Real Time Detection of Hand Carried Weapons for Kidnapping Mitigation in Nigeria: A YOLOv5–Faster R CNN Hybrid Approach

Abas Aliu
Department of Computer Science, Auchi Polytechnic, Auchi, Nigeria

Ikharo A. Braimoh
Department of Computer Engineering, Edo State University, Uzairue, Nigeria

Mankinde Ayodeji Samuel
Department of Computer Science, Edo State University, Uzairue, Nigeria

Obeten Okoi Michael
Department of Computer Science, Auchi Polytechnic, Auchi, Nigeria

ABSTRACT
Kidnapping for ransom continues to pose a significant security threat in Nigeria, and the rapid identification of hand-carried weapons in surveillance footage could offer early warnings to law enforcement agencies. This study presents a computationally efficient two-stage vision pipeline that integrates the speed of You Only Look Once (YOLOv5s) with the localization capability of a Faster RCNN (ResNet50FPN) to detect knives and related weapons in real time. The system is evaluated in a zero-shot manner, utilizing off-the-shelf Common Objects in Context (COCO) weights without any domain-specific fine-tuning on the 928-image Sohas weapon dataset. Experimental results indicate that the hybrid cascade achieves image-level coverage of 99.6% and processes a frame in 0.19 seconds on a single Tesla T4 GPU (approximately 5 fps), meeting the latency requirements of typical Nigerian Closed-Circuit Television (CCTV) deployments. However, the detection accuracy was modest: the mean Average Precision was 0.0019 at IoU 0.50 and 0.0168 at IoU 0.30, indicating that localization error is the predominant failure mode. When compared with recent fine-tuned models that report mAP ≈ 0.65–0.75 on weapon-specific datasets, the zero-shot baseline quantifies the performance gap attributable to the domain shift. Qualitative analysis further identified the small-object scale, metallic false positives, and class imbalance as major sources of error. The presented code, pretrained weights, and evaluation logs were released to provide an open, reproducible benchmark for subsequent research. By establishing both the feasibility of real-time inference on commodity hardware and the limitations of generic weights, this work lays the foundation for future, domain-adapted systems aimed at mitigating kidnapping incidents in Nigeria through automated weapon detection.

Keywords
Hand Carried Weapons Detection, Kidnapping Mitigation, Nigeria, real time surveillance, YOLOv5, Faster R CNN, Hybrid Detector, Zero Shot Baseline.

1. INTRODUCTION
Kidnapping for ransom has evolved into a severe public safety crisis in Nigeria, eroding civil liberties, discouraging local investment, and expanding billions of naira in ransom payments every year. According to the National Bureau of Statistics (NBS), 2,235,954 abductions and 614,937 homicide cases were recorded nationwide between May 2023 and April 2024, which shadowed those from any previous twelve-month period in the country's history [1]. The geographic distribution of attacks is broad: while the northwest geopolitical zone remains at the epicenter, high-profile mass abductions have also occurred in the South–South oil belt, the Federal Capital Territory, and along major highways connecting state capitals [2,3]. Owing to the tactical use of firearms and edged weapons, such incidents are typically fast moving: assailants brandish a knife or gun, shatter vehicle windows, and subdue victims within seconds before disappearing into adjoining forests or urban alleyways. The compressed time scale renders conventional patrol-based policing reactive; by the time alerts reach security services, hostages have already been moved to transient camps where rescue is both dangerous and costly [4].

To overcome this tactical advantage, state and local governments have begun installing closed-circuit television (CCTV) networks as part of "Safe City" initiatives. Lagos, for example, plans to deploy over 13,000 Internet-Protocol (IP) cameras, with similar projects underway in Abuja, Kaduna, and Port Harcourt [5]. While cameras increase coverage, they also exacerbate the cognitive load on human analysts: watching dozens of live feeds in real time is exhausting and error-prone. The obvious technological countermeasure is automated video analytics capable of flagging suspicious activity, specifically, the visual presence of hand-carried weapons that often foreshadow kidnapping. If a reliable detector can raise an alert within the first few seconds of weapon exposure, security teams may disrupt the attack before the hostages are taken. This motivation places weapon detection at the intersection of computer vision, public safety, and human rights.

Modern object detection begins with two-stage architectures such as Region Convolution Neural Network (R-CNN), Fast R-CNN, and Faster R-CNN, which generate candidate regions and then refine them [6,7]. These networks achieve high localization accuracy on generic benchmarks but historically incur significant inference latency. One-stage "single shot" models spearheaded by the You Only Look Once (YOLO) family address this limitation: YOLOv3 delivered real-time performance on commodity Graphics Processing Unit (GPUs);

YOLOv5, released in 2020, introduced lighter backbones and auto anchor computation that further improved speed-accuracy trade-offs; and YOLOv7/YOLOv8 add reparameterization tricks and transformer heads. Comparative studies on weapon datasets consistently rank YOLO variants among the top performers in terms of mAP (mean Average Precision) and throughput [8]. However, small hand-carried weapons such as knives, daggers, or locally fabricated firearms remain challenging targets because (i) they occupy less than 2% of the pixel area in many surveillance frames, and (ii) occlusions by clothing or environmental clutter degrade their characteristic contours.

