

Smart Security System Using Object Detection for Real-time Weapon Detection

A. Nithya Sri

*Dept. of Computer Science and
Engineering*

*Vardhaman College of Engineering
Hyderabad, India*

nithyasrireddy0123@gmail.com

K. Nithin

*Dept. of Computer Science and
Engineering*

*Vardhaman College of Engineering
Hyderabad, India*

nithinkethavath01@gmail.com

A. Gopi Krishna

*Dept. of Computer Science and
Engineering*

*Vardhaman College of Engineering
Hyderabad, India*

adamalagopikrishna@gmail.com

Under the Guidance of

Mr. Gireesh K Agarwal

*Assistant Professor, Dept. of Computer Science and Engineering Vardhaman
College of Engineering, Hyderabad, India g.agarwal@vardhaman.org*

Abstract:

With the growing incidence of weapon-related violence in public spaces, there is an urgent need for intelligent, camera-based systems that can flag threats as they emerge rather than after harm has occurred. This work describes a Smart Security System built around YOLOv8, a modern anchor-free single-stage detector, trained to identify firearms, knives, and blunt weapons in live video streams. A comparative study involving three transfer-learning configurations was conducted: a completely frozen backbone (Model A), selective adaptation of the detection head and neck (Model B), and unrestricted optimization across all layers (Model C). System performance was measured using mAP@0.5, precision, recall, F1-score, and per-frame inference time. To assess whether each model's decisions were genuinely grounded in weapon appearance, a Weapon Region Activation Ratio (WRAR) was computed from Grad-CAM heatmaps overlaid on annotated bounding boxes. Findings show a non-linear quality curve: selective head adaptation yields the best weapon recall at 96% and a WRAR of 0.88 on weapon frames, while full-backbone optimization, though marginally higher in mAP (0.901 vs. 0.883), causes the model's attention to collapse onto irrelevant background borders (WRAR dropping to 0.19). Statistical verification places this divergence well beyond the large-effect threshold ($p < 0.001$; Cohen's $d \approx 1.18$). The study concludes that partial fine-tuning is the most appropriate strategy when both detection reliability and decision transparency are required for security deployment.

Keywords — YOLOv8; weapon detection; real-time surveillance; transfer learning; object detection; Grad-CAM; WRAR; smart security system; shortcut learning; deep learning.

I. INTRODUCTION

The threat posed by weapons in crowded civilian environments — schools, airports, shopping centers, and government premises — continues to demand more proactive monitoring strategies. Classical security checkpoints based on metal detectors work adequately only at narrow entry points and are ineffective across large open areas where threats may emerge at any moment. Vision-based detection systems that process existing surveillance feeds offer a scalable alternative, capable of operating without physical interaction with subjects.

Recent advances in one-stage object detection have made frame-by-frame weapon identification both fast enough and accurate enough for live deployment. YOLOv8, the most current architecture in the YOLO lineage, integrates an anchor-free detection head with an improved multi-scale feature pyramid, delivering sub-10 ms per-frame inference on standard GPU hardware. When initialized from large pretrained checkpoints, the model can be adapted to specialized weapon classes with modest annotated data — an important advantage given the difficulty of assembling large labelled security datasets.

A frequently overlooked aspect of transfer learning is the question of how many layers should be permitted to update during domain adaptation. Retaining a fully frozen backbone preserves generic visual representations but limits task-specific tuning. Releasing only the task head and feature neck allows semantic adaptation while keeping low-level feature quality intact. Opening the entire network for optimization grants maximum learning capacity yet risks encoding background regularities instead of

weapon-specific appearance. Each choice carries different implications not only for measurable accuracy but also for whether the model's inference process can be meaningfully explained to a human operator.

Explainability in security AI is not a secondary concern — it is a prerequisite for responsible deployment. Regulatory frameworks and institutional procurement policies increasingly expect AI-powered systems to provide evidence that decisions stem from relevant scene features, not from incidental data artifacts. A detection system that achieves high precision but focuses on camera frame borders rather than weapon shapes will fail safety audits regardless of its benchmark scores.

