Evaluation of YOLOv12 for Multi-Modal Real-Time Visual UAV Detection

Julian Münzer

Motivation

Defence & Aerospace Safety: Lots of recent research

Interesting Challenges: Small & Fast-Moving, Environmental

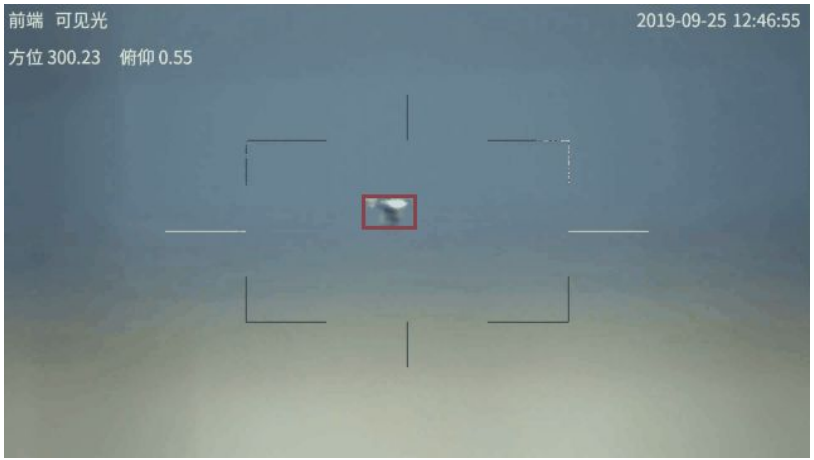
Benchmarkable problem: Regular Object Detection Evaluation Metrics

Approach: Run Object Detection using **Yolov12**

Dataset: AntiUAV-300

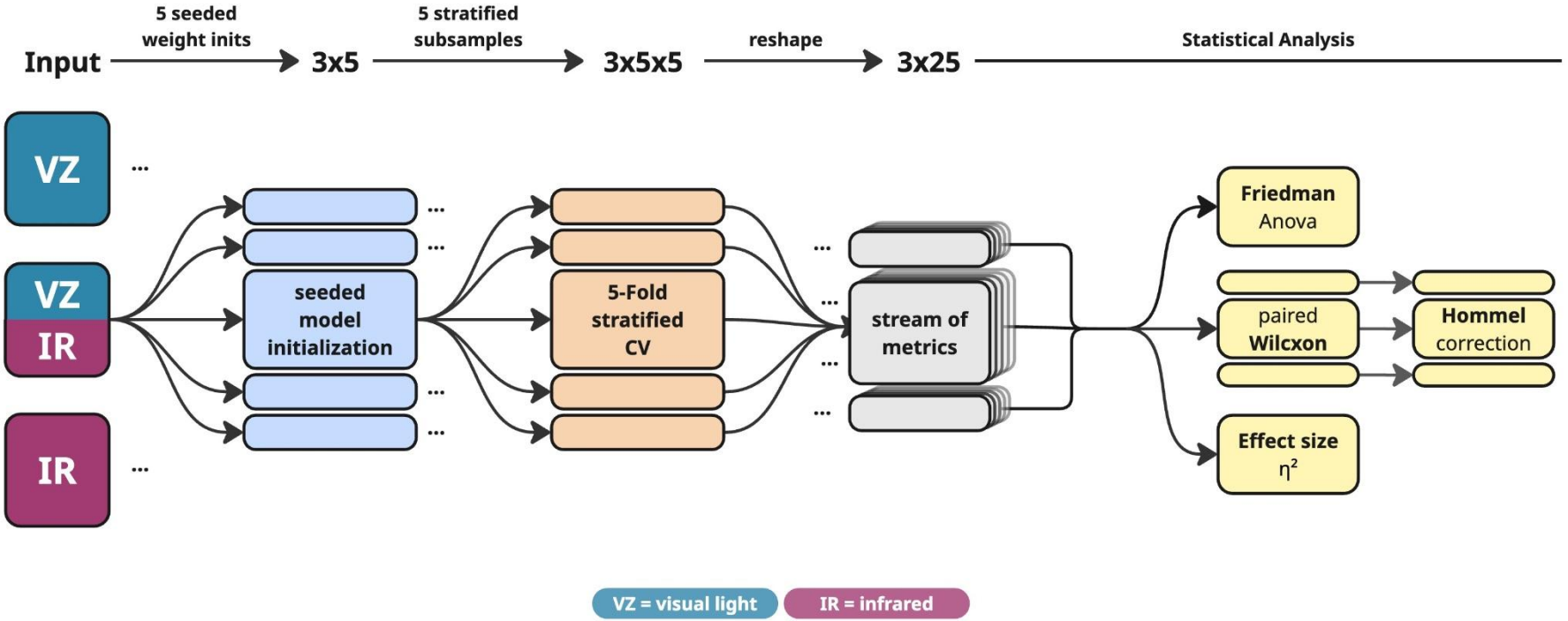
Lu, F., Zeng, C., Shi, H., Xu, Y., & Fu, S. (2025). Real-time detection sensor for unmanned aerial vehicle using an improved YOLOv8s algorithm. Sensors, 25(19), 6246. https://doi.org/10.3390/s25196246
Svanström, F., Alonso-Fernandez, F., & Englund, C. (2021). A dataset for multi-sensor drone detection. *Data in Brief, 39,* 107521. https://doi.org/10.1016/j.dib.2021.107521
Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-Centric Real-Time Object Detectors. arXiv preprint arXiv:2502.12524.
Coluccia, A., Fascista, A., Schumann, A., Sommer, L., Dimou, A., Zarpalas, D., Méndez, M., de la Iglesia, D., González, I., Mercier, J. P., Gagné, G., Mitra, A., & Rajashekar, S. (2021). Drone vs. Bird Detection: Deep Learning Algorithms and Results from a Grand Challenge. *Sensors (Basel, Switzerland)*, *21*(8), 2824. https://doi.org/10.3390/s21082824

