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

Julian Münzer **No. 294094**

January 2026

1 Research

1.1 Research Plan

The primary goal of this research is the comparative analysis of YOLOv12 [6] performance across Visible, Infrared, and Hybrid modalities for real-time UAV detection. This study evaluates the multi-modal fusion approach against single-modality systems. The Hybrid modality is defined as a combination of visible and infrared spectrums within the same training pipeline.

The research objectives are to quantify detection performance across spectral modalities and validate differences using statistical analysis. This includes a comparison of Precision, Recall, and Mean Average Precision alongside non-parametric statistical tests to ensure robust validation across multiple implementations.

Premises and Assumptions:

- Infrared modality provides complementary information to visible light.
- A Hybrid approach may leverage advantages of both spectral domains.
- Repeated stratified k-fold cross-validation ensures robust estimation.
- YOLOv12-nano [6] represents state-of-the-art real-time detection capabilities.

1.2 Datasets

1.2.1 Anti-UAV300 Dataset

The Anti-UAV300 dataset, introduced by Nan Jiang et al. (2021) [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 tracking and detection 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: Statistics of the Anti-UAV300 Dataset [4]

1.2.2 Dataset Pre-processing and Stratification

To ensure efficient experimentation while maintaining statistical robustness, the dataset was processed using a stratified subsampling strategy:

- **Format Conversion:** Source MP4 streams were extracted into individual JPEG frames.
- **Label Cleaning:** Annotations were parsed from the source JSON files. Ground truth bounding boxes were filtered to strictly exclude entries where the `exist` flag was set to 0. Valid boxes were converted to normalized COCO format $(x_{\text{center}}, y_{\text{center}}, w, h)$.
- **Stratified Subsampling:** A subset of $N = 1{,}000$ frames was sampled from the visible spectrum using a cryptographically secure random seed. The sampling maintained the original class balance, resulting in 56 background frames and 944 annotated frames (approximately 1:17 ratio).

1.3 Research Environment

Experiments were executed on macOS (Apple Silicon M4 Chip with 10 CPU Cores and 10 GPU Cores) using Metal Performance Shaders (MPS). The software stack included Python 3.13.5 with PyTorch, scikit-learn, pandas, numpy, scipy, pinguouin for machine learning and statistical analysis; Ultralytics YOLOv12 [6] for object detection; and R 4.5.2 for independent statistical verification.

2 Statistical Methodology and Measures

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.

2.2 Non-Parametric Statistical Testing

Non-parametric tests are appropriate when the assumption of normally distributed data cannot be guaranteed, which is common when comparing machine learning algorithms across multiple folds or datasets. Following the recommendations of Demšar (2006) [2], we employ the Friedman test for omnibus comparison and the Wilcoxon signed-rank test with Hommel correction for pairwise analysis.

2.2.1 Friedman Test

The Friedman test is a non-parametric alternative to repeated-measures ANOVA. It ranks the performance of k algorithms for each of n experimental runs (folds) and tests whether the average ranks differ significantly.

Ranks r_i^j are assigned to modalities j for each fold i, and the test statistic is calculated using:

$$\chi_F^2 = \frac{12n}{k(k+1)} \left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{3}$$

where $R_j = \frac{1}{n} \sum_{i=1}^{n} r_i^j$ is the average rank of the j-th algorithm. Under the null hypothesis that all algorithms perform equivalently, this statistic follows a chi-squared distribution with $k-1$ degrees of freedom.

2.2.2 Pairwise Wilcoxon Signed-Rank Test

When the Friedman test indicates significance, pairwise comparisons are conducted using the Wilcoxon signed-rank test. This test compares two related samples by ranking the absolute differences and computing:

$$z = \frac{\min(R^+, R^-) - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \tag{4}$$

where R^+ and R^- are the sums of ranks for positive and negative differences, respectively.

2.2.3 Hommel's Correction Procedure

When performing multiple pairwise comparisons, the probability of Type I errors increases. Hommel's procedure controls the Family-Wise Error Rate (FWER) while maintaining higher statistical power than the more conservative Bonferroni correction. For m hypotheses at significance level α:

1. Order p-values: $p_1 \leq p_2 \leq \cdots \leq p_m$.
2. Find the largest j such that $p_{m-j+k} > \frac{k\alpha}{j}$ for all $k = 1, \ldots, j$.
3. Reject all hypotheses H_i where $p_i \leq \frac{\alpha}{j}$.

2.3 Effect Size: Eta Squared (η^2)

We utilize η^2 to measure the proportion of variance explained by the modality factor. Effect sizes provide practical significance beyond statistical significance, quantifying the magnitude of differences between groups:

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{5}$$

where the Sum of Squares components are:

$$SS_{\text{effect}} = n \sum_{j=1}^{k} (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 \tag{6}$$

$$SS_{\text{total}} = \sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}_{\cdot\cdot})^2 \tag{7}$$

Notation: n: number of folds; k: number of modalities; x_{ij}: metric value for fold i and modality j; $\bar{x}_{\cdot\cdot}$: grand mean across all observations.

Following Cohen's conventions, η^2 values of 0.01, 0.06, and 0.14 are considered small, medium, and large effect sizes, respectively [1].

3 Multi-Modal UAV Detection Experiments

3.1 General Experimental Design

All experiments share a synchronized architecture to ensure validity and reproducibility. Deterministic behavior was enforced through cryptographically secure seed generation: seven seeds were generated using Python's `secrets` module and hashed through MD5 to produce 32-bit integer seeds. One seed controlled the stratified subsampling, one the cross-validation folding, and five seeds initialized distinct model weight configurations.

The experimental pipeline consisted of: (1) data ingestion and preprocessing, (2) stratified k-fold partitioning, (3) model training with deterministic initialization, and (4) multi-implementation statistical analysis.