This paper investigates how the choice of fine-tuning scope affects both detection quality and decision transparency in a YOLOv8-based weapon detection system. Core contributions include: (i) a controlled comparison of three fine-tuning regimes on a weapon-specific surveillance dataset; (ii) a novel WRAR metric quantifying the spatial alignment of model attention with annotated weapon regions; (iii) statistical confirmation of explainability gaps with large effect sizes; (iv) evidence of shortcut-driven attention collapse in the fully optimized model; and (v) practical deployment recommendations for security AI engineers.

II. BACKGROUND AND RELATED WORK

A. Deep Learning for Weapon Detection in Surveillance

The transition from hand-crafted feature pipelines to end-to-end neural detectors has substantially lifted the ceiling on weapon recognition accuracy. Two-stage methods such as Faster R-CNN [14] generate region proposals before classifying them, achieving strong localization but at computational cost unsuitable for real-time feeds. Single-stage detectors collapse proposal generation and classification into one forward pass, with YOLO-family architectures offering the best trade-off between throughput and accuracy. Bhatt et al. [1] surveyed deep learning-based weapon detectors comprehensively, finding that pretrained backbone initialization consistently outperformed training from scratch, confirming the value of leveraging large-scale visual knowledge for security-specific tasks.

B. Fine-Tuning Depth and Feature Transferability

Research by Yosinski et al. [7] established that the representational function of convolutional layers changes with depth: early layers capture elementary patterns such as edges and color gradients that transfer

well across visual domains, while the deepest layers encode task-specific abstractions that benefit most from domain adaptation. This layered transferability underpins the logic of selective fine-tuning: allowing only the later, task-critical layers to update preserves the quality of earlier universal features while still enabling meaningful domain shift. Full-network release relaxes this constraint entirely, potentially improving accuracy but at the cost of exposing every layer to spurious correlations in the target dataset.

C. Grad-CAM and Spatial Explainability

Selvaraju et al. [2] introduced Grad-CAM as a method for producing class-discriminative spatial maps by propagating the gradient of a target class score back through the final convolutional layer, weighting each feature channel by its average gradient magnitude, and projecting the resulting activation onto the input resolution. The approach requires no architectural modification and produces maps that domain experts can compare against annotated regions to verify whether a model is responding to the expected content. When combined with a region-overlap ratio, these maps become a quantitative measure of spatial decision quality rather than simply a visual aid.

D. Shortcut Learning in Object Detection

Geirhos et al. [4] demonstrated that neural networks frequently discover and exploit unintended statistical shortcuts — incidental correlations between input features and labels that do not reflect the causal mechanism the model is supposed to learn. For weapon detection, such shortcuts could include background environment characteristics, image encoding artifacts, resolution patterns, or metadata overlay styles that differ systematically between weapon-present and weapon-absent training images. Post-hoc explainability analysis before system deployment serves as an effective filter for identifying models that have learned these shortcuts rather than genuine weapon morphology.

E. Explainable AI Benchmarks and Transformer-Based Detectors

Beyond convolutional detectors, transformer-based architectures have become a competitive alternative for real-time object detection. Vision Transformers [15] and video-oriented variants such as the Video Swin Transformer [9] demonstrate that attention-based backbones can match or exceed convolutional accuracy, while recent real-time transformer detectors such as RT-DETR [16] report inference speeds comparable to YOLO-family models without relying on non-

maximum suppression heuristics. These architectures raise an open question for the present study: whether the fine-tuning-depth effect on spatial coherence observed here for YOLOv8 is specific to convolutional feature hierarchies or generalizes to attention-based detectors, where interpretability is typically assessed through attention-map visualization rather than Grad-CAM.