Dataset

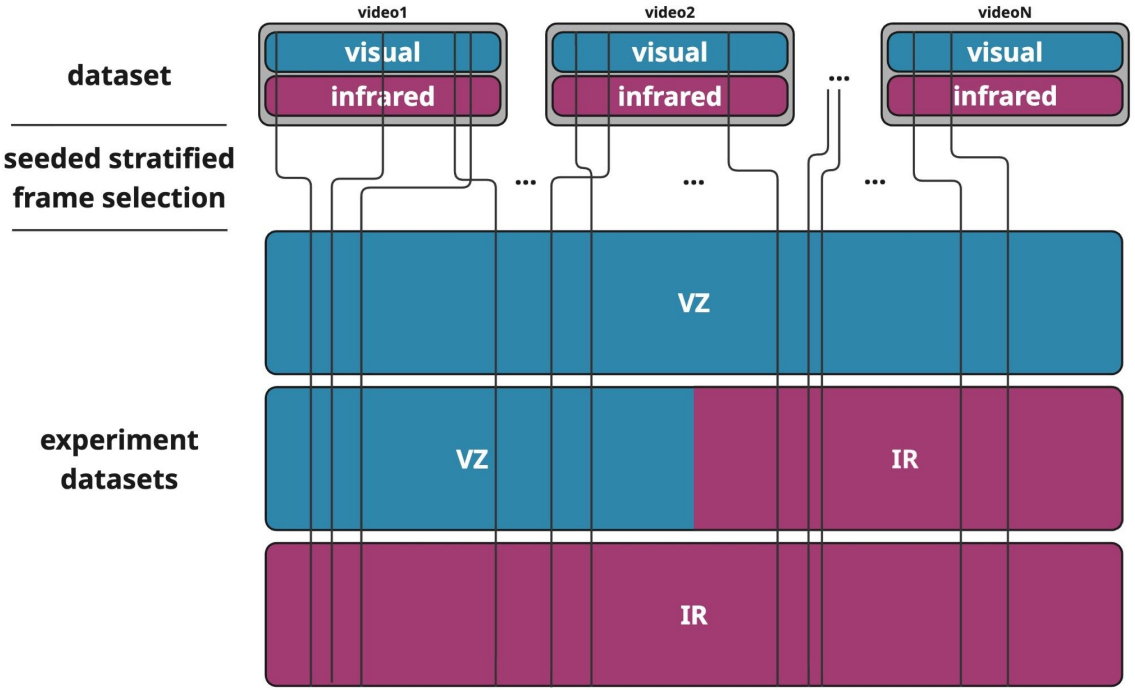


Svanström, F., Alonso-Fernandez, F., & Englund, C. (2021). A dataset for multi-sensor drone detection. *Data in Brief, 39,* 107521. https://doi.org/10.1016/j.dib.2021.107521