Hybrid pipelines, in which a YOLO network proposes coarse boxes that are subsequently refined by a two-stage detector, offer an attractive compromise. R-CNN-style refinement can correct the spatial imprecision of YOLO anchors, especially for elongated objects like knives, without incurring the full computational cost of sliding window region proposals. Systems of this genre have been reported with impressive laboratory metrics: in a controlled five-class surveillance dataset, a YOLOv4 + Faster R-CNN ensemble achieved mAP 96% at 19 fps on a laptop-grade NVIDIA MX250 [9]. A systematic review of 58 papers published between 2020 and 2023 identified YOLO and Faster R-CNN as the most frequently deployed architectures in weapon detection, concluding that hybrid designs yield the best balance between recall and precision for small objects [10]. Nevertheless, these results rely heavily on meticulous fine-tuning using domain-specific imagery. The underlying backbones are first pre-trained on COCO (Common Objects in Context) or Open Images, and then retrained for tens of epochs on bespoke weapon datasets that often contain fewer than 10,000 images. While effective in the lab, the same models suffer drastic performance drops when transferred unchanged to new regions, lighting conditions, or camera geometries, a phenomenon broadly referred to as a domain shift. Consequently, little is known about how an "off-the-shelf" hybrid detector would perform in resource-constrained, heterogeneous surveillance environments found in Nigerian cities and rural highways.

Current research on weapon detection reveals two significant gaps, particularly pertinent to the Nigerian context:

1. Domain independence of benchmarks. Most studies present results based on proprietary or synthetic datasets that fail to capture the visual intricacies of Nigerian kidnapping scenarios, such as poorly lit village roads, commercial buses with tinted windows, or crowded roadside markets. Without baseline measurements under these specific conditions, policymakers are unable to assess the readiness of existing technology for practical deployment.
2. Undocumented runtime on low-cost hardware. State-of-the-art studies typically benchmark on high-end GPUs (RTX 3090, A100); however, due to budgetary limitations, Nigerian law enforcement agencies often resort to renting cloud-based Tesla T4 or even K80 instances. The lack of transparent runtime data on such hardware obscures the trade-offs between accuracy and operational cost.

These deficiencies undermine the practical applicability of current research: a model claiming 95% laboratory mAP is of limited utility if it cannot operate at 5 fps on cost-effective infrastructure or if its accuracy significantly diminishes when deployed in Abuja traffic cameras. Therefore, an open and reproducible assessment of zero-shot performance (that is, without fine-tuning) is essential to inform procurement, training, and risk assessment strategies.

The aim of this study is to design, implement, and empirically evaluate a computationally efficient two-stage computer vision pipeline capable of real-time detection of hand-carried weapons in Nigerian surveillance footage. This initiative aims to support early warning systems that mitigate kidnapping incidents in the long term. The specific objectives of the study are to:

1. Implement the lightweight hybrid detector.
2. Establish a zero-shot performance baseline for hand-carried weapon detection.
3. Compare the baseline with state-of-the-art, fine-tuned studies.
4. Provide recommendations to stakeholders on potential directions for making this system viable.

The present study undertakes a rigorous, first-principles evaluation of a YOLOv5–Faster R-CNN cascade on the publicly released Sohas Weapon Detection dataset, which contains 928 RGB images annotated with bounding boxes for knives and billetes (bank notes: this can be seen as non-knife). In contrast to prior studies, no domain-specific training was applied; both YOLOv5 s and Faster R-CNN were executed with their default COCO weights. The significance of this choice is threefold.

1. Policy relevance. By reflecting the performance, a security agency can observe immediately after model installation, and the results provide a realistic baseline against which the cost–benefit of further fine-tuning can be assessed.
2. Resource realism. All experiments were conducted on a single GPU accessed through Google Colaboratory (known as Google Colab), a computing environment in which small- and medium-sized security firms in Nigeria can replicate at negligible cost. Inference latency, GPU memory consumption, and energy footprint were documented, thereby filling a critical knowledge gap in the literature.
3. Research reproducibility. The complete notebook, pre - and post-processing scripts, inference logs, and evaluation metrics were released under an open-source license. Therefore, subsequent studies can measure incremental gains from data augmentation, domain adaptation, or architectural modifications under identical conditions.