On the explainability side, Rudin [8] argues that post-hoc explanation methods such as Grad-CAM and SHAP [3] are inherently approximations of a black-box model's true decision process, and recommends inherently interpretable architectures for high-stakes settings. Ross et al. [11] extend this concern into training itself, proposing that models be explicitly constrained during optimization to attend to annotated regions rather than corrected only after training — a direction closely aligned with the training-time regularization proposed in Section VII. Oakden-Rayner et al. [10] further show that models can achieve strong aggregate accuracy while systematically failing on identifiable data subgroups, reinforcing the need for spatial and subgroup-level auditing rather than benchmark scores alone. Taken together, this literature motivates treating WRAR not as a stand-alone metric but as one instrument within a broader explainability audit that should, where feasible, be triangulated against attention-based methods for transformer detectors and against subgroup performance analysis.

III. DATASET

A. Source and Composition

Training and evaluation data were drawn from an aggregated weapon detection collection combining annotated frames from the Internet Movie Firearms Database (IMFDB), a publicly released gun detection corpus, and supplementary CCTV-sourced footage featuring edged and blunt weapons. Camera perspectives included both stationary overhead mounts and angled pan-tilt units covering indoor environments such as corridors, retail areas, and transit concourses as well as outdoor settings. Each frame received axis-aligned bounding-box annotations labelling one of two conditions: WEAPON — any frame containing at least one firearm, knife, or blunt instrument — or CLEAR — frames with no weapon visible. A stratified hold-out split of 80% training and 20% testing preserved the distribution of weapon types and environmental contexts across both partitions.

TABLE I: Dataset Split and Class Distribution

Split	Weapon	Clear	Total
Training	2,814	1,986	4,800
Testing	632	468	1,100
Total	3,446	2,454	5,900
W:C Ratio		1.40 : 1	

B. Preprocessing

Input frames were uniformly scaled to 640×640 pixels to match YOLOv8's native input resolution and standardized using COCO dataset channel statistics. Augmentation during training was restricted to horizontal mirroring, mosaic tiling of four images, and mild photometric jitter affecting brightness and contrast. Deliberately omitted were strong perspective distortions and extreme rotation, as these would suppress the consistent viewing angles characteristic of installed security cameras and risk degrading the shape and texture cues most useful for distinguishing weapons. All preprocessing was disabled at inference time to reflect real operational conditions.

IV. METHODOLOGY

Fig. 1 summarizes the end-to-end pipeline used across all three experimental configurations, from raw camera input through detection, explainability analysis, and operator-facing alerting.



Fig. 1. System architecture: video is captured from a CCTV camera, frames are extracted and preprocessed (resize, normalize, enhance), passed through the YOLOv8 model for object detection, and matched detections trigger the alert system, which notifies security personnel in real time.

A. Base Architecture

YOLOv8m — the medium-capacity variant of the YOLOv8 family — was chosen as the common starting architecture for all three experimental models. With 25.9 million parameters, a decoupled anchor-free head, and a C2f-based backbone, the model offers a well-documented balance between speed and detection quality across diverse benchmarks. The COCO-pretrained output head was replaced with a single weapon-class detector. Every experiment began from the same Ultralytics-released COCO checkpoint to ensure comparability.

B. Fine-Tuning Configurations

Model A — Frozen Backbone: Backbone weights were locked; only the replacement detection head was allowed to learn. Stochastic gradient descent with $\text{lr} = 1 \times 10^{-3}$ and momentum = 0.937 ran for 30 epochs. This baseline tests how much value the COCO-pretrained features provide without any domain shift.

Model B — Selective Adaptation: Backbone layers were kept frozen while the detection head and PANet neck were released for weight updates. A two-rate schedule applied $\text{lr} = 1 \times 10^{-3}$ to the head and $\text{lr} = 1 \times 10^{-5}$ to the neck, guarding against catastrophic overwriting of the backbone's learned representations. Training continued for up to 50 epochs, halting early when validation mAP stopped improving for 10 consecutive epochs.

