

Recibido 16 Jun 2026

ReCIBE, Año 15 No. 3, Sep 2026

Aceptado 14 Sep 2026

Robust Reliability for UAV Detection in Edge AI: A Data-Centric Framework via Structural–Semantic Curation, Maximin Selection, and Raspberry Pi Validation

Confiabilidad robusta para la detección con UAV en Edge AI: un marco centrado en datos mediante curación estructural-semántica, selección maximin y validación en Raspberry Pi

.

Luis Ángel Olvera Olvera¹
lolvera68@alumnos.uaq.mx

Luis Antonio Díaz¹
luis.diaz@uaq.edu.mx

Jorge Luis Pérez¹
jorge.luis.perez@uaq.edu.mx

¹ Universidad Autónoma de Querétaro, Querétaro, México

Abstract

UAV-based human detection is vulnerable to domain shift and scale degradation, while average metrics can conceal failures in challenging conditions. This study evaluates whether data-centric optimization improves the worst-case robustness of lightweight detectors without increasing model complexity. The proposed Data-Centric AI (DCAI) framework combines pHash/DCT deduplication, teacher–student semantic auditing with YOLO11x, and controlled oversampling. A hierarchical Design of Experiments through ablation (R0–R4) evaluates data configurations using minimum-recall gating and multicriteria selection under the Wald maximin criterion. Average improvements are modest (approximately 1%), but worst-case gains reach +5.95% in precision and +1.36% in recall. A fixed-effects GLM/ANOVA identifies dominant dataset effects and metric-dependent scenario effects, whereas seed effects are not statistically significant ($p > 0.05$). Deployment on Raspberry Pi 5 using NCNN demonstrates feasibility under edge constraints. Across 5–25 m, the selected hybrid model N10 achieves precision of 0.939–0.968 and recall of 0.939–0.978, with approximately 6.86 FPS and 109.47 ms mean inference time. These findings support data curation as a means of improving robustness across the evaluated domains while retaining a lightweight architecture.

Keywords: Data-centric AI; UAV human detection; dataset curation; semantic auditing; design of experiments; perceptual hashing; edge AI.

Resumen

La detección de personas mediante UAV es vulnerable al cambio de dominio y a la degradación por escala, mientras que las métricas promedio pueden ocultar fallos en condiciones difíciles. Este estudio evalúa si la optimización centrada en datos mejora la robustez en el peor caso de detectores ligeros sin aumentar la complejidad del modelo. El marco propuesto de inteligencia artificial centrada en datos (DCAI) combina deduplicación mediante pHash/DCT, auditoría semántica profesor–estudiante con YOLO11x y sobremuestreo controlado. Un diseño de experimentos jerárquico mediante ablaciones (R0–R4) evalúa configuraciones de datos utilizando una restricción de exhaustividad mínima y selección multicriterio bajo el criterio maximin de Wald. Las mejoras promedio son modestas, aproximadamente del 1%, pero las ganancias en el peor caso alcanzan +5.95% en precisión y +1.36% en exhaustividad. Un GLM/ANOVA de efectos fijos identifica efectos dominantes del conjunto de datos y efectos del escenario dependientes de la métrica, mientras que los efectos de la semilla no son estadísticamente significativos ($p > 0.05$). El despliegue en Raspberry Pi 5 con NCNN demuestra viabilidad en dispositivos de borde. Entre 5 y 25 m, el modelo híbrido seleccionado N10 alcanza precisión de 0.939–0.968 y exhaustividad de 0.939–0.978, con aproximadamente 6.86 FPS y 109.47 ms de inferencia promedio. Los resultados respaldan la depuración de datos para mejorar la robustez en los dominios evaluados manteniendo una arquitectura ligera.

Palabras clave: inteligencia artificial centrada en datos; detección de personas mediante UAV; depuración de datos; auditoría semántica; diseño de experimentos; hashing perceptual; inteligencia artificial en el borde.

1. Introduction

Single-stage detectors such as YOLO support efficient inference, but performance depends on the match between training and deployment distributions (Zhu et al., 2022). UAV human detection combines scale degradation, motion, and limited embedded computing resources. Missed detections under these conditions motivate evaluation of worst-case performance alongside average metrics.

Architectural approaches address small objects through greater capacity and multi-scale processing, often at additional computational cost (Bhusnoor et al., 2024; Hao et al., 2022; Tan et al., 2025). Data-Centric AI (DCAI) instead targets visual redundancy, annotation noise, and cross-domain consistency (Bochkovskiy et al., 2020; Jegham et al., 2024; Zha, 2023).

We evaluate whether curating hybrid training data improves the robustness of a fixed lightweight detector. Hybrid datasets and constrained deployment motivate this question (Hegiste, 2024; Li et al., 2023; Rey et al., 2025), alongside existing detection datasets and UAV systems (Akshatha & others., 2023; Lai & others., 2024; Lin et al., 2014; Taiana et al., 2013; Zhu et al., 2022).

Figure 1 links data acquisition, deduplication, semantic auditing, and controlled oversampling to ablation rounds R0–R4. Recall gating and multicriteria ranking select candidates for statistical analysis and Raspberry Pi 5 distance evaluation.

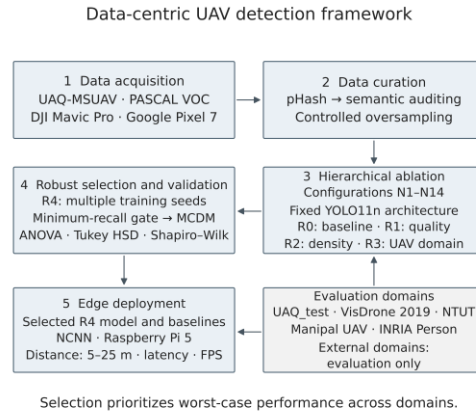


Figure 1. 1General diagram of the proposed DCAI process, from data acquisition to deployment. Source: Own elaboration

2. Related Work

2.1 Small Object Detection in UAVs from a Model-Centric Perspective

Attention mechanisms, multi-scale processing, and high-resolution pipelines improve small-object detection in UAV imagery (Cheng et al., 2024; Zhang et al., 2025). These architectural advances motivate a complementary evaluation of training-data quality with model capacity held constant.

2.2 Data-Centric AI (DCAI) and Training Dataset Quality

DCAI prioritizes dataset quality (OALA, 2023; Zha, 2023), while Confident Learning addresses label errors (Northcutt et al., 2021). Their relevance to UAV detection motivates combining structural and semantic refinement within a controlled evaluation (Cheng et al., 2024).

2.3 Structural Integrity and Deduplication through Perceptual Hashing

DCT-based perceptual hashing detects similar images despite compression or scaling (Zauner, 2010). It can reduce repeated UAV frames before training (Zha, 2023).

2.4 Semantic Auditing, Label Noise, and Signal Validation

Teacher–student methods use a higher-capacity model to assess noisy labels (Tanaka et al., 2018). Here, semantic auditing is evaluated through ablation alongside the architectural approaches prevalent in UAV detection (Cheng et al., 2024).

2.5 Controlled Oversampling and Compensation for Refinement-Induced Scarcity

Oversampling after label refinement tests whether restoring sample density benefits learning without replicating rejected annotations (Northcutt et al., 2021). Its contribution is evaluated through hierarchical DoE (Fontana, 2024).