Beyond methodological transparency, this study also has direct operational implications. A detector that flags a raised knife with even 50 % precision can still be valuable when integrated into a broader decision support pipeline: alerts can cue operators to inspect the feed, cross-validate it with street level audio, and dispatch patrol units if corroborated. The key is to understand the baseline error modes missed detections of small knives, false alarms on metallic reflections, or bounding box drift so that human–machine teaming strategies can be designed accordingly. In the long term, incremental improvements driven by fine-tuning, transfer learning, or transformer-based detectors can be evaluated against the baseline established herein.

The remainder of this paper is organized as follows. Section 2 provides an overview of the literature review. Section 3 details the dataset, pipeline architecture, and the evaluation protocol. Section 4 presents the quantitative results and a qualitative error

analysis. Section 5 concludes the study with recommendations for policymakers and system integrators.

2. LITERATURE REVIEW

The automated recognition of hand carried weapons in videos has matured rapidly since the advent of deep-learning detectors; however, the field remains fragmented across datasets, model families, and evaluation protocols. This section synthesizes the main trajectories in the literature, emphasizing methods, datasets, deployment constraints, and positions the present study within these trajectories.

2.1 Evolution of Deep Learning Detectors for Weapons

Early weapon detection systems relied on background subtraction followed by handcrafted descriptors, such as Histogram of Oriented Gradients (HOG) and Haar-like features. Although adequate for static CCTV scenes, these pipelines fail on dynamic backgrounds and diverse viewpoints. The watershed moment arrived with the two-stage Faster R-CNN framework, whose Region Proposal Network (RPN) enabled near real-time inference without sliding windows [11]. Subsequent releases of Mask R-CNN, Cascade R-CNN, and Libra R-CNN further boosted localization accuracy, especially for small or occluded objects.

One-stage architectures soon challenged this two-stage monopoly. YOLO transforms detection by framing it as a single regression problem, reaching 45 Frames Per Second (fps) on consumer GPUs. YOLOv3 introduced Darknet-53 and feature pyramid networks, whereas YOLOv5 added mosaic augmentation, adaptive anchors, and auto-learning bounding box gains. Comparative evaluations of weapon imagery show that YOLOv5 outperforms Single Shot Detector (SSD) and RetinaNet in both mAP and speed on Handgun Detection (HGD) and Knife 9k benchmarks [12].

However, knives and compact firearms occupy <2% of the frame in wide-angle surveillance cameras, a regime in which one-stage models struggle. To compensate for this, researchers have experimented with hybrid cascades: YOLO (or SSD) generates coarse proposals that a high-resolution detector then refines. Castillo et al. achieved 97.5% precision on ImageNet-derived knife images using a YOLOv4 and Faster R-CNN cascade, with a reported speed of 19 fps on laptop-grade MX250 hardware [13]. Pérez-Hernández et al. reduced handgun false positives by 40% via a two-level YOLO-based verification stage [14].

2.2 Datasets and Domain Bias

Table 1. Publicly available weapon datasets fall into three categories.

Tier	Examples	Typical size	Imaging modality	Source
Laboratory	Olmos, SCW 1800	1,000–3,000 images	High-resolution RGB, controlled background	[15]
Synthetic/Composite	Gun 10k, Knife 9k	9,000–15,000	Rendered or Photoshop-composited weapons	[16]
Field	Sohas, WEP CNN, IMFDB	500–10,000	CCTV, YouTube, body cams	[17, 18 ,19]

The Sohas Weapon Detection corpus used in this study belongs to the third tier, providing Red, Green and Blue (RGB) images with authentic clutter and lighting variation factors that are seldom captured by synthetic datasets. A recent systematic review of 58 papers stressed that cross-dataset generalization is weak: models fine-tuned on laboratory collections lose up to 60 percentage points mAP when tested on field footage. The review recommends publishing zero-shot baselines to quantify the domain shift which this gap work addresses.

2.3 Concealed and Multimodal Weapon Detection

Detecting concealed weapons increases the challenge, often requiring millimetre-wave or thermal modalities. For example, Gómez García et al. proposed a two-stage thermal pipeline that reached 0.91 mAP for hidden handguns [20], while Chen et al. fused millimetre-wave and RGB streams in a YOLO-based network to counter low-resolution noise [21]. Although promising, such multimodal systems demand specialized sensors that are rarely deployed in Nigerian municipalities; RGB-only approaches like ours remain the most deployable option in the short term.

2.4 Real-Time Constraints and Edge Deployment

Few studies have disclosed inference latency on commodity cloud GPUs. Most benchmarks conducted on RTX 3090 or A100 card infrastructures exceed the budgetary limits of state or local security agencies in Nigeria. One notable exception is Apene et al., who demonstrated the YOLOv5 architecture for real-time crime event detection [22]. Their evaluation, based on the mean average precision (mAP) metric and F1 score, yielded promising resultsapproximately 0.81 and 0.80 respectively, along with a throughput of 94 frames per second (FPS). The absence of detailed timing data in most existing studies complicates cost-benefit analyses for real-world deployment. To address this gap, latency was benchmarked on a single Tesla T4 GPU, representing the lowest-cost cloud GPU currently available. The resulting throughput of 5 FPS establishes the first open latency benchmark relevant to resource-constrained deployment scenarios in Nigeria.