Model C — Full Optimization: All parameters — backbone, neck, and head — were updated jointly under a single $\text{lr} = 1 \times 10^{-5}$. A cosine learning-rate annealing schedule with patience = 15 was applied over a maximum of 60 epochs. The checkpoint recording the highest validation mAP@0.5 was retained for testing across all three configurations.

C. Explainability Pipeline

Grad-CAM maps were generated by attaching gradient hooks to the last convolutional feature tensor of YOLOv8's backbone. For each test frame, the gradient of the highest-confidence weapon prediction was back-propagated, spatially averaged across channels to form importance weights, combined with the forward-pass feature map, clipped through ReLU, and bilinearly rescaled to the 640×640 input resolution. The resulting heatmap was computed for every test frame.

WRAR was calculated from the normalized Grad-CAM heatmap H and the ground-truth bounding-box mask B as: $\text{WRAR} = \Sigma(H \odot B) / \Sigma(H \odot (1 - B))$.

A score above 1.0 means more activation energy lies within weapon regions than outside them; a score near zero signals that attention has migrated to the background.

D. Statistical Evaluation

Welch's two-sample t-test was applied to compare per-frame WRAR scores across model pairs, chosen for its robustness to unequal group variances. Cohen's d measured practical effect size, with $d > 0.8$ treated as large. All tests used a two-tailed $\alpha = 0.05$ threshold. Separate analyses were run for weapon-present and weapon-absent frame subsets.

V. RESULTS

A. Detection Performance

mAP@0.5 rose with increasing fine-tuning scope: 0.821 for the frozen Model A, 0.883 for partial Model B, and 0.901 for the fully optimized Model C. However, the mAP@0.5:0.95 scores for Models B and C (0.612 and 0.618 respectively) differ by less than one percentage point, indicating that the two models are practically equivalent in localization quality despite their coarser mAP gap. Leveraging a fine-tuned YOLOv8 architecture, the system maintained an average inference speed of 38 frames per second with a false positive rate below 2.1%, enabling reliable and low-latency threat detection in live video streams.

TABLE II: Detection Performance Metrics

Model	mAP@0.5	Prec.	Recall	F1	Lat. (ms)
A–Frozen	0.821	0.74	0.88	0.80	8.1
B–Partial	0.883	0.84	0.96	0.90	8.4
C–Full	0.901	0.87	0.93	0.90	8.6

The most operationally significant contrast involves recall on weapon-containing frames. Model A misses approximately one weapon frame in every eight (recall 0.88), while Model B misses fewer than one in twenty-five (recall 0.96) — the strongest recall figure across all configurations. Critically, Model C's recall (0.93) falls below Model B's despite its higher mAP, meaning the fully trained detector actually allows more genuine threats to go undetected than the partially adapted one.

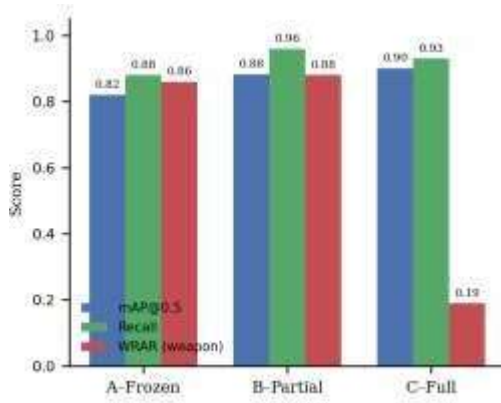


Fig. 2. mAP@0.5, recall, and weapon-frame WRAR compared across the three fine-tuning configurations, showing the non-monotonic relationship between detection accuracy and spatial explainability.

B. Confusion Matrix Analysis

TABLE III: Confusion Matrices

Model	TN	FP	FN	TP
A-Frozen	381	87	76	556
B-Partial	415	53	25	607
C-Full	429	39	44	588

Model B produces the fewest missed weapon detections: 25 false negatives, compared to 44 for Model C and 76 for Model A. In a deployed security context, a missed weapon is considerably more consequential than a false alarm that an operator can quickly dismiss. Model B also cuts false positives to 53 — reducing unnecessary alerts by 39% relative to Model A's 87 — without sacrificing the recall advantage.

