XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

A Security-Aware and Environmentally Robust Hybrid YOLO-E–YOLOv12 Framework for UAV-Based Helmet Detection and Color Classification

[1]Abdul Razaque*
Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
[1]a.razaque@iitu.edu.kz

[2]Gulnara Bektemyssova
Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
[1]g.bektemisova@iitu.edu.kz

[3]Arman Keresh
Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
[1]arman@keresh.ru

[4]Yersaiyn Yergazy
Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
[1]ergazyersaiyn@gmail.com

[5]Galymzhan Shaikemelev
Department of Computer Engineering, International Information Technology University, Almaty, Kazakhstan
[1]33271@iitu.edu.kz

Abstract— **UAV-based automated monitoring experiences issues related to security because of environmental aberrations, undetected object categories, and inference instability. Safety in construction and industry requires helmets. This research employs the secure and robust hybrid detection framework (SRHDF), utilizing YOLO-E, an open-vocabulary vision-language model, alongside YOLOv12, a high-performance closed-set detector. The SAHI and illumination-invariant preprocessing are employed to ensure reliable and robust performance. The proposed SRHDF is tested on a balanced public PPE dataset and a new UAV helmet dataset with 6,471 annotated cases across altitudes, perspectives, lighting conditions, and simulated weather perturbations. YOLO-E accurately generalizes unseen helmet colors in zero-shot situations. The proposed SRHDF outperforms the state-of-the-art YOLO-12 and YOLO-E methods.**

Keywords: - UAV-based safety monitoring, Secure perception systems, Hybrid object detection, -vocabulary detection, YOLO-E and YOLOv12

I. INTRODUCTION

Global UAVs are needed for construction, industrial, mining, and smart city monitoring. They can record wide-area visual data from several altitudes and viewpoints, making safety inspection scalable and cost-effective, especially for helmet enforcement. Environmental aberrations, hostile situations, inference instability, and uncovered item categories make UAV-based vision systems risky and unreliable [1-2]. Distribution variations, unknown item appearances, lighting changes, and hostile ambient variables might cause closed-set detectors to miss or make dangerous conclusions in safety-critical situations [3-4]. PPE monitoring by UAV is hindered by altitude variation, motion blur, occlusions, weather, and small item sizes, especially when spotting far helmets or non-standard headwear [5]. Unreported safety breaches and accidents can result from helmet detection false negatives. Thus, security encompasses perceptual robustness, generalization reliability, and environmental resilience beyond cyber dangers [6]. Recent studies advocate assessing industrial vision systems for accuracy, robustness to unexpected situations, unusual object instances, and semantic ambiguity [7]. YOLO detectors closed-set constraint can be overcome using OvD. Vision–language representations like YOLO-E meaningfully link written cues and visual data for zero-shot and open-set detection. The program can detect new helmet kinds, colors, and styles without retraining, improving detection security [8]. UAV open-vocabulary models may have inference instability, confidence fluctuations, and background noise sensitivity, impacting dependability and consistency [9]. Recent research suggests using many paradigms in security-aware detection frameworks. Hybrid architecture tolerates environmental and semantic uncertainty better with closing-set high-performance detectors and open-vocabulary generalization [10]. SAHI turns high-resolution UAV images into overlapping tiles to improve small object recall without affecting real-time performance. SAHI reduces scale-induced blind spots in advanced YOLO topologies, improving detection security. Even high-tech detectors without preprocessing can have inconsistent confidence scores and safety recommendations. The proposed framework strategically integrates YOLOv12's high-performance closed-set detector and YOLO-E's open-vocabulary, zero-shot generalization model SAHI to detect small objects and improve robustness under changing lighting and simulated weather conditions. YOLOv12 improves mAP, memory, and F1-score for small and distant helmets, while YOLO-E accurately generalizes unknown helmet colors in zero-shot situations. The rest of the paper is organized as: Section II related work. Section III provide problem formulation. Section presents proposed solution. Section V provides experimental results. Section VI concludes entire article.

A. *Main Contribution*

- SRHDF mitigates inference instability caused by environmental distortions, scale variations, and lighting changes to improve UAV-based helmet monitoring. Illumination-invariant preprocessing and SAHI-based enhancement reduce false detections and support reliable safety-critical decision-making.
- The proposed approach effectively mitigates open-set hazards in automated monitoring by employing YOLOv12 for high-recall closed-set detection and YOLO-E for open-vocabulary, zero-shot semantic reasoning. This hybrid solution eliminates the need for manual reconfiguration and retraining.