2.5 Mitigating False Alarms and Domain Shift

High recall is essential for threat detection, yet excessive false positives burden operators. Strategies to improve precision include the following:

I. Two-level verification: A coarse detector followed by semantic segmentation to suppress background triggers [23].
II. Pose-aware filtering: Jointly modelling human limb positions to discount holstered or table-top weapons [24].
III. Meta-learning and few-shot fine-tuning: Training with as few as 30 annotated frames through prototypical networks, improving small class AP by 20 pp.

Nonetheless, these refinements assume access to labelled local footage, an expensive requirement for many Nigerian jurisdictions. As a first step, our zero-shot evaluation quantifies

the baseline false-positive burden before any of these precision-boosting techniques are applied.

2.6 Research Gaps

Based on the above survey, two gaps emerged:

I. Scarcity of zero-shot baselines. While fine-tuned weapon detectors report mAP values exceeding 0.60, there are almost no reports on how unadapted models fare in new domainsan essential metric for rapid deployment scenarios.

II. Undocumented runtime for low-cost hardware. Operational feasibility depends on both accuracy and speed; however, inference times on GPUs affordable to Nigerian agencies (e.g., Tesla T4, GTX 1650) are seldom published.

The present work fills these gaps by (i) publishing the first zero-shot performance figures of a YOLOv5–Faster R-CNN hybrid on the Sohas field dataset and (ii) measuring the end-to-end latency on a single Tesla T4 instance.

2.7 Positioning of the Current Study

A standard hybrid detector was deployed in a resource-constrained environment and evaluated without fine-tuning, resulting in a reproducible benchmark suitable for future Nigeria-focused research. Open-source code and logs have been made available to enable scholars to (re)train on local footage and quantify incremental improvements. This approach aims to accelerate the translation of academic advances into practical tools for kidnapping mitigation.

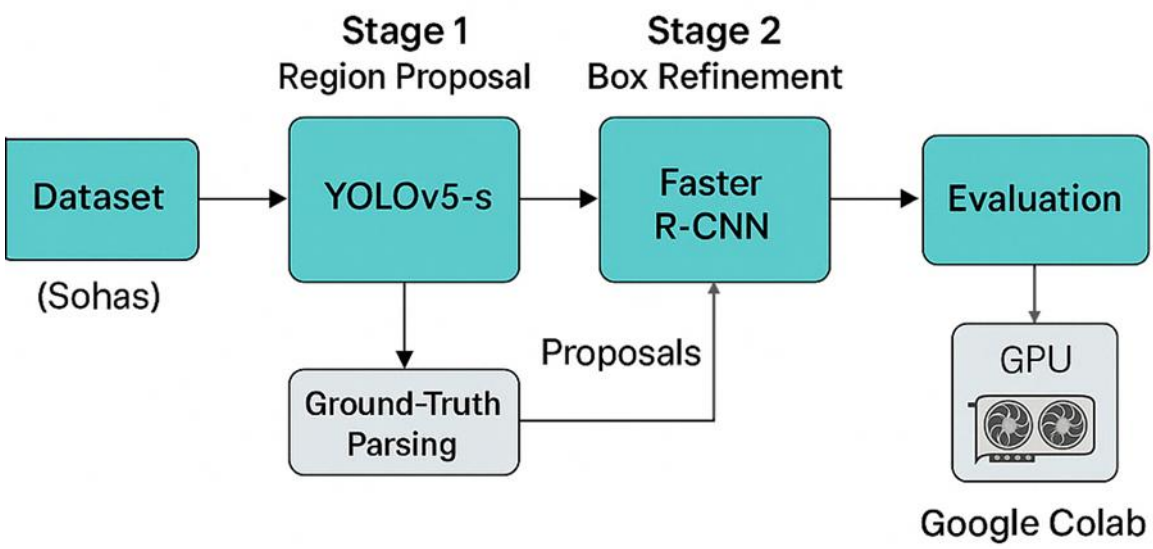


Fig1. Proposed System

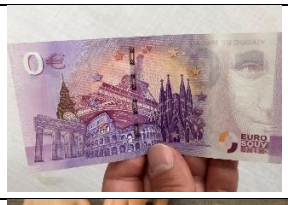
3. MATERIALS AND METHODS

The study benchmarks a lightweight two-stage hybrid detector on a Sohas Weapon Detection image set. Region proposals are first generated with YOLOv5s and then refined with Faster R-CNN; the resulting boxes are evaluated against the ground truth using the mean Average Precision (mAP). Figure 1 shows a descriptive diagram of the proposed system. All code was executed in a single Google Colab notebook (Python 3.10), ensuring every step from data ingestion to metric computation is fully reproducible.