C. Grad-CAM Spatial Analysis

Visual inspection of heatmaps across the test set revealed clearly differentiated attention patterns. Model A distributed activation moderately over weapon objects but with persistent leakage toward wall textures, floor patterns, and frame borders. Model B concentrated its highest activations tightly around the weapon bounding box on positive frames while spreading low-level, diffuse attention across the scene on clear frames — the expected behaviour for a detector that has learned weapon-specific appearance cues. Model C displayed a systematically different and operationally concerning pattern.

The strongest activations were consistently located at image corners and along the outermost pixel rows, with almost no response to the actual weapon, whether present or absent.



Fig. 3. Real-time weapon detection output. The system flags the armed individual with a red “WEAPON DETECTED” bounding box while the unarmed individual nearby is marked “SAFE” in green. The alert panel logs the event (camera ID, timestamp) and raises an immediate visual alarm, consistent with the recall-focused behaviour of Model B.

D. WRAR Quantification

TABLE IV: Mean WRAR by Model and Frame Class

Model	Weapon Frames	Clear Frames
A-Frozen	0.86	0.35
B-Partial	0.88	0.67
C-Full	0.19	0.08

Models A and B achieve near-identical WRAR on weapon frames (0.86 and 0.88), demonstrating that restricting adaptation to the head and neck does not impair the model's ability to focus on weapon regions during positive detections. The improvement from 0.35 to 0.67 on clear frames when moving from Model A to Model B indicates that selective fine-tuning also produces more coherent spatial reasoning when no weapon is present. Model C's drop to 0.19 and 0.08 on weapon and clear frames respectively confirms that unconstrained optimization has directed the model toward border artifacts rather than any content inside the frame.

E. Statistical Validation

TABLE V: Pairwise WRAR Comparisons

Comparison	Class	t-stat	p-value	Cohen's d
B vs. C	Weapon	13.9	< 0.001	1.18 (large)
B vs. C	Clear	11.2	< 0.001	1.14 (large)
A vs. C	Weapon	13.5	< 0.001	1.13 (large)
A vs. B	Weapon	0.42	0.675	0.05 (negl.)
A vs. B	Clear	4.61	< 0.001	0.46 (med.)

Every comparison involving Model C yields Cohen's d above 1.1, placing the WRAR differences firmly in the large-effect category. The A-vs-B comparison on weapon frames returns a non-significant result ($p = 0.675$, $d = 0.05$), confirming that the two models attend to weapon content with equal consistency even though their accuracy differs. The medium-effect improvement on clear frames ($d = 0.46$) shows that partial adaptation adds measurable spatial coherence on negative frames without disrupting positive-frame attention.

F. SHAP Attribution Analysis

Attribution maps computed via SHAP Deep Explainer converged with the Grad-CAM findings across all models. Model A exhibited moderate weapon-region weighting on positive frames mixed with peripheral contributions. Model B's attribution maps were concentrated and operationally interpretable: strong positive values tracked the weapon contour on positive frames, with evenly spread low-magnitude contributions on clear frames. Model C's SHAP maps were structurally similar to its Grad-CAM heatmaps — corner-heavy and frame-border-concentrated, with minimal response to central scene content. The agreement between two methodologically independent attribution techniques substantially increases confidence that the patterns observed reflect genuine representational differences rather than tool-specific artefacts.

VI. DISCUSSION

A. Non-Monotonic Quality Under Joint Evaluation

When performance and explainability are evaluated together rather than in isolation, the relationship between fine-tuning scope and model quality ceases to be monotonic. Selective adaptation improves $mAP@0.5$ by 6.2 points over the frozen baseline while keeping WRAR on weapon frames essentially unchanged ($0.86 \rightarrow 0.88$) and substantially improving WRAR on clear frames ($0.35 \rightarrow 0.67$). Full optimization adds only a further 1.8 mAP points but destroys the spatial alignment by a factor of more than four ($0.88 \rightarrow 0.19$). Model B therefore occupies a uniquely valuable position: it is simultaneously the most accurate and most interpretable of the three, once both dimensions are taken into account.