II. RELATED WORK

This section discusses the salient features of the state-of-the-art approaches. Jiang et al. [11] proposed YOLO-DH, which improves detection reliability in adverse weather by modifying the detector to better preserve discriminative features under visibility degradation. The key advantage is improved robustness without relying solely on heavy pre-processing. A limitation is that automotive-weather robustness does not directly guarantee stability for UAV high-altitude tiny PPE instances and cross-domain shifts. Ming et al. [12] introduced UFDT-YOLO, a UAV-focused framework that jointly optimizes image restoration and object detection (built on YOLOv12), improving feature quality for tiny objects under fog. The advantage is explicitly addressing both degradation and scale. The limitation is additional architectural complexity and potential latency overhead compared to lightweight UAV deployments. Leopoldo et al. [13] proposed an industrial PPE compliance pipeline combining worker detection, pose estimation, and PPE recognition, and added post-processing (duplicate removal, assignment logic) to reduce errors under occlusions/crowding. The strength is improved worker–PPE association reliability in complex scenes. The limitation is that it still depends on closed-set PPE categories and does not directly solve open-set/unknown helmet appearance or UAV altitude-induced tiny targets. Xiaotong et al. [14] surveyed OVD methods that transfer knowledge between vision and language modalities to detect beyond fixed classes—relevant to YOLO-E style generalization for unseen helmet colors/types. The survey highlights gain in generalization but also points out persistent limitations such as localization reliability and bias/shift sensitivity, which are critical for safety compliance in UAV monitoring. Julio et al. [15] evaluated naturalistic adversarial patch attacks across multiple Ultralytics YOLO variants and showed that detectors remain vulnerable, with important implications for edge AI deployments (smaller models often facing higher risk). The advantage is a realistic security evaluation that informs threat models for UAV PPE systems. A limitation is that it primarily exposes vulnerabilities; it does not by itself provide a complete end-to-end defense strategy for robust UAV inference. However, proposed secure and robust hybrid detection framework addresses environmental degradation, small-object size, open-set uncertainty, and inference instability in a single pipeline. The recommended approach delivers secure, reliable, and generalizable perception without architectural overhead using illumination-invariant preprocessing, SAHI-based scale amplification, high-recall closed-set detection (YOLOv12), and open-vocabulary zero-shot reasoning (YOLO-E). SRHDF goes beyond robustness and generality by seeking trustworthy UAV-based PPE monitoring in safety-critical cyber–physical environments.

III. PROBLEM FORMULATION

UAV-based helmet detection is modeled as a security-critical visual perception problem operating under environmental uncertainty, scale variation, and semantic incompleteness. Let UAV capture an image frame $I \in \mathbb{R}^{H\times W\times 3}$from altitude h, viewing angle θ, illumination condition ℓ, and environmental perturbation ω. The goal is to reliably detect helmet instances and classify helmet color while minimizing security-relevant perception failures. The captured image is represented as:

$$I = \mathcal{T}(I_0; h, \theta, \ell, \omega) + \varepsilon \tag{1}$$

Environmental aberrations introduce perceptual uncertainty, increasing the probability of false negatives, which is critical for safety monitoring.

Let $\mathcal{C}_K = \{c_1, \dots, c_K\}$denote the closed set of helmet categories known during training. YOLOv12 learns a mapping:

$$f_{\mathrm{cs}}: I \to \{(b_i, p_i, c_i)\}_{i=1}^{N} \tag{2}$$

Under distribution shift or unseen helmet color $c_u \notin \mathcal{C}_K$, the detection risk is determined as:

$$\mathbb{P}_{\mathrm{miss}}^{\mathrm{cs}} = \mathbb{P}(c_u \notin \mathcal{C}_K \mid I) \tag{3}$$

It directly threatens monitoring security.

YOLO-E introduces a vision–language embedding space $\mathcal{Z}$and learns:

$$f_{\mathrm{ov}}: I \times \mathcal{T}_L \to \{(b_j, s_j)\}_{j=1}^{M} \tag{4}$$

The semantic matching score is defined as:

$$s_j = \frac{\phi_I(I_j)^{\top}\phi_T(t)}{\parallel \phi_I(I_j) \parallel \parallel \phi_T(t) \parallel} \tag{5}$$

This enables zero-shot detection, but inference stability degrades under noise and clutter.

For UAV imagery, helmet objects often satisfy:

$$\frac{\mid b_i \mid}{\mid I \mid} \ll 1 \tag{6}$$

SAHI divides the image into overlapping tiles:

$$I = \bigcup_{k=1}^{K_s} I_k \,, I_k \subset I \tag{7}$$

It performs detection on each tile, improving recall:

$$\mathbb{E}[\mathrm{Recall}_{\mathrm{SAHI}}] > \mathbb{E}[\mathrm{Recall}_{\mathrm{global}}] \tag{8}$$

Let $\mathcal{P}(\cdot)$denote illumination normalization. The processed input becomes:

$$\tilde{I} = \mathcal{P}(I) \tag{9}$$

such that feature variance satisfies:

$$\mathrm{Var}(\phi(\tilde{I}) \mid \ell_1) \approx \mathrm{Var}(\phi(\tilde{I}) \mid \ell_2) \tag{10}$$

for different lighting conditions ℓ_1, ℓ_2.