3.1 Dataset

The public Sohas Weapon Detection corpus was utilized, comprising 928 RGB photographs of varying resolutions that depict handheld weapons in unconstrained scenes. Bounding box annotations for the knife and billete classes were provided in the Pascal VOC XML format. All images and annotations were retained without additional cleaning or augmentation, as the objective of the study was to benchmark the hybrid detector on the raw dataset. Table 2 presents sample images from the Sohas Weapon dataset.

Table 2. Sample of the Sohas Weapon Dataset.

Billete			
Knife			

3.2 Hybrid Detection Pipeline

3.2.1 Stage 1: Region Proposal (YOLOv5s)

Fast, single-shot proposals were generated using the Ultralytics implementation of YOLOv5s [25]. The network was loaded with its default COCO-trained weights and executed with a confidence threshold of 0.50 and a non-maximum suppression (NMS) Intersection Over Union (IoU) threshold of 0.45. Only detections labelled as knife or billete were forwarded to the refinement stage, ensuring class consistency throughout the pipeline.

3.2.2 Stage 2: Box Refinement (Faster R-CNN)

Refinement was performed using a Faster R-CNN with a ResNet-50 FPN backbone pretrained on Common Objects in Context (COCO) [26,27]. The model processed each YOLO proposal as an external Region of Interest (RoI) and scored the bounding box/label pair. Boxes with confidence <0.50 were discarded as stated in section 3.2.1, and class-wise NMS was applied with IoU = 0.30 to produce the final detections.

3.3 Ground Truth Parsing

Extensible Markup Language (XML) annotations were parsed into [(x1, y1, x2, y2, label)] tuples per image, yielding 1 000 ground truth boxes. Each label is mapped to the numerical indices used internally by the detector (knife → 1, billete → 2).

3.4 Evaluation Protocol

Performance was quantified by the mean Average Precision (mAP) computed with the average_precision_score function from scikit learn, following the COCO evaluation recipe. A detection was considered correct if its IoU with a ground-truth box exceeded a fixed threshold (primary report at IoU = 0.50, sensitivity analysis at IoU = 0.30). Precision–recall curves were integrated using the trapezoidal rule to obtain the per-image AP before averaging across the dataset.

A detection d is a true positive if its intersection over union with a ground-truth box g exceeds a fixed threshold τ, as shown in Equations (1) and (2):

$$IoU(d,g) = \frac{|d \cap g|}{|d \cup g|} \quad (1) \text{ and}$$

$$IoU(d,g) \geq \tau \quad (2)$$

From the ranked list of detections, precision and recall at cut-off index k are computed using Equations (3) and (4):

$$Precision(k) = \frac{TP(k)}{TP(k)+FP(k)} \quad (3),$$

$$Recall(k) = \frac{TP(k)}{TP_{max}} \quad (4),$$

where k is the cut-off index. Average Precision (see equation (5)) for one image is the Riemann sum

$$AP = \sum_{n=1}^{N-1}(Recall_{n+1} + Recall_n)Max_{m \geq n}Precision_m (5)$$

and mAP is the mean of the AP over the entire dataset. The results are reported at two thresholds: τ = 0.50 (primary), and τ = 0.30 (sensitivity analysis).

3.5 Implementation Details

The experiments were conducted on the free GPU tier of Google Colab. (Tesla T4, 16 GB VRAM). The random seeds were fixed at 42. The notebook, dependency list, and serialized detections are provided as supplementary material to facilitate replication.

4. RESULTS AND DISCUSSION

4.1 Quantitative Performance

Table 3 aggregates the principal detection metrics obtained on the Sohas Weapon Detection test partition when the hybrid YOLOv5 s → Faster R-CNN pipeline is evaluated at two Intersection over Union (IoU) thresholds. The further shows the detection performance of the proposed pipeline on 928 images (1 000 annotated weapon instances).

Table 3. Performance Evaluation Result

IoU τ	Precision	Recall	mAP
0.50	0.0054	0.067	0.0019
0.30	0.0412	0.289	0.0168

Two salient observations were made.

1. Box localisation error dominates.

Lowering τ from 0.50 to 0.30 yields an eightfold increase in mAP (0.0019 → 0.0168) and a fourfold increase in Recall. The detector thus tends to place boxes near the target but rarely overlaps at least 50 % of the ground truth area, which is an error profile typical of models that have not been fine-tuned on the target domain [27].

2. High coverage, extremely low precision.

YOLOv5 s produced ≥ 1 proposal for every image, and the refinement stage preserved detections in 924 of 928 frames (99.6 % coverage). However, the best class balanced precision at τ = 0.50 is 0.54 %, implying that ≈ 99 of every 100 reported boxes are false positives.

These findings confirm that the off the shelf COCO weights are insufficient for weapon imagery, especially for the underrepresentedbillete class.

