org/10.1186/s40537-019-0197-0.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." arXiv, Apr. 19, 2014, https://doi.org/10.48550/arXiv.1312.6034.

[18] M. Hammad, M. ElAffendi, and S. Meshoul, "HybridFusionNet with Explanability: A Novel Explainable Deep Learning-Based Hybrid Framework for Enhanced Skin Lesion Classification Using Dermoscopic Images," *Computer Modeling in Engineering & Sciences*, vol. 145, no. 1, pp. 1055–1086, 2025, https://doi.org/10.32604/cmes.2025.072650.

[19] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, 2015, vol. 28, Accessed: Feb. 26, 2026.

[20] G. Jocher, "YOLOv5 by Ultralytics." Feb. 2020, https://doi.org/10.5281/zenodo.3908559.

[21] G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao, "A Survey of Object Detection for UAVs Based on Deep Learning," *Remote Sensing*, vol. 16, no. 1, Dec. 2023, https://doi.org/10.3390/rs16010149.