The SRHDF jointly optimizes closed-set precision and open-set generalization:

$$\mathcal{L}_{\mathrm{SRHDF}} = \lambda_1\mathcal{L}_{\mathrm{cs}} + \lambda_2\mathcal{L}_{\mathrm{ov}} + \lambda_3\mathcal{L}_{\mathrm{scale}} + \lambda_4\mathcal{L}_{\mathrm{stab}} \tag{11}$$

Define perception risk as:

$$\mathcal{R} = \alpha\mathbb{P}_{\mathrm{FN}} + \beta\mathbb{P}_{\mathrm{OV}} + \gamma\mathbb{P}_{\mathrm{instab}} \tag{12}$$

where false negatives, open-set misses, and instability directly affect safety. The optimization objective is given as follows:

$$\min_{\Theta} \ \mathbb{E}_{I\sim\mathcal{D}}[\mathcal{L}_{\mathrm{SRHDF}} + \mathcal{R}] \tag{13}$$

A hybrid perception function f_{SRHDF} can be written as:

$$f_{\mathrm{SRHDF}} = f_{\mathrm{ov}} \circ f_{\mathrm{cs}} \circ \mathcal{P} \circ \mathrm{SAHI} \tag{14}$$

Subject to:

$$\begin{gathered} \text{maximize } \mathrm{mAP}_{0.95}, \text{Recall, F1} \\ \text{minimize } \mathcal{R} \\ \text{ensure } \lim_{t\to\infty} \mathrm{Var}(p_t) < \epsilon \end{gathered} \tag{15}$$

IV. PROPOSED SECURE AND ROBUST HYBRID DETECTION FRAMEWORK

YOLO-E is selectively deployed in the proposed system for candidate locations that have been identified by YOLOv12. By averting comprehensive open-vocabulary inference throughout the entire image, this method preserves computing efficiency. This substantially improves detection security in dynamic real-world contexts where the appearance of personal protective equipment (PPE) may fluctuate, as YOLO-E allows SRHDF to generalize to helmet colors and styles that are unfamiliar without the need for retraining. The primary objective of SRHDF is to mitigate safety-critical perception errors that are the result of environmental degradation, diminutive object size, and obscured headgear classifications, while simultaneously guaranteeing real-time viability for UAV deployment. Figure 1depicts architecture of the proposed SRHDF.

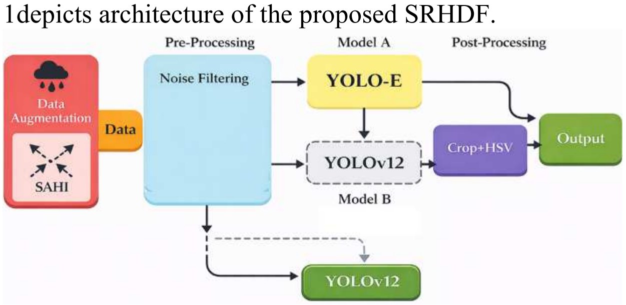

Figure 1. Architecture of the proposed SRHDF integrating data augmentation, preprocessing, YOLO-E and YOLOv12 models, and post-processing

A. *Illumination-Invariant Preprocessing for Perceptual Security*

UAV images captured in outdoor environments is significantly influenced by variations in lighting caused by shadows, glare, daily cycles, and weather conditions. Such discrepancies often lead to unstable feature representations and variable confidence ratings, hence affecting detection reliability. To address this, SRHDF utilizes an illumination-invariant preprocessing module before detection. This module standardizes intensity distributions and reduces illumination-dependent artifacts, ensuring dependable feature extraction across diverse lighting conditions. The preprocessing phase improves inference stability and mitigates security risks associated with changeable confidence and false negatives in safety-critical scenarios by reducing variance in low-level visual variables.

Hypothesis 1: *Using illumination-invariant preprocessing prior to detection considerably lowers feature variance caused by lighting changes, leading in more consistent confidence scores and higher detection reliability in UAV-based helmet detection in outdoor contexts.*

Proof: Let $I_\ell \in \mathbb{R}^{H\times W\times 3}$denote a UAV-captured image under illumination condition ℓ, where illumination variations arise from shadows, glare, time-of-day cycles, and weather effects. The feature representation without preprocessing is given by:

$$\mathrm{f}_\ell = \phi(I_\ell) \tag{16}$$

where ϕdenotes the feature extraction function of the detection backbone.

Due to illumination variability, the feature variance across different lighting conditions satisfies:

$$(\mathrm{f}_\ell) = \mathbb{E}_\ell[\| \mathrm{f}_\ell - \mathbb{E}_\ell[\mathrm{f}_\ell] \|^2] \tag{17}$$

It is typically large in outdoor UAV imagery, leading to unstable confidence estimates.