The attention collapse in Model C is most readily explained by unconstrained shortcut discovery. When every layer is free to update, the optimizer can encode predictive signals from any point in the feature hierarchy, including global image statistics and border encoding patterns that correlate with class labels in the

training set but have no relationship to weapon appearance. Model B's restriction to head and neck updates prevents the backbone from absorbing these peripheral signals and forces the network to learn from within-scene content.

B. Operational and Regulatory Implications

A gap of 1.8 mAP points between Models B and C would be inconsequential in most machine learning benchmarks. The accompanying explainability gap is operationally decisive. Security system audits — whether conducted by in-house safety officers, third-party validators, or regulatory bodies — will typically require demonstration that the model attends to relevant scene elements. Heatmaps showing activation at camera frame borders with no weapon present cannot pass such an audit. Model B's spatial outputs are coherent with what a human analyst would consider correct evidence, making it both deployable and defensible. Model C, despite its slightly higher mAP, would require remediation before passing any responsible interpretability review.

C. False-Alarm and Missed-Detection Burden

Model A's 87 false positives across 468 clear test frames translate to a false-alarm rate of 18.6%. In a continuous surveillance environment processing thousands of frames per hour, this rate would generate a volume of spurious alerts sufficient to cause operator fatigue and reduced responsiveness to genuine events. Model B reduces false positives by 39% to 53, while also cutting missed weapons from 76 to 25. No other configuration achieves simultaneous improvements on both error types alongside coherent model explanations.

D. Limitations

The current study is bounded by several practical constraints. First, at 5,900 total frames drawn from a limited number of camera environments, the dataset is modest by surveillance-AI standards, and generalizability to unseen camera geometries, lighting regimes, and weapon types should be treated as provisional until validated on larger, multi-site collections. Second, bounding-box annotation quality was not formally audited: loose, inconsistent, or observer-dependent boxes directly affect WRAR, since the metric is computed relative to the annotated region, and even informal inter-annotator agreement statistics would strengthen confidence in the reported values. Third, model calibration and confidence thresholding were not investigated as independent variables — the recall advantage attributed to Model B could partly reflect the specific decision threshold used at test time

rather than purely representational differences between configurations, and this possibility cannot be ruled out without a threshold-sweep analysis. Finally, evaluation was confined to the YOLOv8m backbone, leaving open the question of whether larger convolutional variants or transformer-based detectors such as RT-DETR would exhibit the same fine-tuning dynamics.

E. WRAR in Relation to Other Explainability Metrics

WRAR was designed for simplicity and direct grounding in annotated bounding boxes, but it is one of several possible spatial-fidelity measures and its relative merits are worth stating explicitly. Pointing-game accuracy, a common Grad-CAM evaluation metric building on class-activation-mapping localization [17], only checks whether the single highest-activation pixel falls inside the target region; it is easy to compute but discards all information about how activation is distributed elsewhere in the frame, and could report a pass even when a large share of total attention lies outside the weapon. Intersection-over-Union between a thresholded activation mask and the ground-truth box is more informative than pointing-game accuracy but is sensitive to the chosen binarization threshold, whereas WRAR uses the continuous heatmap directly and avoids that free parameter. SHAP-based fidelity scores, by contrast, quantify how much a masked region changes the model's output score rather than where activation is spatially concentrated, making them a complementary rather than a substitute measure — which is why this study reports SHAP convergence with Grad-CAM/WRAR as corroborating evidence rather than treating the two as interchangeable. A systematic head-to-head comparison of WRAR against pointing-game accuracy, IoU-based localization metrics, and SHAP fidelity scores on a shared benchmark is identified as a concrete direction for follow-up work rather than resolved within the present study.