2.6 Robust Evaluation, Multi-Domain Validation, and Multicriteria Selection

Multi-domain evaluation exposes failures hidden by averages (Gulrajani & Lopez-Paz, 2021). MCDM combines metrics (Triantaphyllou, 2000), while maximin selection emphasizes the weakest evaluated domain (Ben-Tal et al., 2009). Multiple seeds assess sensitivity to training initialization (Feurer & Hutter, 2020).

2.7 Research Gap and Positioning of the Proposal

The contribution is a joint evaluation of deduplication, semantic auditing, and oversampling under fixed architecture, using hierarchical ablation and robust selection (Karczmarczyk & others., 2025). Table 1 positions this combination relative to prior work.

2.8 Comparative Synthesis of the State of the Art

Table 1. 1Comparative synthesis of the state of the art.

Criteria	M1	M2	M3	M4	DCAI proposed
Paradigm	Model	Model	Data	Data	Data
UAV Domain	Yes	Yes	No	No	Yes
Data Curation	No	No	Yes	Yes	Yes
Semantic Auditing	No	No	Yes	No	Yes
Multi-Domain Eval.	No	No	No	No	Yes
Statistical Analysis	No	No	Yes	Yes	Yes
Arch. Modification	Yes	Yes	No	No	No

Note. M1–M4 identify the comparison methods described below. DCAI = data-centric artificial intelligence.

Models:

M1: CSSDet (Cheng et al., 2024).

M2: YOLO-MARS (Zhang et al., 2025).

M3: Confident Learning (Northcutt et al., 2021).

M4: Perceptual image hashing (Zauner, 2010).
DCAI-UAV: Our Proposed Framework.

3. Robust Multicriteria Framework (MCDM) for Round 4

MCDM applies baseline-referenced recall gating before a score combining worst-case and mean performance. The reproduced R4 scores use all six evaluation datasets: COCO_TEST, DOMINIO_DRON, IRINA, MANIPAL_UAV, NTUT, and VISDRONE. Thus, the selection covers both ground and aerial domains.

In R0–R3, 14 configurations are evaluated using a single seed ($s = 0$). In R4, the selected models are retrained using multiple seeds $S = \{0, 7, 42, 123, 999\}$ to estimate stability.

$$M(i, d, s) \quad (1)$$

For R4, each metric is averaged across training seeds:

$$\widetilde{M}(i, d) = \frac{1}{|S|} \sum_{s \in S} M(i, d, s) \quad (2)$$

where $\widetilde{M}(i, d)$ represents the average performance across multiple training runs, S is the set of seeds employed, and $|S|$ its cardinality.

3.1 Worst-Case and Average Performance

Let $\mathcal{D}_{\text{target}}$ denote the set of UAV domains:

$$M_{\min}(i) = \min_{d \in \mathcal{D}_{\text{target}}} \widetilde{M}(i, d) \quad (3)$$

$$\overline{M}(i) = \frac{1}{|\mathcal{D}_{\text{target}}|} \sum_{d \in \mathcal{D}_{\text{target}}} \widetilde{M}(i, d) \quad (4)$$

where $M_{\min}$ represents worst-case performance and $\overline{M}$ the mean performance. Metrics are normalized within the range $[0, 1]$.

3.2 Safety Constraint (Activation/Gating)

A feasibility condition based on worst-case recall is imposed:

$$\text{Recall}_{\min}(i) \geq \gamma \cdot \text{Recall}_{\min}(P_{\text{Raw}_0}) \quad (5)$$

where:

$$\text{Recall}_{\min}(i) = \min_{d \in D_{\text{target}}} \text{Recall}(i, d) \quad (6)$$

With $\gamma = 1.0$, the gate requires worst-case recall at least equal to the unrefined UAQ-MSUAV baseline N11. This is a constraint over the evaluated domains, not a guarantee for unseen deployment conditions.

3.3 Robust Scoring Function

For each metric $M \in \mathcal{M} = \{\text{mAP}_{50:95}, \text{mAP}_{50}, \text{Precision}, \text{Recall}\}$, the robustness score is defined as:

$$S_M(i) = \alpha \cdot \bar{M}(i) + (1-\alpha) \cdot M_{\min}(i) \quad (7)$$

with $\alpha = 0.5$ fixed across all rounds, balancing average performance and robustness.

The parameter $\alpha \in [0, 1]$ controls the trade-off between average performance and worst-case robustness. This scoring function is applied only to configurations that satisfy the gating criterion. $\bar{M}(i)$ represents the average performance of metric M across all domains, while $M_{\min}(i)$ corresponds to the worst observed performance (lower robustness bound).

3.4 Global Score

The final global ranking score for each alternative i is computed as:

$$S_{\text{global}}(i) = \sum_{M \in \mathcal{M}} w_M S_M(i) \quad (8)$$

where $S_M(i)$ is the robustness score associated with metric M , defined in the previous stage, and w_M represents the relative weight assigned to each metric, satisfying the constraint $\sum_{M \in \mathcal{M}} w_M = 1$:

The set of considered metrics is:

$$\mathcal{M} = \{\text{mAP}_{50:95}, \text{mAP}_{50}, \text{Precision}, \text{Recall}\} \quad (9)$$

The following weight vector is employed:

$$w = (0.50, 0.20, 0.15, 0.15) \quad (10)$$

The weights assign 0.50 to mAP50–95, which evaluates localization across IoU thresholds, 0.20 to mAP50, and 0.15 each to precision and recall. Recall also has an independent feasibility constraint through gating. These fixed weights express the selection priorities used throughout the experiments.

4. Model Selection and Validation through Exploratory Stability Analysis (R0–R4)

During the exploratory rounds R0–R3, each configuration is evaluated using a single seed ($s = 0$) in order to isolate dataset effects and reduce computational cost. At this stage, the MCDM framework is applied directly to $M(i, d, s)$ with $s = 0$, assuming a point estimate of performance.

In contrast, R4 introduces confirmatory validation using multiple seeds $S = \{0, 7, 42, 123, 999\}$, aiming to capture the variability induced by stochastic training.

Additionally, statistical stability is estimated through:

- mean $\bar{x}$
- standard deviation (SD)
- coefficient of variation (CV)
- confidence interval (CI95)

$$SD = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (11)$$

$$CV = \left(\frac{SD}{\bar{x}} \right) \times 100 \quad (12)$$

$$CI95 = \bar{x} \pm t_{0.975, df} \cdot \frac{SD}{\sqrt{N}} \quad (13)$$

5. General Linear Model of Fixed Effects (ANOVA)

A fixed-effects general linear model (GLM/ANOVA) separates the observed contributions of scenario, evaluation domain, seed, and scenario–domain interaction. This tests whether configuration and domain differences exceed residual variation; it does not independently establish causality. The metric response is defined as follows: The fitted design contains 120 observations: four configurations $\times$ six datasets $\times$ five seeds, using sequential sums of squares.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + \varepsilon_{ijk} \quad (14)$$

where:

- μ is the global mean.
- α_i , β_j , and γ_k represent the main effects of scenario, domain, and seed, respectively.
- $(\alpha\beta)_{ij}$ models the interaction between scenario and domain.

- ε_{ijk} represents the random error term, assumed to have zero mean, constant variance, and independence across observations.