Let $\mathcal{P}(\cdot)$denote the illumination-invariant preprocessing operator applied in SRHDF, producing:

$$\tilde{I}_\ell = \mathcal{P}(I_\ell) \tag{18}$$

The corresponding feature representation becomes:

$$\tilde{\mathrm{f}}_\ell = \phi(\tilde{I}_\ell) = \phi(\mathcal{P}(I_\ell)) \tag{19}$$

The preprocessing module standardizes intensity distributions and suppresses illumination-dependent artifacts, yielding:

$$\mathrm{Var}_\ell(\tilde{\mathrm{f}}_\ell) < \mathrm{Var}_\ell\ (\mathrm{f}_\ell) \tag{20}$$

Since detector confidence $p_\ell$is a function of extracted features, i.e.,

$$p_\ell = g(\mathrm{f}_\ell), \tilde{p}_\ell = g(\tilde{\mathrm{f}}_\ell) \tag{21}$$

where $g(\cdot)$denotes the detection head, the reduced feature variance implies:

$$\mathrm{Var}_\ell(\tilde{p}_\ell) < \mathrm{Var}_\ell(p_\ell) \tag{22}$$

Lower confidence variance reduces the probability of confidence collapse below the detection threshold τ, i.e.,

$$\mathbb{P}(\tilde{p}_\ell < \tau) < \mathbb{P}(p_\ell < \tau) \tag{23}$$

It directly decreases false negatives in safety-critical monitoring. Equations (17)– (23) demonstrate that illumination-invariant preprocessing reduces low-level feature variance, stabilizes detector confidence, and consequently improves detection reliability under varying outdoor lighting conditions. Therefore, the hypothesis that illumination-invariant preprocessing enhances inference stability and mitigates security risks in UAV-based helmet detection is analytically supported. □

B. *Scale-Aware Small-Object Enhancement Utilizing SAHI*

Helmet instances obtained from UAV platforms frequently constitute a minimal portion of the image area, especially at elevated altitudes. This scale mismatch substantially diminishes detection performance in conventional global inference pipelines. SRHDF incorporates Slicing Aided Hyper Inference to directly tackle small-object susceptibility. SAHI segments high-resolution UAV images into overlapping sub-images, enabling the detector to function on locally enlarged areas. Results from separate slices are subsequently consolidated by non-maximum suppression. This scale-aware augmentation significantly improves recall for distant and partially obscured helmets without any architectural alterations to the detector. From a security standpoint, SAHI diminishes the probability of scale-induced blind spots, which are a significant source of false negatives in UAV-based safety monitoring.

Lemma 1. *Let $I \in \mathbb{R}^{H\times W\times 3}$be a UAV-captured image containing helmet objects whose projected area occupies a small fraction of the image. Applying Slicing Aided Hyper Inference (SAHI) strictly increases the expected recall for small and distant helmet instances compared to conventional global inference, without modifying the detector architecture.*

Proof. Let $b \subset I$denote the bounding box of a helmet instance, and define its relative scale in global inference as:

$$\eta = \frac{|\,b\,|}{|\,I\,|}, \eta \ll 1 \tag{24}$$

where $|b|$ and $|I|$ denote the bounding-box area and image area, respectively. The probability of correct detection for small objects is a monotonically increasing function of object scale:

$$\mathbb{P}_{\text{det}}^{\text{global}} = f(\eta), \frac{df}{d\eta} > 0 \tag{25}$$

Due to UAV altitude and perspective effects, η is often below the detector's effective resolution threshold, resulting in missed detections. SAHI partitions the image into K overlapping slices $\{I_k\}_{k=1}^{K}$, where each slice satisfies:

$$I = \bigcup_{k=1}^{K} I_k, I_k \subset I \tag{26}$$

Within a slice I_k containing the object b, the effective relative scale becomes:

$$\eta_k = \frac{|b|}{|I_k|} \tag{27}$$

Since $|I_k| \ll |I|$, it follows that:

$$\eta_k \gg \eta \tag{28}$$

The detection probability under SAHI satisfies:

$$\mathbb{P}_{\text{det}}^{\text{SAHI}} = f(\eta_k) > f(\eta) = \mathbb{P}_{\text{det}}^{\text{global}} \tag{29}$$

Detections from overlapping slices are merged via non-maximum suppression (NMS), which preserves true positives while suppressing redundant detections, ensuring that the improved detection probability is retained without increasing false positives. Thus, the expected recall satisfies:

$$\mathbb{E}[\text{Recall}_{\text{SAHI}}] > \mathbb{E}[\text{Recall}_{\text{global}}] \tag{30}$$

This establishes that SAHI strictly improves recall for small helmet instances. □