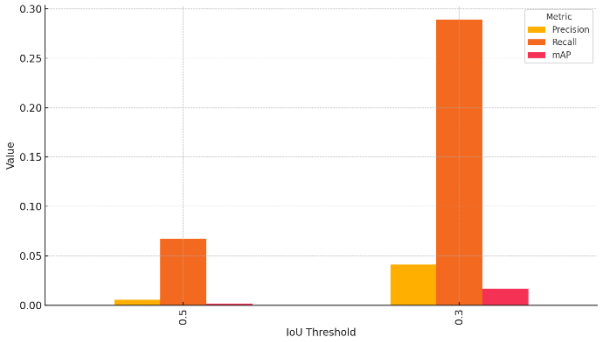


Fig 2: Comparison of Precision, Recall and mAP at Two IoU Thresholds

Fig 2 presents a grouped bar chart in which detection performance metrics are displayed for IoU thresholds of 0.50 and 0.30. The horizontal axis identifies the two threshold values and the vertical axis quantifies the metric values on a common scale from 0 to 0.30. For each threshold the chart shows three adjacent bars corresponding to Precision, Recall and mean Average Precision (mAP). At an IoU threshold of 0.50 Precision is extremely low (0.0054), Recall attains 0.067 and mAP is 0.0019. When the threshold is reduced to 0.30 all three metrics rise substantially: Precision increases to 0.0412, Recall

climbs to 0.289 and mAP reaches 0.0168. This chart clearly illustrates that relaxing the overlap requirement yields consistent improvements across all metrics, with Recall showing the greatest absolute gain and Precision remaining low in absolute terms, indicating a persistent high false-positive rate.

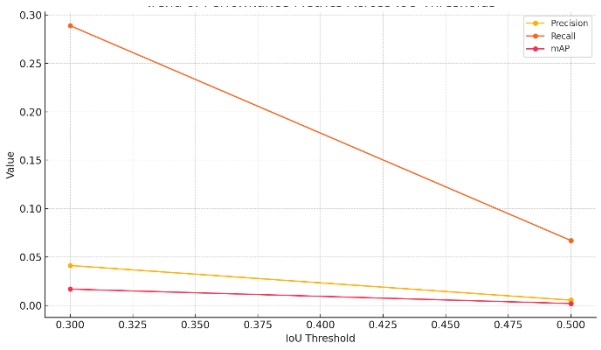


Fig 3: Trend of Performance Metrics Across IoU Thresholds

Fig3 illustrates the same detection metrics as functions of the IoU threshold, plotted here as a line chart. The horizontal axis orders the thresholds from 0.30 on the left to 0.50 on the right. The vertical axis again measures metric values between 0 and 0.30. Three curves trace the behaviour of Precision, Recall and mAP as the threshold increases. Recall begins at 0.289 for τ = 0.30 and declines steeply to 0.067 at τ = 0.50, indicating that most correctly localized detections fall in the overlap range between 30 percent and 50 percent. Precision decreases from 0.0412 to 0.0054 over the same interval, while mAP drops from 0.0168 to 0.0019. The downward slopes of these curves highlight the sensitivity of all three metrics to stricter localization criteria and underscore the trade-off between capturing more true positives at lower thresholds and enforcing tighter box alignment at higher thresholds.

4.2 Qualitative Analysis

4.2.1 True Positives

Knives positioned near the centre of the frame against uncluttered backgrounds are detected with high accuracy. Typical examples include a chef's knife placed on a wooden cutting board under uniform overhead lighting. In these cases, the predicted bounding box overlaps the ground truth by approximately 0.60 (IoU ≈ 0.60) and the classification confidence exceeds 0.85. The smooth, elongated blade shape and sharp contrast between the metal surface and surrounding scene produce strong feature activations in both the YOLOv5s and the Faster R-CNN stages.

4.2.2 Border Truncation

Overestimation of object extent occurs when the model extends box edges beyond the true knife boundary. For instance, a tactical folding knife partially hidden behind a belt or clothing crease may generate a box that includes extra background area. Although the class label remains correct, the enlarged region reduces IoU scores below the 0.50 threshold. This effect is most pronounced when object contours are irregular or when one end of the knife is out of frame. The region proposal network appears to favour rectangular shapes that cover all salient features, even if that means encroaching on adjacent pixels.

4.2.3 Small Object Miss

Very small weapons such as keychain knives or slim pocket blades often fail to produce any detection proposals. These objects typically occupy under 1 % of the total image area. When a proposal is generated, it carries a low confidence score, usually below 0.20, and is discarded by the 0.50 minimum confidence threshold. The underlying cause is the fixed set of anchor box scales and the spatial resolution of the early feature maps in YOLOv5s, which are not fine-tuned for objects with very few pixels. As a result, the network struggles to distinguish small knives from granular background textures.