For each factor, the F-test compares its mean square with the residual mean square. Statistical significance is reported with p-values and effect magnitude with partial eta squared:

$$\eta_p^2 (\text{Seed}) = \frac{SS_{\text{Seed}}}{SS_{\text{Seed}} + SS_{\text{Residual}}} \quad (15)$$

5.1 Multiple Comparisons

After detecting significant effects through ANOVA ($p < 0.05$), the Tukey HSD post hoc test is applied to identify significant differences between model pairs:

$$\text{HSD} = q_{\alpha, k, df} \sqrt{\frac{MS_{\text{Residual}}}{n}} \quad (16)$$

Where $q_{\alpha, k, df}$ is the critical value of the studentized range distribution, k is the number of compared groups, and n is the number of observations per group (balanced design), enabling control of the family-wise error rate.

5.2 Assumption Validation (Shapiro–Wilk)

Residual normality is evaluated using the Shapiro–Wilk test:

$$W = \frac{(\sum_{i=1}^n a_i e_{(i)})^2}{\sum_{i=1}^n (e_i - \bar{e})^2} \quad (17)$$

where:

- e_i = i -th residual (unordered)
- $e_{(i)}$ = ordered residual (i -th smallest)
- a_i = Shapiro–Wilk coefficients
- $\bar{e}$ = mean residual
- n = total number of residuals

A non-significant Shapiro–Wilk test, considered together with Q–Q plots, indicates no detected departure from normality; it does not establish that the assumption holds.

Only five training seeds are available. Confidence intervals and assumption tests therefore require caution, and non-significant seed effects should be interpreted as limited evidence of initialization sensitivity within this experiment.

6. Proposed DCAI Process Flow

6.1 Deduplication through the Discrete Cosine Transform (DCT)

During ETL ingestion, pHash reduces repeated visual content while tolerating illumination and compression changes. Images are resized to 32×32 , converted to grayscale, and represented by a 64-bit signature derived from low-frequency DCT coefficients.

The 2D DCT of an image $I(x, y)$ of size $N \times N$ is defined as:

$$F(u, v) = \alpha(u)\alpha(v) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} I(x, y) \cos \left[\frac{\pi(2x+1)u}{2N} \right] \quad (18)$$

where $I(x, y)$ denotes the pixel intensity at spatial coordinates (x, y) , and $F(u, v)$ the coefficient in the frequency domain. The image has dimensions $N \times N$, with $x, y, u, v \in \{0, \dots, N-1\}$, where x, y are spatial indices and (u, v) frequency indices. The factors $\alpha(u)$ and $\alpha(v)$ correspond to orthonormal normalization, with $\alpha(0)$ associated with the DC coefficient:

$$\alpha(0) = \sqrt{\frac{1}{N}}, \alpha(k) = \sqrt{\frac{2}{N}}, k > 0 \quad (19)$$

The upper-left 8×8 DCT block is thresholded at its mean to produce the 64-bit signature, retaining global structure while suppressing high-frequency variation.

6.1.1 Deduplication Criterion

Image similarity is evaluated through the Hamming distance between pHash signatures:

$$d_H(h_i, h_j) \leq \tau \quad (20)$$

The sensitivity analysis compared Hamming thresholds of 1, 3, 5, and 10. We selected 1 because larger thresholds removed more than 10% of the custom dataset, risking loss of diversity. Figure 2 illustrates pHash, and Table 2 reports threshold-dependent retention.

Original pixels								DCT coefficients								Perceptual hash								
52	55	61	66	70	61	64	73	-415	-30	-61	27	56	-20	-2	0	0	0	0	1	1	0	0	1	1
63	59	55	90	109	85	69	72	2	-22	-61	10	13	-7	-9	5	1	0	0	1	1	0	0	1	1
62	59	68	113	144	104	66	73	-47	7	77	-25	-29	10	5	-6	0	1	1	0	0	1	1	0	0
63	58	71	122	154	106	70	69	-49	12	34	-15	-10	6	2	2	0	1	1	0	0	1	1	1	1
67	61	68	104	126	88	68	70	12	-7	-13	-4	-2	2	-3	-3	1	0	0	0	1	1	0	1	1
79	65	60	70	77	68	58	75	-8	3	2	-6	-2	1	4	2	0	1	1	0	0	1	1	1	1
85	71	64	59	55	61	65	83	-1	0	0	-1	-4	-3	1	2	1	1	1	0	1	0	1	1	1
87	79	69	68	65	76	78	94	-1	0	-1	-4	-1	0	-1	2	1	1	1	0	1	1	1	1	1

Figure 2. Visualization of the pHash/DCT deduplication Note. Source: Own elaboration

Table 2. 2Simulation table of Hamming thresholds.

C (τ / ID)	Dataset	Initial	Final	Ret. %	% Δ
1 / N4	HYBRID	15.2k	14.9k	97.7	2.2
1 / N12	UAQ—MSUAV	5.6k	5.3k	94.4	5.6
3 / N4	HYBRID	15.2k	14.1k	93	7
3 / N12	UAQ-MSUAV	5.6k	4.6k	81.5	18.4
5 / N4	HYBRID	15.2k	13.6k	89.7	10.2
5 / N12	UAQ-MSUAV	5.6k	4.1k	72.7	27.3
10 / N4	HYBRID	15.2k	12.6k	83.2	16.7
10 / N12	UAQ-MSUAV	5.6k	3.1k	56	44

Note. Ret. = retained images; Δ = change; k = thousand images. Thresholds and configuration identifiers are reported in the first column.

6.2 Semantic Auditing: The Auditor Model Approach

After deduplication, YOLO11x audits existing PASCAL VOC labels and UAQ-MSUAV labels generated by YOLO11m and checked in CVAT. The higher-capacity auditor supplies a semantic consistency check that perceptual hashing cannot provide.

COCO pretraining may bias the auditor against small objects and unusual aerial views. Ambiguous cases therefore require human validation; auditor disagreement is not sufficient evidence of annotation error.

6.2.1 Inference Strategy

Inference is performed using low confidence thresholds in order to prioritize recall. Let $u_i \in [0, 1]$ denote the auditor confidence associated with the prediction of instance i . A configurable threshold is defined as:

$$\beta_u \in (0.35, 0.65) \quad (21)$$

The filtering decision is based on:

$$\text{Accept instance: } u_i \geq \beta_u \quad (22)$$

Two operational regimes are considered:

- $\beta_u = 0.65$: strict threshold, prioritizes precision (reduces false positives).
- $\beta_u = 0.35$: relaxed threshold, prioritizes recall (higher tolerance to noise).

Both regimes are integrated into the Design of Experiments (DoE) framework to evaluate their impact on the precision–recall trade-off.

6.2.2 Geometric Consistency Verification

Annotations in YOLO format are transformed into absolute coordinates (x_1, y_1, x_2, y_2) . Consistency is evaluated through Intersection over Union (IoU), where A denotes the ground truth (GT) and B the prediction:

$$\text{IoU}(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B| + \epsilon}, \epsilon = 10^{-6} \quad (23)$$

An instance is considered misaligned if:

$$\max_{\text{GT}, \text{Pred}} (\text{IoU}) < \text{IoU}_{\text{TH}}, \text{IoU}_{\text{TH}} = 0.5 \quad (24)$$

6.2.3 Noise Taxonomy

Each sample is classified as:

- ok: semantic consistency
- missing label: detection without GT
- bad label: GT without detection
- poor alignment: insufficient IoU

Non-*ok* samples are removed. If an image remains without valid annotations, it is discarded to avoid gradients dominated by excessive noise.

6.2.4 Scope and Experimental Separation

The auditor is applied only to compatible domains (PASCAL, UAQ-MSUAV, hybrid), where the spatial distribution (5–25 m, 0°–60°) ensures higher filtering reliability.