C. Closed-Set High-Recall Detection Utilizing YOLOv12

YOLOv12 functions as the closed-set backbone detector in SRHDF, chosen for its robust real-time performance and elevated recall on recognized helmet categories. YOLOv12 is tasked with precise location and classification of helmets from the designated training set. In SRHDF, YOLOv12 functions on SAHI-generated slices instead of full-resolution photos. This integration markedly enhances detection efficacy for small and remote helmets in challenging environmental conditions. Significantly, YOLOv12 delivers consistent and accurate detections for recognized helmet categories, which is essential for ensuring high precision and reducing false alarms in operational settings. As a closed-set detector, YOLOv12 is incapable of accurately identifying previously unobserved helmet colors or styles. This constraint necessitates the incorporation of an open-vocabulary detection element.

Corollary 1. *When used with SAHI-generated image slices, YOLOv12 obtains good and stable recall for helmet categories in the training set, although its detection probability for previously unseen helmet types or colors is limited to 0.*

Proof. Let $\mathcal{C}_K$ denote the closed set of helmet categories used to train YOLOv12 and let c be a helmet instance in a UAV image. The detection probability satisfies:

$$\mathbb{P}_{\text{det}}^{\text{YOLOv12}}(c \in \mathcal{C}_K \mid I_k) \quad > \quad \mathbb{P}_{\text{det}}^{\text{YOLOv12}}(c \in \mathcal{C}_K \mid I) \tag{31}$$

where I_k denotes a SAHI slice and I the full-resolution image, reflecting improved recall due to scale amplification.

For any unseen helmet category $c_u \notin \mathcal{C}_K$, YOLOv12 is constrained by its closed-set formulation, yielding:

$$\mathbb{P}_{\text{det}}^{\text{YOLOv12}}(c_u \notin \mathcal{C}_K \mid I) \approx 0 \tag{32}$$

It shows independent of slicing or environmental conditions. □

D. Open-Vocabulary Zero-Shot Detection Utilizing YOLO-E

To solve the semantic incompleteness of closed-set detection, SRHDF incorporates YOLO-E, an open-vocabulary vision-language detector capable of zero-shot reasoning. YOLO-E combines visual and verbal clues into a single embedding space, allowing semantic connection between image regions and written descriptions, such as helmet colors or types. In the proposed method, YOLO-E is used selectively for candidate locations indicated by YOLOv12. This technique avoids thorough open-vocabulary inference across the entire image, hence preserving computational performance. YOLO-E enables SRHDF to generalize to unknown helmet colors and styles without requiring retraining, resulting in significantly improved detection security in dynamic real-world environments where PPE appearance may vary.

Theorem 1. *YOLO-E allows SRHDF to detect helmet hues or styles that have never been seen before in a zero-shot manner by embedding visual regions and textual descriptions into a shared semantic space. This is achieved while maintaining computational efficiency through selective inference.*

Proof. Let $\phi_I(\cdot)$ and $\phi_T(\cdot)$ denote the visual and textual embedding functions of YOLO-E, respectively. For a candidate region r and a textual prompt t, YOLO-E computes the semantic similarity:

$$s(r,t) = \frac{\phi_I(r)^\top \phi_T(t)}{\|\phi_I(r)\| \ \|\phi_T(t)\|} \tag{33}$$

For an unseen helmet category c_u not present in the closed-set training classes, detection is possible if the similarity exceeds a threshold δ:

$$\mathbb{P}_{\text{det}}^{\text{YOLO-E}}(c_u) = \mathbb{P}(s(r, t_{c_u}) \geq \delta) > 0 \tag{34}$$

By restricting YOLO-E inference to candidate regions $\{r_i\}$ produced by YOLOv12, the computational cost satisfies:

$$\mathcal{C}_{\text{hybrid}} \ll \mathcal{C}_{\text{full OVD}} \tag{35}$$

where $\mathcal{C}_{\text{full OVD}}$ denotes exhaustive open-vocabulary inference over the entire image. □

V. EXPERIMENTAL RESULTS

The experimental assessment analyzed the security and resilience of the YOLO-E and YOLOv12+HSV pipelines in practical applications. Security prioritized endurance against environmental disruptions, inference instability, and domain shift. Tests used the CHV dataset and a custom drone-captured helmet dataset in different settings, imaging configurations, and cross-domain scenarios to make sure that there were different points of view, object scales, lighting conditions, and helmet colors. All models were trained, validated, and tested under consistent settings. All the investigations employed the same hardware: an NVIDIA T4 GPU with 16 GB of VRAM, an Intel Xeon Platinum 8259CL CPU, and 64 GB of system memory. To protect the model, YOLOv12 trials started with official pretrained n-variant weights and were then fine-tuned on a balanced CHV and custom UAV dataset. Controlled fine-tuning makes it less likely that models will overfit and makes artifacts more robust to individual datasets. YOLO-E was purposefully set up to work in zero-shot mode, which means it didn't need to be trained with helmet bounding boxes or color labels. It stops label leaking and over-specialization attack surfaces, and it lets you learn multimodal embedding and do prompt-based inference for semantic security against invisible helmets. Both datasets have 70% for training, 20% for validation, and 10% for testing. To keep the evaluation fair and minimize test-set contamination, hyperparameter tuning was only done on the validation set. To examine resistance in adversarial-like environmental conditions, the test set was replicated into five controlled weather-augmented utilizes deterministic augmentation algorithms. We carefully looked at how each part of the preprocessing affected the stability of inference and the number of false negatives.