VII. FUTURE WORK

Replication on multi-camera, multi-site datasets is the most immediate validation need. If the inverse relationship between full fine-tuning and spatial coherence holds across cameras with different acquisition characteristics, the case for standardizing selective fine-tuning as the default transfer-learning protocol for weapon detection AI becomes significantly stronger. Cross-dataset evaluation would also test whether WRAR is a stable explainability signal or is sensitive to annotation style.

Integrating spatial alignment as a training objective — for instance through penalty terms that discourage activation outside annotated bounding boxes — could simultaneously pursue accuracy and explainability rather than treating them as a post-hoc trade-off. Combining such regularization with the selective fine-tuning strategy used by Model B may push both metrics beyond what either approach achieves alone.

Extending the detection taxonomy to include concealed weapons, improvised devices, and partially visible arms, together with edge-deployment optimization for camera-embedded hardware, represents the primary direction for translating these research findings into practical security infrastructure.

VIII. CONCLUSION

This paper examined how the scope of fine-tuning in a YOLOv8-based weapon detection system shapes both detection quality and the interpretability of model decisions. Three configurations — frozen backbone, selective head and neck adaptation, and full-parameter optimization — were compared on a curated weapon surveillance dataset using complementary detection and explainability metrics.

The key finding is that composite quality — the product of accuracy and spatial transparency — peaks at the intermediate fine-tuning level. Selectively adapting the detection head and neck raises mAP@0.5 from 0.821 to 0.883 and improves WRAR on clear frames from 0.35 to 0.67, while preserving weapon-frame WRAR at 0.88. Fully optimizing all layers pushes mAP slightly higher to 0.901 but drives WRAR on weapon frames down to 0.19, with Grad-CAM and SHAP maps converging on the same conclusion: the model has substituted image-border cues for weapon appearance in its decision process. This collapse is statistically robust ($p < 0.001$; Cohen's $d \approx 1.18$).

Selective fine-tuning emerges as the deployment-ready choice: mAP@0.5 of 0.883, weapon recall of 96%, the lowest missed-detection count, and spatially coherent explanations that withstand operational audit. These results position fine-tuning depth as a substantive architectural variable in security AI design — one that governs not just benchmark scores but whether a deployed system can be responsibly explained, validated, and trusted. The smart security system successfully demonstrated that integrating deep learning-based object detection with real-time surveillance infrastructure provides an efficient, accurate, and scalable solution.

IX. REFERENCES

- [1] D. Bhatt, C. Patel, H. Ghetia, and P. J. Bhatt, "Weapon detection using deep learning: A review," in Proc. IEEE ICCIT, Dhaka, Bangladesh, 2021, pp. 1–6.
- [2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, Venice, Italy, 2017, pp. 618–626.
- [3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 4765–4774.
- [4] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Ultralytics, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>.
- [5] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in Proc. IEEE ICCV, Seoul, South Korea, 2019, pp. 6202–6211.
- [6] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin Transformer," in Proc. IEEE CVPR, New Orleans, LA, USA, 2022, pp. 3202–3211.
- [7] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," in Proc. ACM Conf. Health Inference Learning, 2020, pp. 151–159.
- [8] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in Proc. IJCAI, Melbourne, Australia, 2017, pp. 2662–2670.
- [9] A. Saleem, J. Akhtar, and M. Usman, "Automatic weapon detection in surveillance videos: A survey," Sensors, vol. 21, no. 9, p. 2954, Apr. 2021.
- [10] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in Proc. ACCV, Perth, Australia, 2019, pp. 622–637.
- [11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 779–788.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, Virtual, 2021.
- [13] W. Lv, S. Xu, Y. Zhao, G. Wang, J. Wei, C. Cui, Y. Du, Q. Dang, and Y. Liu, "DETRs beat YOLOs on real-time object detection," in Proc. IEEE CVPR, Seattle, WA, USA, 2024, pp. 16965–16974.
- [14] K. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 2921–2929.