External datasets (VisDrone, Manipal, INRIA, NTUT) are used exclusively for evaluation, without refinement, preserving a clean reference for measuring generalization under domain shift.

This restriction limits exposure to extreme viewpoints for which the auditor was not adapted, but does not eliminate domain bias. Figure 3 summarizes its filtering logic.

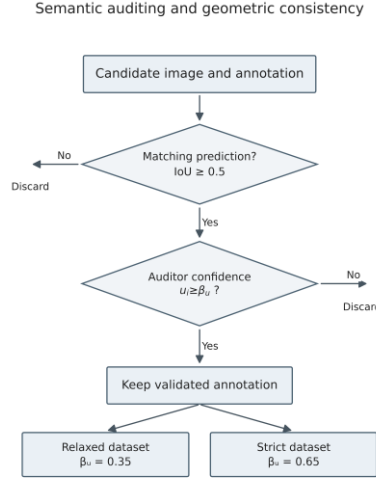


Figure 3. Auditor model logic. Source: Own elaboration

6.3 Hierarchical Ablation and Oversampling (OS)

Oversampling tests whether sample density lost during curation can be recovered by duplicating validated images. Applying it after pHash and auditing avoids intentionally replicating rejected samples, although duplication adds no new semantic information.

The ablations compare 0%, 15%, and 40% oversampling across training datasets and auditor regimes:

The experimental variables considered are:

- Quality, through pHash + auditor (Strict / Relaxed).
- Density, with OS $\in \{0, 15, 40\}\%$.
- Training dataset, including PASCAL, UAQ-MSUAV, and hybrid configurations.

Figure 4 summarizes the operational workflow of the proposed DCAI process for image cleaning.

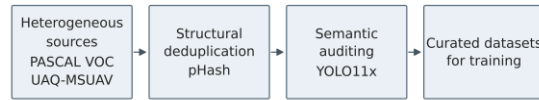


Figure 4. Proposed DCAI workflow. Source: Own elaboration.

7. Experimental Configuration

Source domains support training and ablation; target domains are reserved for evaluation (Tables 3 and 4). The split protocol is 80% training, 10% validation, and 10% testing.

Rounds R0–R3 use seed 0. R4 retrain selected configurations with seeds 0, 7, 42, 123, and 999 (Table 5).

Source Domains

Table 3. 3Image domains used for training.

Dataset	Description	Images
PASCAL VOC (person)	Ground-based scenario	9,583
UAQ-MSUAV	Multi-scale UAV scenario (5–25 m, 0°–60°)	5,783
Hybrid (PASCAL + UAQ)	Mixed distribution	15,246

Target Domains (Evaluation)

Table 4. 4Image domains used for evaluation.

Dataset	Scenario	Images
VisDrone2019	High altitude, high scale variation	1,267
NTUT 4K	High resolution, top-down view	382
Manipal UAV	Unstructured environments	49
UAQ_test	Multi-scale (5–25 m)	450
INRIA Person	Classical ground-based dataset	491

Randomness Control

Table 5. 5Random seeds employed per training round.

Round	Seeds
R0–R3	0
R4	0, 7, 42, 123, 999

7.1 Implementation Details

All experiments train YOLO11n by transfer learning, using an independent dataset YAML for each scenario and the fixed hyperparameters in Table 6.

YOLO11x is used without fine-tuning only for offline auditing. Keeping the detector architecture and training settings constant focuses the comparison on data configuration without adding auditor cost to deployed inference.

Constant Experimental Configuration

Table 6. 6Fixed hyperparameters used during training.

Category	Parameter	Value
Training	Initialization	YOLO11n pretrained weights
	Epochs	100 (early stopping = 15)
	Input	640×640
	Batch	8
Hardware	GPU / CPU	RTX 3060 (6GB), 4 workers
Optimizer	Algorithm	AdamW
	Determinism	Enabled
Optimization	lr_0 / lrf	0.01 / 0.01
	Momentum	0.937
	Weight decay	5×10^{-4}
	Warmup	3 epochs
Losses	Box / Cls / DFL	7.5 / 0.5 / 1.5
Augmentations	Mosaic	1.0 (close_mosaic = 10)
	MixUp	0
	Flip	0.5
	HSV	(0.015, 0.7, 0.4)
Evaluation	Validation	Enabled (split = val)
	Logging	Checkpoints + metric curves

Category	Parameter	Value
Camera (UAV/Mobile)	Resolution	4K (3840×2160)
	Shutter	1/1000s (Fixed Shutter)
	Sensitivity	ISO 100 (Fixed)
	Aperture	f/2.8 – f/4.0
	Focus	Infinity / Locked

8. Experimental Design (DoE through Ablation)

The hierarchical DoE varies pHash, semantic auditing, and oversampling with YOLO11n held fixed. Table 7 defines the scenarios and contrasts; the rounds progressively screen configurations before multi-seed confirmation in R4.

Round Description

- R0 (control): establishes lower bounds (zero-shot) and a transfer-learning baseline.
- R1 (quality): isolates the effects of pHash, the auditor, and their interaction.
- R2 (density): evaluates the impact of OS and its interaction with semantic filtering.
- R3 (generalization): validates the transferability of improvements within the UAQ domain.

Key Contrasts (Defined in Table 7)

- pHash: N4 vs. N3
- Auditor: N5 vs. N3
- Interaction: N6
- Threshold: N7 vs. N6

Main Scenario Table

Table 7. 7All controlled experimental scenarios separated by round.

Round	ID	Scenario	Domain	pHash	Auditor (β_u)	OS
R0	N1	Z_COCO	COCO	No	—	0%
R0	N2	B_Raw	PASCAL	No	—	0%
R1	N3	H_Raw_0	Hybrid	No	—	0%
R1	N4	H_pH_0	Hybrid	Yes	—	0%
R1	N5	H_JR_0	Hybrid	No	0.35	0%

Round	ID	Scenario	Domain	pHash	Auditor (β_u)	OS
R1	N6	H_pH_JR_0	Hybrid	Yes	0.35	0%
R1	N7	H_pH_JS_0	Hybrid	Yes	0.65	0%
R2	N8	H_pH_JR_15	Hybrid	Yes	0.35	15%
R2	N9	H_pH_JR_40	Hybrid	Yes	0.35	40%
R2	N10	H_pH_JS_40	Hybrid	Yes	0.65	40%
R3	N11	P_Raw_0	UAQ-MSUAV	No	—	0%
R3	N12	P_pH_JR_0	UAQ-MSUAV	Yes	0.35	0%
R3	N13	P_pH_JR_15	UAQ-MSUAV	Yes	0.35	15%
R3	N14	P_pH_JR_40	UAQ-MSUAV	Yes	0.35	40%

Note. OS = oversampling. The auditor threshold is the confidence threshold used for semantic filtering; an dash indicates that auditing was not applied.

8.1 Distance Degradation Experiment

An evaluation is conducted on the UAQ_test set, covering a distance range from 5 to 25 m (450 images in total, 90 per distance level), using models N1, N2, N11, and the winning model from round R4.

Inference is executed on a Raspberry Pi 5 (8 GB), employing model conversion from.pt format to NCNN and considering reduced-precision deployment scenarios (FP32 to FP16/INT8).

9. Results

9.1 pHash and Auditor Model Results

In the hybrid dataset, pHash retained 14,903 of 15,246 images, removing 343 duplicates (2.25%; Table 8). Redundancy mainly arose from consecutive frames in the proprietary UAV data.