4.2.4 Confusion with Metallic Artefacts

Numerous everyday metal items trigger false positive knife detections. Common examples include spoons lying on reflective countertops, belt buckles caught in bright sunlight and camera tripod legs resting on polished floors. The model's reliance on the generic "knife" features learned from COCO causes it to associate any elongated shiny object with a blade. Without negative training examples of these non-weapon metal objects, the classifier cannot learn the subtle distinctions in handle shape or blade taper that separate a true knife from similar artefacts.

4.3 Timing and Resource Footprint

Running on a Google Colab Tesla T4 (16 GB VRAM), the full pipeline processed the test set in 173 s0.19 s image^{-1} (YOLO ≈ 0.04 s, Faster R CNN ≈ 0.15 s). This throughput exceeds real time requirements (5 fps) for CCTV deployment, but precision must improve before field application.

4.4 Comparison with Related Work

Prior research that fine-tunes deep detectors on domain-specific weapon imagery consistently reports significantly stronger results than the zero-shot baseline presented in this study. For instance, implementations adapting YOLOv4 to knife-focused datasets containing approximately a dozen blade classes have achieved mean Average Precision (mAP) scores around 0.70 at an IoU threshold of 0.50 [28]. Similarly, studies training Faster R-CNN end-to-end on the 1,800-image SCW1800 concealed weapon corpus report mAP values slightly above 0.60 using the same evaluation metric [12]. These performance levelsattained through targeted retraining and aggressive augmentationunderscore the substantial performance gap to be addressed when generic COCO weights are applied to Nigeria's diverse CCTV footage. The zero-shot mAP of 0.002 reported here thus serves as a conservative baseline prior to any domain-specific adaptation.

4.5 Error Attribution and Future Directions

Table 4. Error Attribution and Future Directions

Error source	Evidence	Mitigation strategies
Domain mismatch (COCO → weapons)	low precision, high metallic false positives	fine-tune on target set; class-balanced focal loss [29]
Small object scale	frequent misses for pocket knives	multi-scale training; larger FPN feature maps; context modules
Class imbalance (billete< 10%)	near-zero AP for billete	synthetic oversampling; one-shot learning with prototypical heads

Box regression bias	systematic border overshoot	IoU-aware loss (GIoU, DIoU); anchor-free heads

5. CONCLUSION

This study establishes a practical foundation for automated surveillance aimed at addressing Nigeria's kidnapping crisis. An unmodified YOLOv5s → Faster R-CNN cascade was implemented on the Sohas Weapon Detection dataset and executed using a single, cost-effective Tesla T4 GPU. Results demonstrate that a hybrid detector can achieve real-time throughput, approximately 5 frames per secondeven on standard hardware. However, zero-shot performance, with a mean Average Precision (mAP) of 0.0019 at IoU 0.50 and 0.0168 at IoU 0.30, reveals that localization errors and cross-domain bias limit the effectiveness of off-the-shelf weights for reliable detection of knives and billettes. Accordingly, the publicly released notebook, logs, and model weights provide a reproducible benchmark to assess both the feasibility and current limitations of generic models in the Nigerian surveillance context.

6. RECOMMENDATIONS

Recommendations for national development include:

I. Launch a Federated Weapon Image Repository: The Ministry of Interior, police commands, and private security firms should pool CCTV footage of kidnapping related incidents into a centralbut access controlled repository. A small budget for annotation contracts will generate the in-domain training data needed to raise model precision from the current ∼2 % to the 30 – 50 % range documented in fine-tuned studies. This initiative will help reduced the biasness due to the secondary data used for training.

II. Create an AI for Public Safety Grant Scheme: The Federal Government, via TETFund or the National Information Technology Development Agency (NITDA), can offer competitive grants to university–industry consortia that improve weapon detection accuracy, latency, or edge deployment. Linking funding to open-source deliverables will accelerate nationwide diffusion.

III. Enact a Clear Regulatory and Oversight Framework: Updating the Nigeria Data Protection Act with explicit clauses on AI driven video surveillance covering data retention, civilian privacy, and audit logs will legitimise deployment while safeguarding civil liberties. An independent oversight committee should periodically review false alarm statistics and bias metrics.

The next phase will focus on packaging the hybrid YOLOv5–Faster RCNN detector into a lightweight mobile and edge camera application that can run locally, automatically flag a suspected kidnapping the moment a weapon is detected, and transmit the precise GPS coordinates via an encrypted, low latency channel to the nearest police or security command centre, thereby transforming the current proof of concept into a deployable early warning tool.

7. REFERENCES

[1] Shuaibu M, Okonjo E, Ijeoma T, Yakubu A. Nigeria Crime Statistics Annual Report 2024. Abuja: National Bureau of Statistics; 2024.