The shared configuration across all experiments was:

- **Model:** YOLOv12-nano, pre-trained on COCO [6]
- **Input Resolution:** 640 pixels (images resized preserving aspect ratio)
- **Validation:** Repeated Stratified K-Fold with 5 splits and 5 repeats ($5 \times 5 = 25$ folds per seed)
- **Weight Initialization:** 5 distinct seeds, yielding $25 \times 5 = 125$ training runs per modality
- **Training:** 20 epochs per fold, AdamW optimizer with learning rate 0.01, batch size 16
- **Augmentation:** Standard YOLO augmentations including mosaic, random flip, and HSV adjustments

Final metrics were computed by averaging across the 5 seed initializations for each of the 25 folds, producing 25 aggregated samples per modality for statistical analysis.

3.2 Experiment 1: Visible Light (VZ)

Input Data: Standard RGB camera imagery extracted from the visible stream of the Anti-UAV300 dataset. Images have a native resolution of 1920×1080 pixels, resized to 640×360 during training.

Goal: Establish baseline detection performance under standard lighting conditions.

3.3 Experiment 2: Infrared (IR)

Input Data: Thermal infrared imagery synchronized to the visible subset frames. Images have a native resolution of 640×512 pixels.

Goal: Evaluate detection capabilities in the thermal spectrum, which captures heat signatures independent of ambient lighting.

3.4 Experiment 3: Hybrid (HY)

Input Data: An interleaved dataset comprising the first 50% of visible frames and the second 50% of infrared frames within the same training batches.

Goal: Assess whether early-fusion multi-modal training improves generalization by exposing the model to both spectral domains simultaneously.

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

Table 2: Performance Comparison Across Modalities

Values reported as mean $\pm$ standard deviation across 25 cross-validation folds.

4 Statistical Analysis Results

To guarantee reliability and guard against implementation-specific artifacts such as value approximations, results were cross-verified across five independent statistical implementations: SciPy and Pingouin in Python, R's native statistical functions, STAC (Statistical Tests for Algorithms Comparison) [5], and Statsmodels for Hommel correction. Not all implementations support all tests; Table 3 summarizes coverage.

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

Table 3: Statistical Test Implementation Coverage

4.1 Friedman Test Results

The Friedman test was applied to determine whether significant differences exist among the three modalities. Results were consistent across implementations (Table 4).

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.

The Friedman test revealed a significant difference among modalities only for mAP50-95 ($p = 0.013$, significant at $\alpha = 0.05$). Precision, Recall, and mAP50 showed no significant omnibus differences, indicating that modality choice does not meaningfully affect these metrics.

4.2 Pairwise Wilcoxon Tests with Hommel Correction

For mAP50-95, where the Friedman test indicated significance, pairwise Wilcoxon signed-rank tests were conducted between all modality pairs: VZ vs IR, IR vs HY, and VZ vs HY. Raw p-values were corrected using Hommel's procedure.

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.

While the Friedman test indicated an overall effect of modality on mAP50-95, subsequent Hommel-corrected pairwise comparisons did not identify specific significant pairs. The IR vs HY and VZ vs HY comparisons approached significance (uncorrected $p \approx 0.06$), but corrections for multiple testing rendered these non-significant. This suggests that the Hybrid modality may introduce subtle degradation in localization accuracy, but the effect is not strong enough to survive correction.

4.3 Effect Size Analysis

Effect sizes (η^2) quantify the practical magnitude of modality effects independent of sample size (Table 6).

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

The effect size for mAP50-95 ($\eta^2 = 0.173$) exceeds Cohen's threshold of 0.14 for a large effect, indicating that modality explains a substantial proportion of variance in localization accuracy. This aligns with the significant Friedman result and suggests that while all modalities achieve similar detection rates, precision of bounding box localization varies meaningfully across spectral inputs.

5 Summary of Research

This study evaluated YOLOv12-nano performance across three spectral modalities for UAV detection using a rigorous experimental design with 125 training runs per modality and multi-implementation statistical validation.

Modality Equivalence: Visible and Infrared modalities achieved statistically equivalent performance across Precision (96.53%), Recall (≈94%), and mAP50 (≈96.7%). This confirms that uncalibrated thermal imagery is as effective as RGB for UAV detection in this benchmark, validating infrared as a viable alternative for conditions where visible light is insufficient.

Hybrid Performance: The interleaved Hybrid approach resulted in slight degradation, particularly in localization accuracy (mAP50-95: 59.60% vs 60.20% for VZ). The Friedman test detected a significant overall modality effect on mAP50-95 ($p = 0.013$) with a large effect size ($\eta^2 = 0.173$). However, pairwise comparisons did not survive Hommel correction, suggesting the degradation is subtle. This may be attributed to the model struggling to generalize across alternating spectral inputs without a dedicated fusion architecture.

Conclusion: For the Anti-UAV300 benchmark using YOLOv12, single-modality systems (either Visible or Infrared) are sufficient and may even be preferable to naive multi-modal fusion. Future work should explore late-fusion or attention-based multi-modal architectures that can better leverage complementary spectral information.

References

[1] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, 1988.

[2] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[3] Ronald L. Iman and James M. Davenport. Approximations of the critical region of the Friedman statistic. *Communications in Statistics - Theory and Methods*, 9(6):571–595, 1980.

[4] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Jian Zhao, Guodong Guo, and Zhenjun Han. Anti-uav: A large multi-modal benchmark for uav tracking, 2021.

[5] Ismael Rodríguez-Fdez, Adrián Canosa, Manuel Mucientes, and Alberto Bugarín. STAC: a web platform for the comparison of algorithms using statistical tests. In *Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015.

[6] Yunjie Tian, Qixiang Ye, and David Doermann. Yolo12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.