Across hybrid configurations, retention ranged from 92.8% to 97.7%.

Strict YOLO11x auditing retained 14,206 validated samples, and subsequent oversampling increased dataset size to as much as 126.1% of the original. These counts describe changes in training volume; their detection effects are evaluated in Table 9.

Table 8. 8pHash deduplication results in hybrid scenarios.

Metric	Value
Total Processed Images	15,246
Unique Images (Hash)	14,903
Duplicate Images	343
Redundancy Rate	2.25%

9.2 Ablation Study Results (R0–R3)

Table 9 reports the R0–R3 ablations used to screen data configurations for R4.

Table 9. 9Ablation study results by round

Round	Evaluated Configuration	Δ Average (mAP / Prec / Rec)	Δ Worst Case	Key Finding	Conclusion
R0	Fine-tuning vs pretrained YOLO	↓ Prec (−2.67%), ↓ Rec (−5.63%)	↓ Prec (−15.53%), ↓ Rec (−5.36%)	Fine-tuning on PASCAL degrades generalization	Transfer learning between similar datasets does not guarantee improvement
R1	DCAI (pHash + auditor without OS)	≈ stable (<0.6% mAP, ~1.2% Prec/Rec)	↑ Prec (+5.95%), ↑ Rec (+1.36%)	$\beta_u = 0.65$ improves critical cases; $\beta_u = 0.35$ provides no benefit	Impact depends on auditor confidence
R2	DCAI + Oversampling (OS)	≈ stable (<0.2% mAP, <1.45% Prec)	Trade-off (Prec vs Rec ± 1 –2%)	OS improves specific scenarios, but not consistently	Marginal benefit, not essential
R3	UAV-only (UAQ-MSUAV) + DCAI	≈ stable in mAP	↓ Prec (up to −13.5%), ↑ Rec (~+0.3%)	Curation reduces false positives; OS partially compensates	Improves reliability at the cost of global precision

Note. Prec = precision; Rec = recall; OS = oversampling. Arrows indicate the direction of the reported change.

9.3 Model Selection (R0–R3)

The MCDM metrics, weights, and recall gate defined in Section 3 were applied to R0–R3.

Hybrid configurations lead the mAP50–95 and global rankings, while the UAQ-MSUAV-only model favors precision at lower recall. Configurations with and without oversampling perform similarly at this screening stage (Table 10).

The selected candidates proceed to R4 for multi-seed evaluation.

Table 10. 10Most robust models selected from R0–R3 through MCDM.

Scenario	Round	S mAP50–95	S mAP50	S Precision	S Recall	S global
H_pH_JS_0	R1	0.192	0.294	0.423	0.292	0.262
H_pH_JS_40	R2	0.193	0.292	0.417	0.289	0.261
P_Raw_0	R3	0.152	0.259	0.436	0.253	0.231
B_Raw_0	R0	0.155	0.254	0.380	0.252	0.223

Note. S denotes the robustness score for each metric; the global score combines the specified metric weights.

9.4 R4 — Robust Validation (MCDM + Multi-Seed)

R4 retraines the selected candidates with the five seeds listed in Table 5 and applies the unchanged MCDM rule from Section 3.

Table 11 and Figure 5 summarize R4 selection. Hybrid models with pHash and auditing lead the robust ranking.

N10 (hybrid, pHash, auditor, 40% OS) has the highest global score but performs similarly to N7 without OS. The small difference provides limited evidence of an additional oversampling benefit.

N11 has the highest worst-case precision, but N7 has the highest mean precision across the six datasets. N2 fails the baseline-referenced recall gate and is excluded from ranking.

Table 11. 11R4 MCDM results for selecting the most robust model.

Model	Configuration	Score_global	mAP50–95	Precision	Recall	Observation
N10	Hybrid + pHash + auditor + OS 40%	0.2626	↑	↑	↑	Best overall performance
N7	Hybrid + pHash + auditor (without OS)	0.2614	↑	↑	↑	Practical tie with N10
N11	UAQ-MSUAV (raw)	0.2296	↓	↑	↓	Highest worst-case precision; lower recall

Mode 1	Configuration	Score_globa l	mAP50– 95	Precision	Recall	Observation
N2	PASCAL VOC	—	—	—	—	Does not pass gating

Note. N2 is excluded because its worst-case recall falls below the N11-based gating threshold.

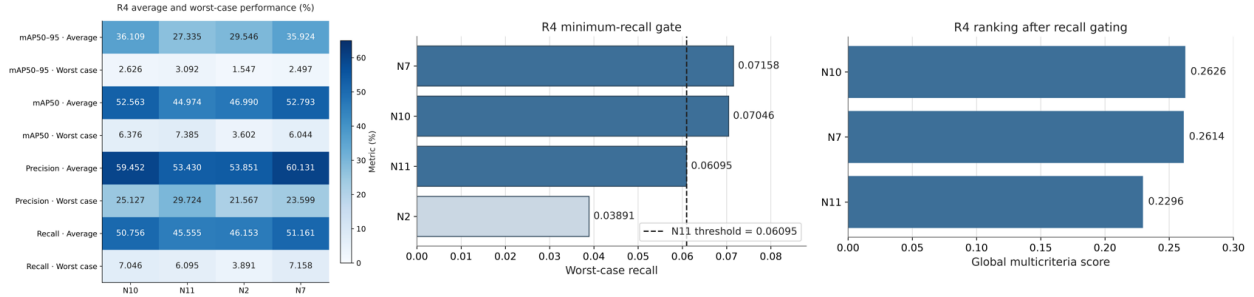


Figure 5. Final model selection and robustness validation (R4): (a) metric summary; (b) recall gating; (c) ranking of eligible models. Source: Own elaboration

9.5 DoE Results through GLM/ANOVA

Table 12 reports significant dataset, scenario, and scenario $\times$ dataset effects. Dataset has the largest effect sizes, while the interaction indicates that configuration performance depends on evaluation domain. The seed effect is not significant, providing no detected evidence of systematic seed differences within the tested runs.

Table 12. 12Multi-seed ANOVA results for R4.

Metric	Scenario F(3,92)	Dataset F(5,92)	Seed F(4,92)	Scenario $\times$ Dataset F(15,92)
mAP50	832.14*** ($\eta_p^2 = 0.964$)	49021.11*** ($\eta_p^2 = 0.999$)	0.66 ($\eta_p^2 = 0.028$)	573.40*** ($\eta_p^2 = 0.989$)
mAP50–95	2231.03*** ($\eta_p^2 = 0.986$)	67136.28*** ($\eta_p^2 = 0.999$)	0.48 ($\eta_p^2 = 0.020$)	2479.88*** ($\eta_p^2 = 0.998$)
Precision	22.74*** ($\eta_p^2 = 0.426$)	787.91*** ($\eta_p^2 = 0.977$)	1.36 ($\eta_p^2 = 0.056$)	14.99*** ($\eta_p^2 = 0.710$)
Recall	73.17*** ($\eta_p^2 = 0.705$)	7246.40*** ($\eta_p^2 = 0.997$)	0.46 ($\eta_p^2 = 0.020$)	50.38*** ($\eta_p^2 = 0.891$)

- *** $p < 0.001$
- η_p^2 = partial eta squared (effect size)

Table 13 summarizes dispersion across configurations and domains.

Table 13. 13Multi-seed descriptive statistics for R4 models.