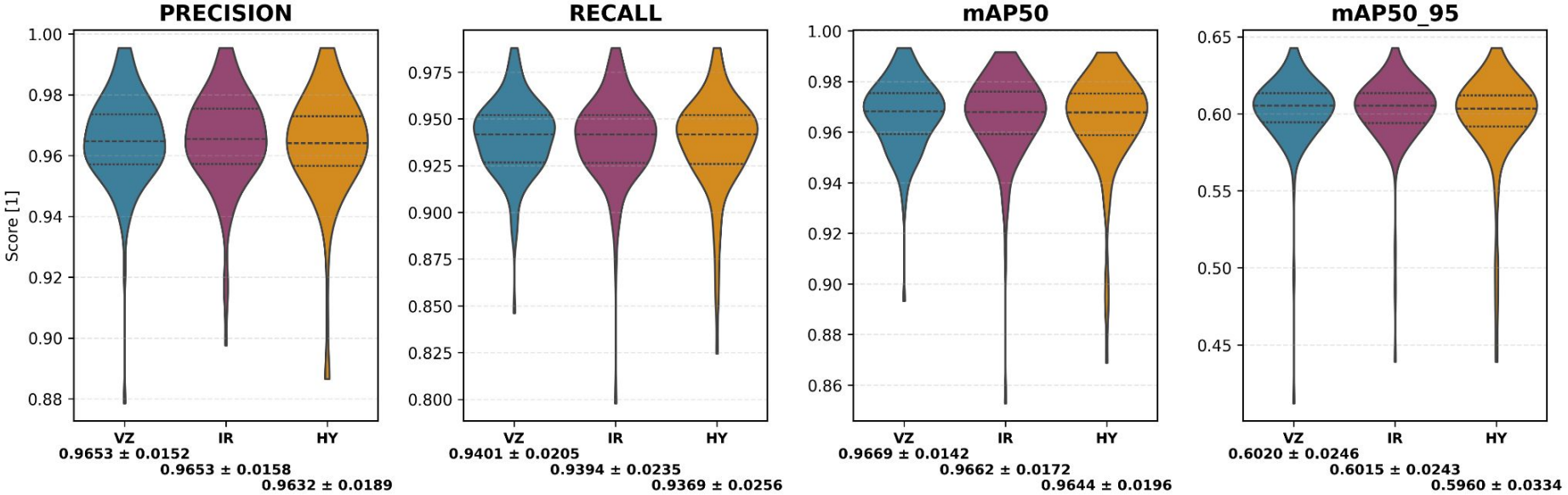
Experimental Plan

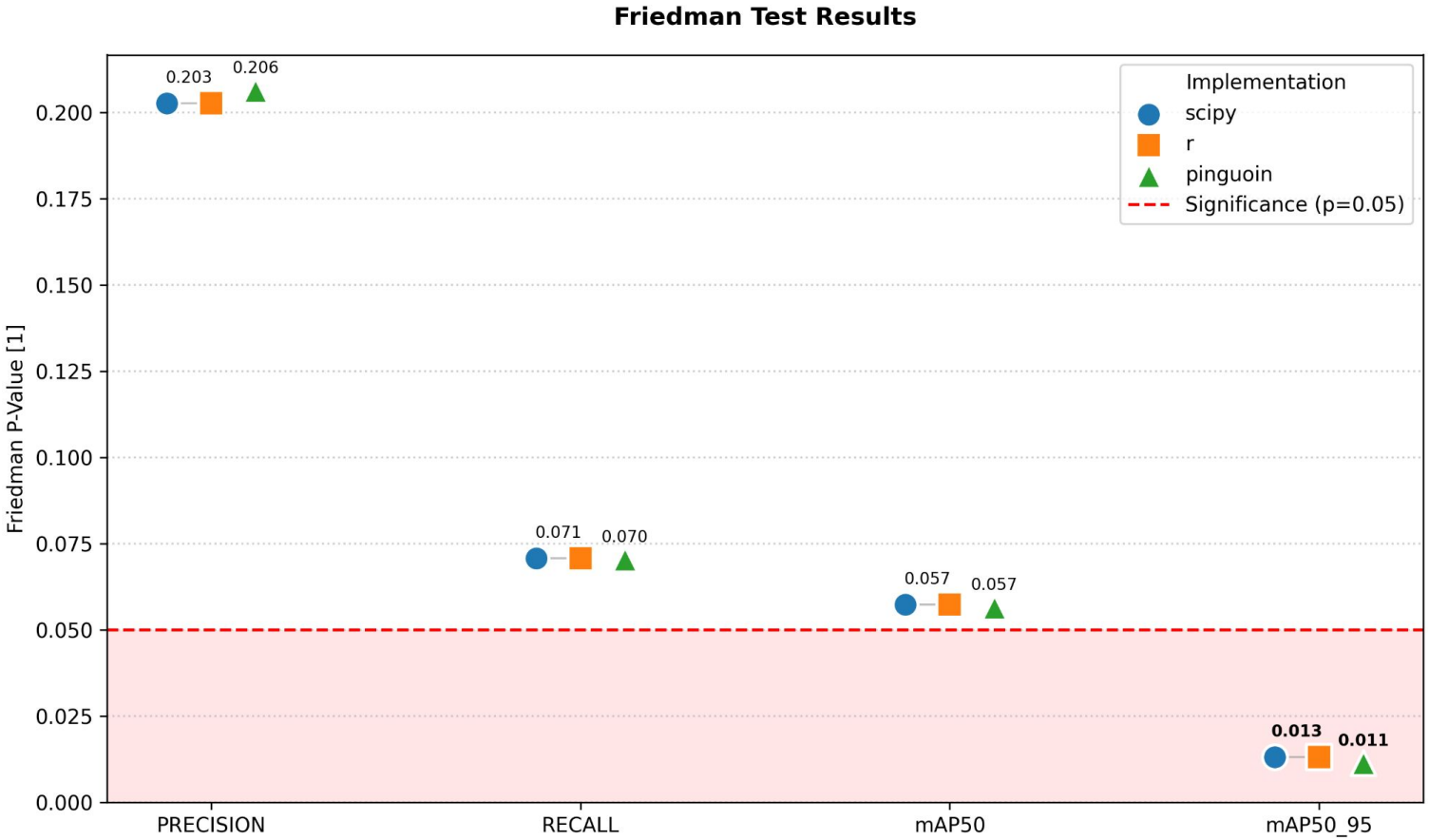


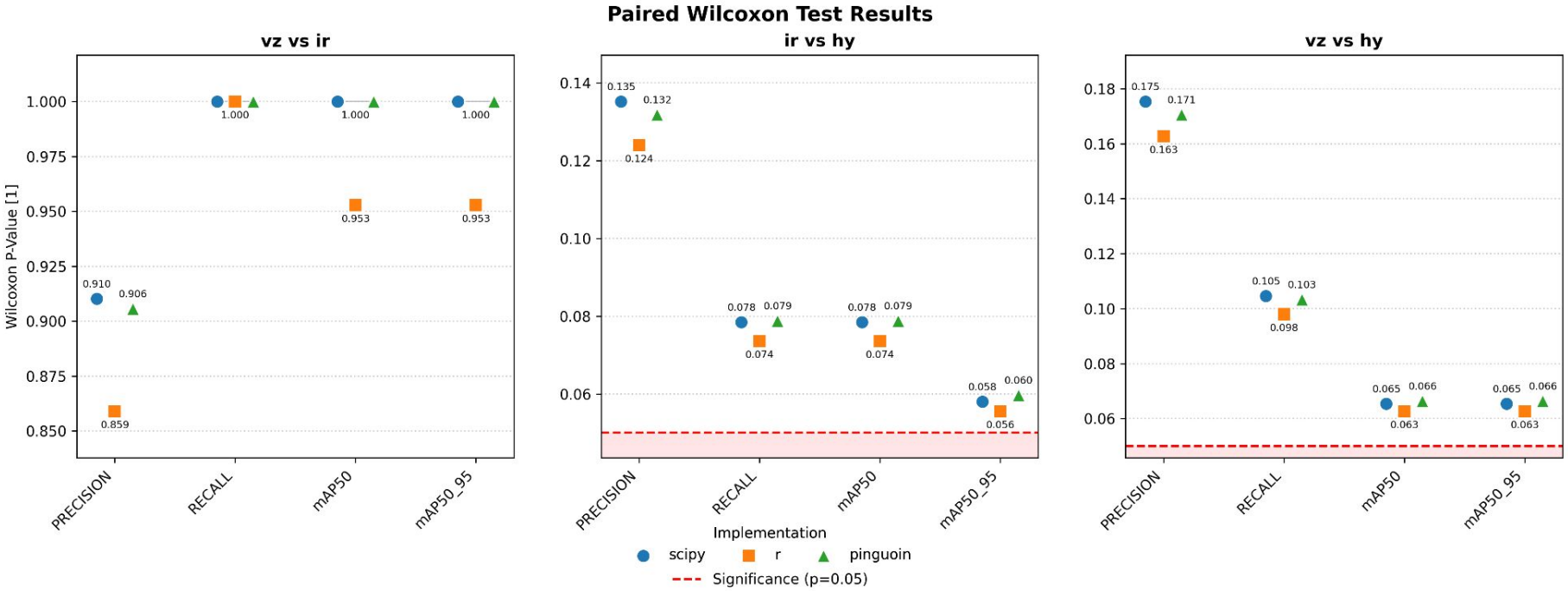
Data Sampling



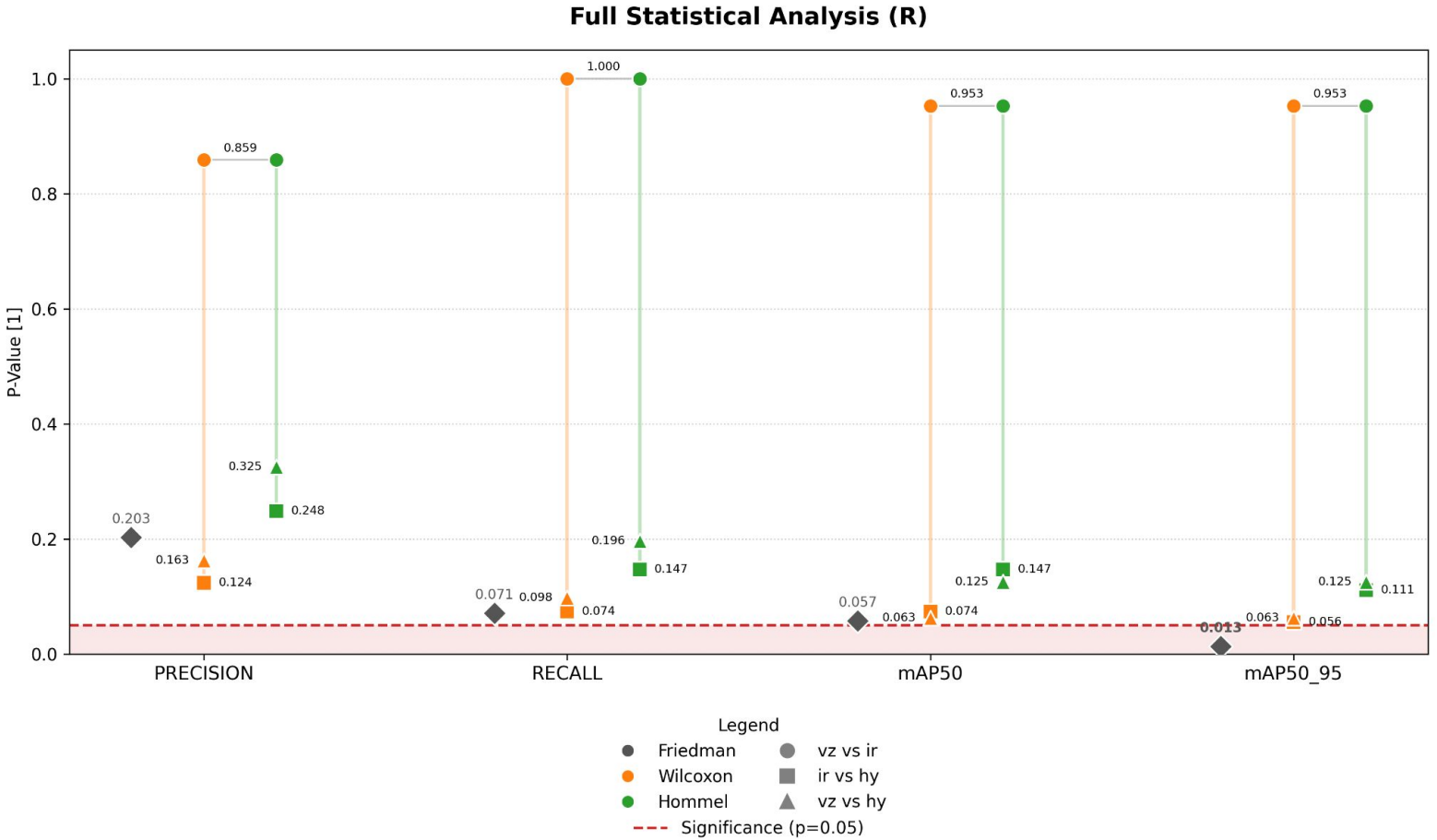
Performance Distribution




Friedman Test Results
Friedman P-Value [1]
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.203
0.206
0.071
0.070
0.057
0.057
0.013
0.011
PRECISION
RECALL
mAP50
mAP50_95
Implementation
scipy
r
pinguoin
Significance (p=0.05)



Precision measures how many of the predictions were correct, while recall measures how many of the actual objects were found.


Full Statistical Analysis (R)
P-Value [1]
1.0
0.8
0.6
0.4
0.2
0.0
PRECISION
RECALL
mAP50
mAP50_95
0.859
0.325
0.248
0.203
0.163
0.124
1.000
0.196
0.147
0.071
0.098
0.074
0.953
0.147
0.125
0.057
0.063
0.074
0.953
0.125
0.111
0.063
0.056
0.013
Legend
Friedman
Wilcoxon
Hommel
vz vs ir
ir vs hy
vz vs hy
Significance (p=0.05)

Effect Size

Metric	η^2	Interpretation
Precision	0.064	Medium
Recall	0.106	Medium
mAP50	0.114	Medium
mAP50-95	0.173	Large

Table 6: Effect Sizes (η^2) by Metric

0.01 small, 0.06 medium, 0.14 large

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates.