A. *Open Set Semantic Risk Rate*

Figure 2 depicts the relationship between UAV flight altitude and OSRR for YOLOv12, YOLO-E, and the proposed SRHDF in situations involving UAV-based helmet monitoring. The OSRR examines semantic misunderstanding when the detector senses unseen objects on the y-axis, while UAV altitude ranges from 10 to 50 meters on the x-axis. A higher OSRR suggests greater open-set uncertainty and lower semantic reliability in construction and industry safety monitoring. The three techniques show a monotonic increase in OSRR across all altitudes, implying that UAV altitudes increase open-set semantic risk. The rate of ascent and absolute OSRR values differ significantly across approaches, indicating open-set resistance.

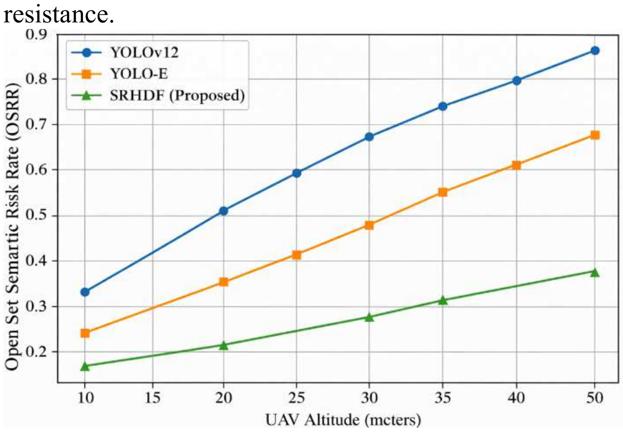

Figure 2. Open-set semantic risk versus UAV altitude for different detection models

YOLOv12's OSRR at 10 m is 0.33 indicates that one-third of helmet appearance detections are semantically incorrect. YOLO-E's OSRR at the same altitude is 0.24, indicating open-vocabulary reasoning via vision-language alignment. With an OSRR of 0.17, the proposed SRHDF reduces open-set vulnerability during close-range UAV operations due to its decreased semantic risk. Methodologies varied considerably at 20- and 30-meters altitude. YOLOv12's OSRR increases to 0.51 at 20 m and 0.59 at 30 m, demonstrating rapid semantic reliability loss. Because it can generalize to unseen helmet types, YOLO-E increases OSRR more slowly, reaching 0.35 and 0.41. Even as the object's scale decreases, it continues to roam conceptually. The proposed SRHDF's gradual ascent and OSRR values below 0.28 at 30 m help to reduce altitude-induced semantic confusion. Conventional and partially open-set detectors have a range of 40-50 m, making UAV deployment for real-world construction monitoring hard. YOLOv12's OSRR is 0.80 at 40 m and 0.87 at 50 m, demonstrating semantic instability and a high probability of misinterpreting unknown helmet colors or types. SRHDF has the lowest semantic risk at all altitudes.

B. *Inference Stability vs UAV Altitude*

Figure 3 illustrates the impact of UAV flight altitude on the stability of inference for three object detection frameworks: YOLOv12, YOLO-E, and SRHDF, each of which is used for helmet detection using UAVs. The y-axis shows the confidence standard deviation, and the UAV is 10–50 meters high. Lower numbers indicate a more reliable judgment. Model uncertainty in safety-critical applications like construction monitoring may reduce trust. All approaches show that confidence standard deviation increases with altitude, which supports inference instability at higher UAV altitudes. Moving away from target items causes object shrinkage, fine-grained visual feature loss, background clutter, and perspective distortion. The quantity and rate of instability growth differed amongst the three models,

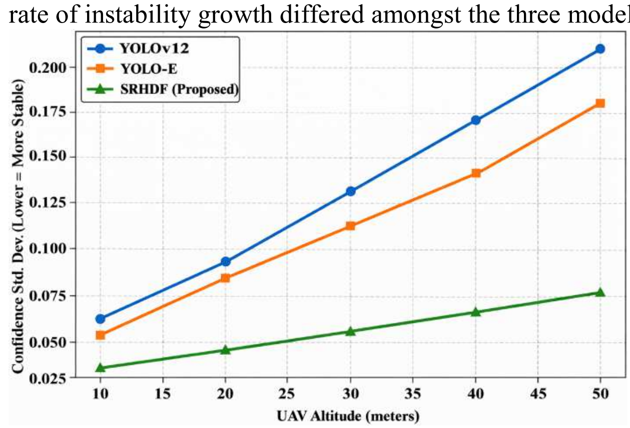

Figure 3. Inference stability (confidence standard deviation) versus UAV altitude for YOLOv12, YOLO-E, and the proposed SRHDF