Scenario	Model	mAP50–95 ($\mu \pm \text{SD}$)	CV (%)	CI95	Precision ($\mu \pm \text{SD}$)	Recall ($\mu \pm \text{SD}$)
H_pH_JS_0	N7	0.359 $\pm$ 0.313	87.1	[0.242, 0.476]	0.601 $\pm$ 0.244	0.512 $\pm$ 0.338
H_pH_JS_40	N10	0.361 $\pm$ 0.315	87.2	[0.243, 0.479]	0.595 $\pm$ 0.250	0.508 $\pm$ 0.342
B_Raw_0	N2	0.295 $\pm$ 0.275	93	[0.193, 0.398]	0.539 $\pm$ 0.262	0.462 $\pm$ 0.353
P_Raw_0	N11	0.273 $\pm$ 0.269	98.5	[0.173, 0.374]	0.535 $\pm$ 0.231	0.456 $\pm$ 0.323

Note. Values are reported as mean $\pm$ standard deviation where indicated. CV = coefficient of variation; CI95 = 95% confidence interval. Aggregated variability includes differences across evaluation domains.

Interpretation of Variability (CV)

Aggregated CV values of approximately 87–98% reflect substantial domain heterogeneity. Within-dataset seed variability is not uniformly below 10%: maximum CV values are 10.2% for mAP50–95, 8.1% for mAP50, 60.4% for precision, and 28.7% for recall. Non-significant seed main effects therefore do not imply uniformly low variability in every condition. Figure 6 presents mAP50–95 effect sizes; Table 12 contains all metrics.

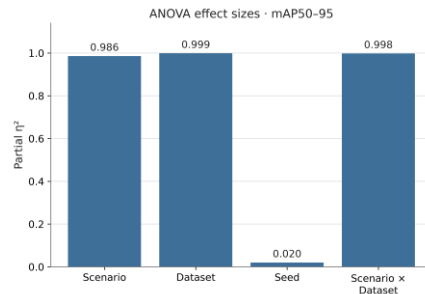


Figure 6. ANOVA effect sizes for mAP50–95 across scenario, dataset, seed, and their interaction. η_p^2 Source: Own elaboration

9.6 Multiple Comparisons (Tukey HSD)

The reported Tukey HSD analysis identifies differences favoring N7 and N10 over the unrefined baselines, particularly for mAP and recall. It detects no significant difference between N7 and N10, consistent with the limited oversampling benefit. Non-significance does not establish equivalence.

Table 14. Multi-seed average metric results for the evaluated models.

Group	mAP50	mAP50–95	Precision	Recall
N7 (H_JS_0)	0.528 $\pm$ 0.003	0.359 $\pm$ 0.001	0.601 $\pm$ 0.007	0.512 $\pm$ 0.006

Group	mAP50	mAP50-95	Precision	Recall
N10 (H_JS_40)	0.526 ± 0.001	0.361 ± 0.001	0.595 ± 0.005	0.508 ± 0.004
N2 (B_Raw)	0.470 ± 0.002	0.295 ± 0.002	0.539 ± 0.009	0.462 ± 0.004
N11 (P_Raw)	0.450 ± 0.002	0.273 ± 0.001	0.535 ± 0.033	0.456 ± 0.009

Note. Values are reported as mean $\pm$ standard deviation across seed-level mean metrics.

Figure 7 shows the distributions of seed-level mean metrics, and Figure 8 presents the corresponding Tukey HSD pairwise comparisons.

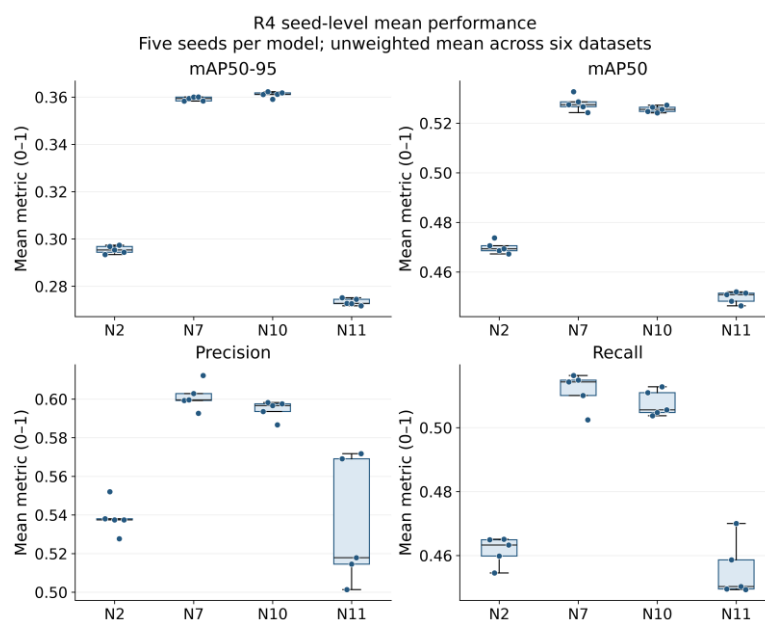


Figure 7. Seed-level mean performance across six evaluation datasets (five seeds per model). Points show all observations; boxes summarize their distribution. Source: Own elaboration

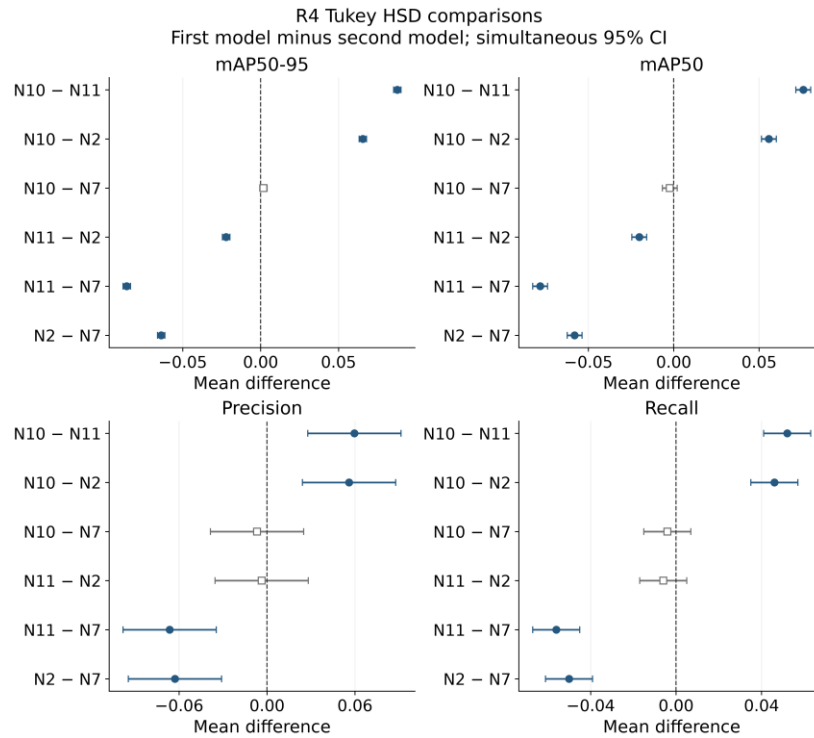


Figure 8. Tukey HSD differences between seed-level model means with simultaneous 95% confidence intervals. Differences are first model minus second model; open squares indicate non-significant comparisons. Source: Own elaboration

9.7 Assumption Validation through Shapiro–Wilk

Table 15 and Figure 9 concern the 20 residuals obtained by centering each seed-level multidomain mean on its model mean. Normality is not rejected for these residuals. A separate check of the 120-observation factorial GLM gives $p = 0.128$ for mAP50–95 and 0.136 for mAP50, but rejects normality for precision ($p = 7.51 \times 10^{-14}$) and recall ($p = 1.27 \times 10^{-5}$). Classical GLM inference for the latter metrics requires caution and robust sensitivity analysis.