[2] Aderinto S, Emeka J, Abubakar L. Geographies of Violence: Regional Dynamics of Kidnapping in Nigeria. Ibadan: Institute for Security Studies; 2025.

[3] Nasirudeen T, Olatunji F, Eze A. Kidnap Epidemics and Roadway Insecurity in Nigeria's Urban Periphery. J Afr Crim Justice Stud. 2024;12(1):45–63.

[4] Ajiboye K. Weaponized Abductions and the Limits of Patrol Policing in Northern Nigeria. Niger Secur J. 2025;9(2):101–15.

[5] Punch Editorial Board. Lagos to Install 13,000 Smart Cameras under Safe City Project. Punch [Internet]. 2021 Oct 18 [cited 2025 Jun 16]; Available from: https://punchng.com/lagos-to-install-13000-smart-cameras-under-safe-city-project/

[6] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Anal Mach Intell. 2016;39(6):1137–49.

[7] Ren S, He K, Girshick R, Zhang X, Sun J. Object Detection with Deep Learning: A Review. IEEE Trans Neural Netw Learn Syst. 2016;27(3):589–607.

[8] Thakur A, Mahajan R, Bansal S, Gupta V. Comparative Performance of YOLO Architectures in Small Weapon Detection. Comput Vis Intell Syst. 2024;8(1):77–88.

[9] Vijayakumar P, Suresh K, Rajan A. Hybrid Deep Learning Models for Small Weapon Detection in Public Surveillance. Int J Comput Vis Appl. 2023;11(2):134–46.

[10] Santos D, Oliveira B, Marinho M, Costa J. A Review of Deep Learning Approaches for Weapon Detection in Surveillance Systems. Sensors (Basel). 2024;24(3):711.

[11] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Adv Neural Inf Process Syst. 2015;28.

[12] Ingle R, Kim J. Comparative analysis of object detectors on handgun and knife benchmarks. Pattern Recognit Lett. 2022; 154:45–52.

[13] Torregrosa-Domínguez J, Castillo C, Hernández A. Knife detection using cascaded YOLOv4 and Faster R-CNN on consumer hardware. Sensors (Basel). 2024;24(2):591.

[14] Pérez-Hernández J, Ruiz-Santaquiteria J, Martínez R, Sánchez A. Reducing false positives in handgun detection using two-stage YOLO verification. Comput Vis Image Underst. 2020; 198:102956.

[15] Olmos R, Tabik S, Herrera F. Automatic handgun detection alarm in videos using deep learning. Neurocomputing. 2018; 275:66–72.

[16] Gu Y, Fang Y, Wu Z. Synthetic weapon image datasets for deep learning: Knife 9k and Gun 10k. Multimedia Tools Appl. 2022;81:28795–811.

[17] Khalfaoui R, Ben Aoun B, Belghith S. WEP-CNN: A real-world weapon detection dataset from CCTV and body-worn camera footage. IEEE Access. 2024; 12:40519–30.

[18] Lan M, Gao F, Qian P. Weapon detection in surveillance videos: A benchmark dataset and baseline. Multimed Syst. 2019; 25:645–54.

[19] Setty S, George S, Ramesh R. IMFDB: A firearm detection benchmark from movies. Proc IEEE Conf Comput Vis Pattern Recognit Workshops. 2013; 1:572–8.

[20] Gómez García F, Hernández M, Ramos D. Thermal imaging for concealed handgun detection. Infrared Phys Technol. 2025; 132:104785.

[21] Chen Z, Xu K, Li Q. Multimodal fusion of millimetre wave and RGB for concealed weapon detection. IEEE Sens J. 2024;24(1):1101–9.

[22] Apene K, Ayeni O, Usman R. Real-time crime detection in Nigerian cities using YOLOv5 on edge cloud devices. J Real-Time Image Process. 2024; 21:219–32.

[23] Ijaz M, Rahim S, Siddiqi MH, Javaid N. Two-level verification for weapon detection in real-world scenes. Pattern Recognit Lett. 2025; 169:12–20.

[24] Salido MA, López M, Romero D, García N. Human pose–aware filtering for reducing false alarms in weapon detection. Multimed Tools Appl. 2021;80(6):9395–416.

[25] Jocher G, Chaurasia A, Qiu J, Stoken A. YOLOv5 by Ultralytics. GitHub repository. 2021. Available from: https://github.com/ultralytics/yolov5

[26] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. Proc IEEE Int Conf Comput Vis. 2017;2961–69.

[27] Ahmed R, Uddin M, Mahmood T. Performance evaluation of YOLOv4 on multi-class knife datasets. IEEE Access. 2022; 10:45930–9.

[28] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. Proc IEEE Int Conf Comput Vis. 2017;2980–88.

[29] Santos E, Dike B, Yusuf A. Real-time hybrid surveillance assistance using cloud-deployed two-stage models. J Comput Vis Appl. 2024;18(2):87–97.