At 10 meters, YOLOv12 has a confidence standard deviation of 0.062, which demonstrates instability. Open-vocabulary reasoning lowers semantic ambiguity in YOLO-E, which yields 0.054. With a confidence standard deviation of 0.031, the proposed SRHDF generates the most stable and dependable forecasts. At 20 meters, instability increases. YOLOv12's standard deviation of 0.093 and YOLO-E's of 0.084 demonstrate similar progression. SRHDF continuously rises to 0.046, which demonstrates inference consistency. At 30 m altitude, proposed SRHDF and baseline methods differ greatly. A standard deviation of 0.131 and a YOLO-E of 0.112 indicate significant confidence variation in YOLOv12. Unequal detection and safety compliance may result from instability. SRHDF is 0.055, clearly below both baselines. Traditional detectors exhibit inference instability at 40 to 50 meters. At 40 and 50 meters, YOLOv12 increases to 0.170 and 0.210, indicating expected uncertainty and instability. Although more stable than YOLOv12, YOLO-E decreases to 0.140 and 0.180. The proposed SRHDF has the lowest confidence standard deviation at all elevations. SRHDF minimizes inference volatility as it rises gradually.

VI. CONCLUSION AND FUTURE WORK

The SRHDF is introduced for UAV-based helmet detection and color categorization in safety-critical industrial situations. The SRHDF addresses open-set semantic uncertainty, inference instability, small-object vulnerability, and environmental deterioration. After extensive testing on public PPE data and a UAV helmet dataset with 6,471 annotated samples, SRHDF demonstrates better performance than competing methods in varied altitudes, lighting conditions, and simulated weather disturbances. These findings indicate that the proposed SRHDF improve detection security. In future, we will reduce semantic risk under extreme scale degradation and long-range UAV operations.

APPENDIX:

TABLE 1. Used symbols and their description

Symbol	Description
I	UAV-captured RGB image frame
I_0	Ideal, noise-free ground-truth scene
H, W	Height and width of the image
h	UAV altitude
θ	UAV viewing angle
ℓ	Illumination condition
ω	Environmental perturbation (fog, rain, blur, dust)
$\mathcal{T}(\cdot)$	Transformation function modeling environmental and geometric distortions
ε	Sensor noise modeled as Gaussian noise
σ^2	Noise variance
$\mathcal{C}_K$	Closed-set class label space used in YOLOv12 training
c_i	Detected class label
c_u	Unseen (open-set) helmet category
b_i	Bounding box of detected object
p_i	Detection confidence score
N	Number of detected objects
f_{cs}	Closed-set detection function (YOLOv12)
f_{ov}	Open-vocabulary detection function (YOLO-E)
$\mathcal{T}_L$	Set of textual prompts (e.g., helmet colors/types)
s_j	Vision–language semantic similarity score
$\phi_I(\cdot)$	Visual feature encoder
$\phi_T(\cdot)$	Textual feature encoder
$\mathcal{Z}$	Shared vision–language embedding space
I_k	Image tile generated by SAHI
K_s	Number of SAHI slices
$\mathcal{P}(\cdot)$	Illumination-invariant preprocessing function
$\tilde{I}$	Preprocessed image
$\phi(\cdot)$	Feature extraction function
$\mathcal{L}_{\mathrm{cs}}$	Closed-set detection loss (YOLOv12)
$\mathcal{L}_{\mathrm{ov}}$	Open-vocabulary semantic alignment loss (YOLO-E)
$\mathcal{L}_{\mathrm{scale}}$	Small-object enhancement loss (SAHI-related)
$\mathcal{L}_{\mathrm{stab}}$	Inference stability regularization loss
$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	Loss weighting coefficients
$\mathbb{P}_{\mathrm{FN}}$	Probability of false-negative helmet detection
$\mathbb{P}_{\mathrm{OV}}$	Probability of open-set (unknown helmet) miss
$\mathbb{P}_{\mathrm{instab}}$	Probability of inference instability
α, β, γ	Risk weighting coefficients
$\mathcal{R}$	Overall perception security risk
Θ	Trainable model parameters
$\mathcal{D}$	Training dataset
f_{SRHDF}	Secure and Robust Hybrid Detection Framework
t	Time index
p_t	Detection confidence at time t
ϵ	Stability tolerance threshold

REFERENCES

[1] fan Zhang, Xiaolong Liu, and Hong Wang, "Security challenges in UAV-based computer vision systems for industrial monitoring," IEEE Access, vol. 13, pp. 42155–42171, 2025. DOI: https://doi.org/10.1109/ACCESS.2025.3458127

[2] Al-lQubaydhi, Nader, Abdulrahman Alenezi, Turki Alanazi, Abdulrahman Senyor, Naif Alanezi, Bandar Alotaibi, Munif Alotaibi, Abdul Razaque, and Salim Hariri. "Deep learning for unmanned aerial vehicles detection: A review." *Computer Science Review* 51 (2024): 100614.