Figure 9 shows the seed-mean comparison diagnostics corresponding to Table 15.

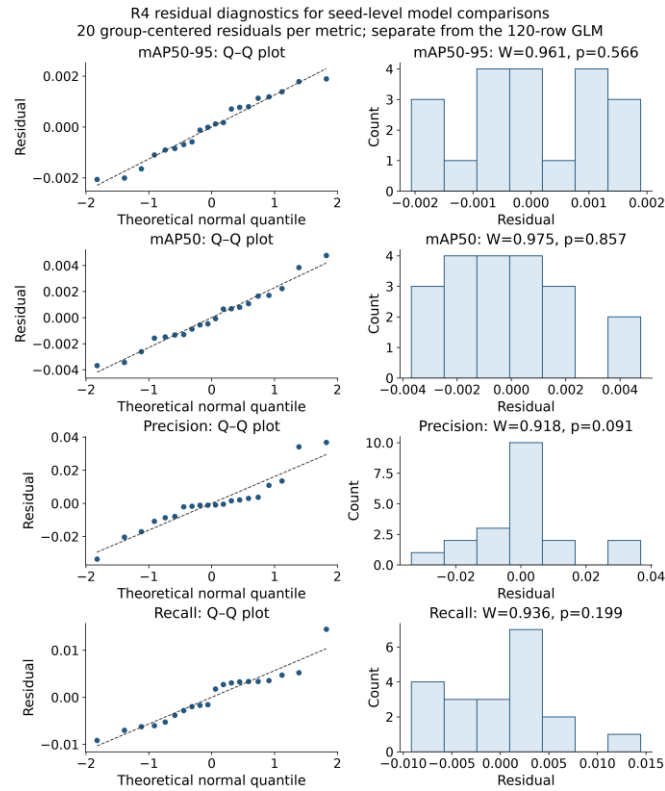


Figure 9. Q–Q plots and histograms of 20 group-centered residuals from seed-level mean comparisons. These are not the residuals of the 120-observation factorial GLM. Source: Own elaboration.

Table 15. Shapiro–Wilk results for 20 group-centered seed-mean residuals per metric.

Test	n_residuals	Statistic	p_value	Alpha	Verdict
Shapiro-Wilk – Recall	20	0.936	0.199	0.05	Not rejected
Shapiro-Wilk – Precision	20	0.918	0.091	0.05	Not rejected
Shapiro-Wilk – mAP50	20	0.975	0.857	0.05	Not rejected
Shapiro-Wilk – mAP50–95	20	0.961	0.566	0.05	Not rejected

9.8 Distance Degradation Results (Raspberry Pi)

Tables 16 and 17 and Figure 10 report performance over 5–25 m on Raspberry Pi 5. mAP50–95 decreases from approximately 0.80 to 0.32 for YOLO11 and from 0.62 to 0.23 for PASCAL, compared with 0.89 to 0.51 for N10.

N10 retains mAP50 of at least 0.91, precision of 0.939–0.968, and recall of 0.939–0.978 across the tested distances. These results complement its multi-domain selection by demonstrating on-device performance under scale variation.

N11 reaches 0.638 mAP50–95 at 25 m, compared with 0.508 for N10 (Table 16), consistent with the value of UAV-domain specialization. This distance experiment and the multi-domain ranking address different conditions; the ranking does not imply that N10 outperforms N11 at every distance.

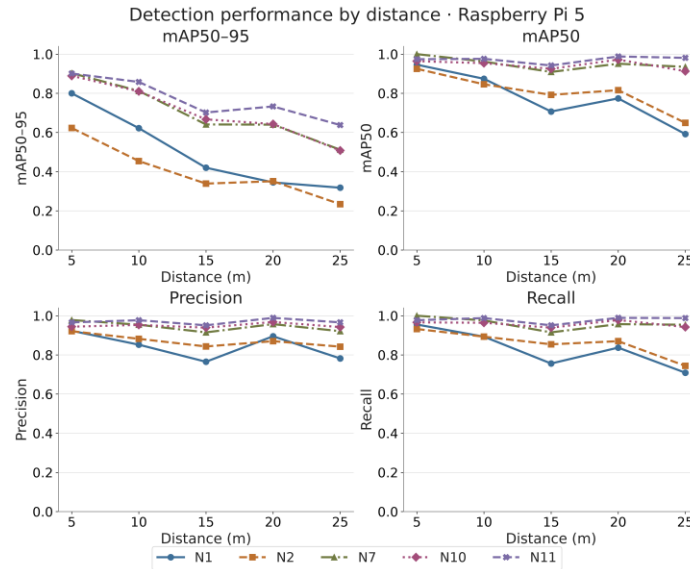


Figure 10. Performance degradation as a function of distance in edge deployments (Raspberry Pi 5). Values are reproduced from Tables 16 and 17. Source: Own elaboration

mAP50–95 and Precision

Table 16. Distance degradation results on Raspberry Pi 5: mAP50–95 and precision.

Distance (m)	mAP50–95				
	N1	N2	N7	N10	N11
5	0.800	0.623	0.903	0.889	0.901
10	0.622	0.454	0.812	0.810	0.858
15	0.420	0.339	0.641	0.667	0.702
20	0.345	0.351	0.640	0.643	0.733
25	0.318	0.234	0.512	0.508	0.638
Distance (m)	Precision				

Distance (m)	mAP50-95				
	N1	N2	N7	N10	N11
5	0.923	0.921	0.978	0.944	0.966
10	0.852	0.882	0.954	0.953	0.977
15	0.765	0.843	0.915	0.939	0.951
20	0.895	0.870	0.957	0.968	0.989
25	0.782	0.842	0.921	0.942	0.966

mAP50 and Recall

Table 17. Distance degradation results on Raspberry Pi 5: mAP50 and recall.

Distance (m)	mAP50				
	N1	N2	N7	N10	N11
5	0.946	0.925	1.000	0.964	0.974
10	0.874	0.846	0.962	0.954	0.976
15	0.707	0.792	0.909	0.925	0.942
20	0.774	0.816	0.951	0.972	0.988
25	0.592	0.649	0.935	0.913	0.981
Distance (m)	Recall				
	N1	N2	N7	N10	N11
5	0.955	0.932	1.000	0.966	0.977
10	0.893	0.893	0.976	0.964	0.988
15	0.756	0.854	0.915	0.939	0.951
20	0.837	0.870	0.957	0.978	0.989

Distan ce (m)	mAP50				
25	0.709	0.744	0.954	0.942	0.988

Table 18. Inference performance results on Raspberry Pi 5.

	N1	N2	N7	N10	N11
Latency (ms)	42.730	36.633	36.462	37.859	37.750
FPS (Proc)	6.475	6.769	6.796	6.861	6.849
Detection Conf.	0.855	0.726	0.897	0.897	0.921
Average inference. (ms)	121.675	107.032	108.172	109.474	108.466

Note. FPS = frames per second; Proc = processing; Conf. = confidence. Latency and mean inference time are retained as distinct measurements, as reported in the source manuscript.

Table 18 reports on-device throughput and timing. N10 achieves approximately 6.86 FPS with 109.47 ms mean inference time; inference time and pipeline throughput are distinct measurements.