Backup Slides

1.2.1 Anti-UAV300 Dataset

The Anti-UAV300 dataset [4], contains 318 video sequences with two synchronized streams: Visible and Infrared. The dataset comprises 593,802 combined frames featuring real-valued pixel intensities, designed specifically for UAV detection and tracking research.

Spectrum	Total Frames	Annotated Frames	Coverage
Visible	296,901	280,218	94.38%
Infrared	296,901	293,209	98.76%
Total	**593,802**	**573,427**	**96.57%**

Table 1: AntiUAV-300 Dataset composition and annotation rates per modality [4]

2.1 Performance Measures

Precision (P) and Recall (R) measure the trade-off between false positives and false negatives:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN} \tag{1}$$

Mean Average Precision (mAP) is defined as the mean of the Average Precision across all classes and/or IoU thresholds. AP is calculated as the area under the Precision-Recall curve:

$$AP = \int_0^1 p(r)\, dr \tag{2}$$

where $p(r)$ is the precision at recall r. We report mAP50 (IoU threshold of 0.50) and mAP50-95 (averaged across IoU thresholds from 0.50 to 0.95 in steps of 0.05), with the latter providing a more strict measure of localization accuracy.

2.3 Effect Size: Eta Squared (η^2)

We utilize η^2 to measure the proportion of variance explained by the modality factor::

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{7}$$

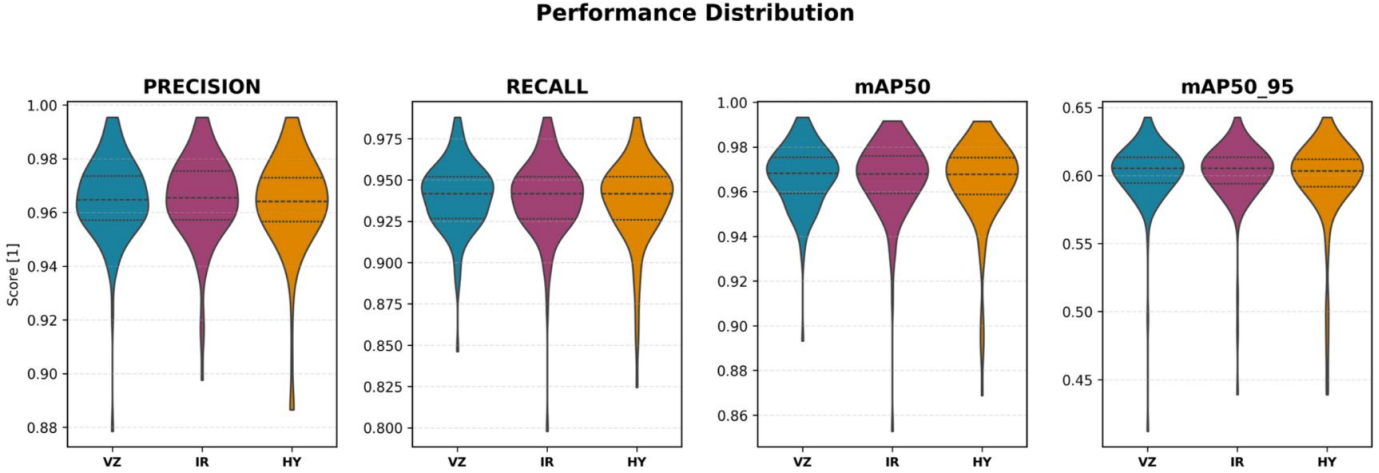
where the Sum of Squares components are calculated on the ranks:

$$SS_{\text{effect}} = n \sum_{j=1}^{k} (\bar{r}_{\cdot j} - \bar{r}_{\cdot\cdot})^2 \tag{8}$$

$$SS_{\text{total}} = \sum_{i=1}^{n} \sum_{j=1}^{k} (r_{ij} - \bar{r}_{\cdot\cdot})^2 \tag{9}$$

Notation: n: number of folds; k: number of modalities; r_{ij}: rank of fold i and modality j; $\bar{r}_{\cdot j}$: mean rank of modality j; $\bar{r}_{\cdot\cdot}$: grand mean rank ($(k+1)/2$).