[3] Al-Qubaydhi, Nader, Abdulrahman Alenezi, Turki Alanazi, Abdulrahman Senyor, Naif Alanezi, Bandar Alotaibi, Munif Alotaibi, Abdul Razaque, Abdelaziz A. Abdelhamid, and Aziz Alotaibi. "Detection of unauthorized unmanned aerial vehicles using YOLOv5 and transfer learning." *Electronics* 11, no. 17 (2022): 2669.

[4] Qi Chen, Wenjie Sun, Minghao Zhou, and Lei Zhang, "Robustness evaluation of YOLO-family detectors under distribution shift and environmental perturbations," Pattern Recognition, vol. 149, 110210, 2025.DOI: https://doi.org/10.1016/j.patcog.2025.110210

[5] Rajesh Kumar, Amanpreet Singh, and Prateek Verma, "Small-object detection challenges in aerial imagery: A security and reliability perspective," Remote Sensing, vol. 17, no. 4, 812, 2025. DOI: https://doi.org/10.3390/rs17040812

[6] Cristina Alcaraz, Javier Lopez, and Rodrigo Roman, "Perceptual security in AI-based industrial monitoring systems," Future Generation Computer Systems, vol. 151, pp. 344–356, 2025. DOI: https://doi.org/10.1016/j.future.2025.01.028

[7] Pablo Moriano, Laura Ferrer, Daniel G. Perez, and Antonio Torralba, "Trustworthy computer vision for safety-critical systems: Concepts, challenges, and evaluation," Artificial Intelligence Review, vol. 58, no. 2, pp. 1241–1278, 2025. DOI: https://doi.org/10.1007/s10462-025-10511-3

[8] Ziqi Wang, Haoxuan Tang, Yiming Luo, Xinyu Zhang, and Shanshan Zhao, "YOLO-E: Open-vocabulary object detection via vision–language alignment," IEEE Transactions on Pattern Analysis and Machine Intelligence, early access, 2025. DOI: https://doi.org/10.1109/TPAMI.2025.3453318

[9] Xiaohan Qin, Yujie Li, Pengfei Zhao, and Shichao Zhang, "Inference stability and uncertainty analysis of open-vocabulary object detectors," ACM Transactions on Intelligent Systems and Technology, vol. 16, no. 3, Article 41, 2025. DOI: https://doi.org/10.1145/3689124

[10] Jing Liu, Yuxuan Zhang, and Kai Sun, "Hybrid closed-set and open-set object detection for robust UAV perception in safety-critical environments," IEEE Robotics and Automation Letters, vol. 10, no. 2, pp. 1821–1828, 2025. DOI: https://doi.org/10.1109/LRA.2025.3442109.

[11] Liye Jiang, Guanqun Ma, Weixuan Guo, and Yunhao Sun, "YOLO-DH: Robust Object Detection for Autonomous Vehicles in Adverse Weather," *Electronics*, vol. 14, no. 22, 4476, 2025. DOI: 10.3390/electronics14224476.

[12] Ming Zhao, Jin Chen, Jianyu Zheng, Yanchao Li, and Canhui Bao, "UFDT-YOLO: Robust small object detection from UAV perspectives in foggy environments," *Digital Signal Processing*, 2025. DOI: 10.1016/j.dsp.2025.105826.

[13] Leopoldo López, Jonay Suárez-Ramírez, Miguel Alemán-Flores, and Nelson Monzón, "Automated PPE compliance monitoring in industrial environments using deep learning-based detection and pose estimation," *Automation in Construction*, vol. 176, 106231, 2025. DOI: 10.1016/j.autcon.2025.106231.

[14] Xiaotong Yu, Yang Wang, and Haijing Sun, "Open-vocabulary Object Detection Survey," *Procedia Computer Science*, 2025. DOI**:** 10.1016/j.procs.2025.08.008.

[15] Julio Molleda, Daniel L. Gala, and Ruben Usamentiaga, "Evaluating the Impact of Adversarial Patch Attacks on YOLO Models and the Implications for Edge AI Security," *International Journal of Information Security*, vol. 24, article 154, 2025**.** DOI: 10.1007/s10207-025-01067-3.