10. Conclusions

The R0–R4 ablations support data curation as a means of improving worst-case UAV human detection with a fixed YOLO11n architecture. Dataset and scenario effects dominate the fitted analysis, while no significant seed effect is detected within the five evaluated initializations. These findings are conditional on the tested domains and experimental design.

N10 leads the robust ranking, but its similarity to N7 provides limited evidence that oversampling adds value after deduplication and semantic auditing. This finding does not establish a general saturation law.

On Raspberry Pi 5 with NCNN, N10 sustains precision of 0.939–0.968 and recall of 0.939–0.978 over 5–25 m at approximately 6.86 FPS. UAV-specific N11 remains competitive in the distance test, distinguishing domain specialization from multi-domain selection.

The main limitations are the small number of seeds, the auditor’s domain dependence, and non-normal factorial-GLM residuals for precision and recall. Robust inferential sensitivity checks and additional domains are needed before extending the conclusions beyond this experiment.

Acknowledgements

To the Universidad Autónoma de Querétaro for the technical guidance and the use of its facilities and laboratories.

Data availability:

<https://www.kaggle.com/datasets/luisngeld/person-detection-uav-pascal-voc-uaq-msuav>

Code:

https://github.com/LuisAngelOlveraOlvera/UAQ_MSUAV_DOE_ABLATION

References

- Akshatha, K. R., & others. (2023). Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194, 328–347. <https://doi.org/10.1016/j.isprsjprs.2022.10.015>
- Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). Robust optimization. Princeton University Press. <https://www2.isye.gatech.edu/~nemirovski/FullBookDec11.pdf>
- Bhusnoor, M., Patel, J., Mehta, A., Patel, D., Sainkar, S., & Mehendale, N. (2024). Investigating the impact of distance on object detection accuracy in unmanned aerial vehicle systems using MobileNetV3. TechRxiv. <https://doi.org/10.36227/techrxiv.24155598.v1>
- Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv. <https://arxiv.org/abs/2004.10934>
- Cheng, G., Ding, Q., Cai, B., Dang, C., Wang, Y., Zuo, X., & Shao, Z. (2024). CSSDet: Small object detection via cross-scale feature enhancement on drone-view images. *International Journal of Digital Earth*, 17(1), Article 2414848. <https://doi.org/10.1080/17538947.2024.2414848>
- Feurer, M., & Hutter, F. (2020). Auto-Sklearn 2.0: Hands-free AutoML via meta-learning. arXiv. <https://arxiv.org/abs/2007.04074>
- Fontana, R. (2024). Design of experiments and machine learning with application to industrial experiments. *Computational Statistics*, 39(5), 1–25. <https://doi.org/10.1007/s00362-023-01437-w>
- Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. International Conference on Machine Learning. <https://arxiv.org/abs/2007.01434>
- Hao, Y., Pei, H., Lyu, Y., Yuan, Z., Rizzo, J.-R., Wang, Y., & Fang, Y. (2022). Understanding the impact of image quality and distance of objects to object detection performance. arXiv. <https://arxiv.org/abs/2209.08237>
- Hashimoto, T. B., Sorensen, M. J., Raghunathan, R., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. arXiv. <https://arxiv.org/abs/1806.08010>
- Hegiste, V. (2024). Enhancing object detection with hybrid dataset in manufacturing environments: Comparing federated learning to conventional techniques. arXiv. <https://arxiv.org/abs/2408.08974>
- Jegham, N., Koh, C. Y., Abdelatti, M., & Hendawi, A. (2024). Evaluating the evolution of YOLO (You Only Look Once) models: A comprehensive benchmark study of YOLO11 and its predecessors. arXiv. <https://arxiv.org/abs/2411.00201>
- Karczmarczyk, A., & others. (2025). New robust multi-criteria decision-making method for wind farm location problems. *Applied Energy*, 357, Article 123456. <https://doi.org/10.1016/j.apenergy.2025.03.1316>
- Lai, K.-T., Chen, W.-Y., & others. (2024). AI wings: An AIoT drone system for commanding ArduPilot UAVs. <https://openreview.net/pdf?id=597f4106075460becdbf0157281cb213fe2f622c>
- Li, L., Lu, J., Zhang, S., Mohaisen, L., & Emam, M. (2023). Frame duplication forgery detection in surveillance video sequences using textural features. *Electronics*, 12(22), Article 4597. <https://doi.org/10.3390/electronics12224597>
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. arXiv. <https://arxiv.org/abs/1405.0312>
- Northcutt, C. G., Athalye, A., & Mackelprang, J. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70, 1–37. <https://arxiv.org/abs/1911.00068>
- OALA. (2023). DMLR: Data-centric machine learning research. arXiv. <https://arxiv.org/abs/2311.13028>
- Rey, L., Bernardos, A. M., Dobrzycki, A. D., Carramiñana, D., Bergesio, L., Besada, J. A., & Casar, J. R. (2025). A performance analysis of You Only Look Once models for deployment on constrained computational edge devices in drone applications. *Electronics*, 14(3), Article 638. <https://doi.org/10.3390/electronics14030638>
- Sheikholeslami, S., Meister, M., Wang, T., Payberah, A. H., Vlassov, V., & Dowling, J. (2021). AutoAblation: Automated parallel ablation studies for deep learning. In Proceedings of the 1st Workshop on Machine

- Learning and Systems (pp. 55–61). Association for Computing Machinery. <https://doi.org/10.1145/3437984.3458834>
- Solawetz, J. (2024). YOLOv11: An overview of the key architectural enhancements. arXiv. <https://arxiv.org/abs/2410.17725>
- Taiana, M., Nascimento, J. C., & Bernardino, A. (2013). An improved labelling for the INRIA person dataset for pedestrian detection. Iberian Conference on Pattern Recognition and Image Analysis. https://isr.tecnico.ulisboa.pt/wp-content/uploads/2015/05/2999_2013_IbPRIA-Taiana-Nascimento-Bernardino.pdf
- Tan, L., Zhang, C., Bai, H., Liu, Z., & Li, Y. (2025). A real-time and efficient detector for small object in UAV aerial images. *Scientific Reports*, 15, Article 39233. <https://doi.org/10.1038/s41598-025-22855-w>
- Tanaka, D., Ikami, D., Yagi, T., & Satoh, Y. (2018). Joint optimization framework for learning with noisy labels. arXiv. <https://arxiv.org/abs/1803.11364>
- Triantaphyllou, E. (2000). Multi-criteria decision making: An operations research approach. In Multi-criteria decision making methods (pp. 1–28). Springer. https://bit.csc.lsu.edu/trianta/EditedBook_CHAPTERS/EEEE1.pdf
- Zauner, C. (2010). Implementation and benchmarking of perceptual image hash functions [Bachelor's thesis, University of Applied Sciences]. https://www.phash.org/docs/pubs/thesis_zauner.pdf
- Zha, D. (2023). Data-centric AI: Perspectives and challenges. arXiv. <https://arxiv.org/abs/2311.05733>
- Zhang, G., Peng, Y., & Li, J. (2025). YOLO-MARS: An enhanced YOLOv8n for small object detection in UAV aerial imagery. *Sensors*, 25(8), Article 2534. <https://doi.org/10.3390/s25082534>
- Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7380–7399. <https://doi.org/10.1109/TPAMI.2021.3119563>



Esta obra está bajo una licencia de Creative Commons
Atribución-NoComercial-CompartirIgual 4.0 Internacional.