Following Cohen's conventions for the effect size index f [1], we derive the corresponding η^2 thresholds using the relation $\eta^2 = f^2/(1+f^2)$. Consequently, values of 0.01, 0.06, and 0.14 are considered small, medium, and large effect sizes, respectively.



Modality	Precision	Recall	mAP50	mAP50-95
Visible (VZ)	0.9653 ± 0.0057	0.9401 ± 0.0060	0.9669 ± 0.0039	0.6020 ± 0.0095
Infrared (IR)	0.9653 ± 0.0054	0.9394 ± 0.0066	0.9662 ± 0.0059	0.6015 ± 0.0096
Hybrid (HY)	0.9632 ± 0.0063	0.9369 ± 0.0078	0.9644 ± 0.0058	0.5960 ± 0.0111

Table 2: Distribution and summary statistics of detection performance. The violin plots visualize the kernel density estimation across 25 data blocks, highlighting the stability of results for Visible (VZ), Infrared (IR), and Hybrid (HY) modalities. The table below details the corresponding quantitative metrics (Mean ± Std).

Implementation	Friedman	Wilcoxon	Hommel
SciPy	✓	✓	–
R	✓	✓	✓
Pingouin	✓	✓	–
STAC	✓	–	–
Statsmodels	–	–	✓

Table 3: Statistical Test Implementation Coverage

Metric	SciPy	R	Pingouin	STAC
Precision	0.203	0.203	0.206	0.411
Recall	0.071	0.071	0.070	0.224
mAP50	0.057	0.057	0.057	0.198
mAP50-95	**0.013***	**0.013***	**0.011***	0.082

Table 4: Friedman Test Results Across Implementations

*Significant at $\alpha = 0.05$. STAC uses the Iman-Davenport correction [3] with an F-distribution, whereas SciPy, R, and Pingouin use the standard chi-squared approximation. We report both approaches; however, because our sample size (n=25) is sufficiently large for the chi-squared approximation to be accurate, we base our primary conclusions on the results from SciPy, R, and Pingouin.

Implementation	VZ vs IR	IR vs HY	VZ vs HY
SciPy + Statsmodels	1.000 / 1.000	0.058 / 0.116	0.065 / 0.131
R	0.953 / 0.953	0.056 / 0.111	0.063 / 0.125
Pingouin	1.000 / 1.000	0.060 / 0.119	0.066 / 0.133

Table 5: Pairwise Wilcoxon P-Values for mAP50-95 (Raw / Hommel-Corrected)

No pairwise comparison reached significance at $\alpha = 0.05$ after Hommel correction.

Implementation	VZ vs IR	IR vs HY	VZ vs HY
SciPy + Statsmodels	1.000 / 1.000	0.058 / 0.116	0.065 / 0.131
R	0.953 / 0.953	0.056 / 0.111	0.063 / 0.125
Pingouin	1.000 / 1.000	0.060 / 0.119	0.066 / 0.133

Table 5: Pairwise Wilcoxon P-Values for mAP50-95 (Raw / Hommel-Corrected)

No pairwise comparison reached significance at $\alpha = 0.05$ after Hommel correction.

Metric	η^2	**Interpretation**
Precision	0.064	Medium
Recall	0.106	Medium
mAP50	0.114	Medium
mAP50-95	0.173	Large

Table 6: Effect Sizes (η^2) by Metric