Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Does Quality-Aware Late Fusion Help? Visual–Thermal Monkey Detection Across Times of Day

Izatul Lail Mohd Yasar

ellail@mardi.gov.my

Malaysian Agricultural Research and Development Institute

Muhammad Izzat Farid Musaddin

Malaysian Agricultural Research and Development Institute

Research Article

Keywords: late fusion, visual–thermal imaging, wildlife detection, weighted box fusion, thermal crossover, domain shift

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11793202/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When Does Quality-Aware Late Fusion Help? Visual–Thermal Monkey Detection Across Times of Day

Izatul Lail Mohd Yasar[1]*, Muhammad Izzat Farid Musaddin[1]

[1]ICT Management Centre, Malaysian Agriculture Research and Development Institute (MARDI), Serdang, Selangor, Malaysia

Corresponding author: ellail@mardi.gov.my

Abstract Round-the-clock detection of monkeys near human settlements is hampered by complementary sensor failures: RGB cameras fail in low light, and thermal cameras fail when animals and background have similar temperatures. Late fusion of RGB and thermal detectors is a common remedy, and quality-aware weighting is often proposed to improve it. We ask when such weighting helps. Using paired RGB–thermal frames of long-tailed macaques recorded in Selangor, Malaysia, at dawn, midday, dusk and night (290 training and 120 test pairs from different scenes), we fuse two YOLO11n detectors with Weighted Box Fusion using fixed weights or adaptive weights computed from image brightness and thermal Michelson contrast (Adaptive Late Fusion, ALF). Under session-grouped cross-validation with session-level bootstrap intervals, modality reliability reversed with time of day: thermal reached 94.4% mAP@0.5 versus 10.0% for visual at dawn/dusk, whereas visual reached 82.4% versus 17.0% for thermal at midday. Equal-weight fusion reached 70.8% pooled mAP@0.5 and ALF 69.2% (difference −1.7 points, 95% CI −5.2 to 3.8). ALF was numerically equal or better within every time window (up to +3.3 points at night) but not when scores were pooled, because condition-dependent weights shift confidence scales; the thermal-contrast cue never fired. Oracle weights indicate about 4 points of headroom. Equal-weight fusion is a strong baseline, and simple brightness and contrast cues help only where they track the failing modality.

Keywords late fusion · visual–thermal imaging · wildlife detection · weighted box fusion · thermal crossover · domain shift

1 Introduction

Monkeys increasingly enter residential and agricultural areas in Malaysia, motivating automated, round-the-clock detection for early warning and deterrence [1]. Cameras face complementary failure modes. RGB cameras degrade in low light and under canopy shadows. Thermal infrared (TIR) cameras do not depend on illumination, but they lose contrast when the background temperature approaches that of the animal (thermal crossover), which is reported around sunrise and in the late afternoon [2]. Combining the two modalities is therefore attractive. In late fusion, each modality has its own detector and the outputs are merged at box level, for example by Weighted Box Fusion (WBF) [3]. This reuses off-the-shelf detectors and allows independent training, and visible–thermal fusion has been shown to improve automated wildlife detection [4].

Most late-fusion pipelines apply one fixed weight ratio to every frame, which implicitly assumes that the relative reliability of the sensors is constant. Quality-aware weighting is a natural refinement and has been explored, for instance, with illumination cues for multispectral pedestrian detection [5]. Whether simple image-level cues suffice for wildlife in tropical settings, and how much they gain over a well-chosen fixed weight, remains unclear. The question is difficult to answer with small, site-specific datasets, because evaluating on frames from the training sessions inflates results [6, 7].

We implement a lightweight quality-aware scheme, Adaptive Late Fusion (ALF), which derives frame-specific WBF weights from the brightness of the RGB frame and the global Michelson contrast

of the thermal frame, and use it as a probe into the question in the title. We address three questions. **RQ1:** how does the reliability of RGB and thermal detectors vary across the day, and how much does it depend on the training scene? **RQ2:** does ALF improve over single-modality detectors and fixed-weight fusion? **RQ3:** where does ALF help or fail, and what limits it? Our contributions are (i) a paired RGB–thermal dataset of long-tailed macaques recorded at four times of day, (ii) a session-grouped evaluation with session-level bootstrap intervals, (iii) a quantification of condition-dependent complementarity and of the headroom for condition-level weighting, and (iv) an ablation-based analysis showing where the brightness cue helps, why the thermal-contrast cue did not fire, and how condition-dependent weights can distort pooled score calibration.

2 Related Work

Region-based and single-stage detectors [8, 9, 10] are standard tools in wildlife monitoring, including camera-trap analysis [11]; we use YOLO11n [12]. Multimodal detectors are commonly grouped into early, feature-level and late fusion. In late fusion, overlapping boxes from different detectors are merged by non-maximum suppression [13], Soft-NMS [14] or WBF [3], which averages coordinates and scores within clusters instead of discarding boxes. Multispectral benchmarks [15] and illumination-aware weighting [5] motivate adaptive schemes, and recent work reports gains from visible–thermal fusion for wildlife [4]. Two gaps remain. First, adaptive weights act on detector confidences, whose calibration across conditions is not guaranteed [16], so their effect on a pooled ranking is rarely examined. Second, results are rarely broken down by time of day or evaluated across recording sessions. This paper addresses both.

3 Method

3.1 Adaptive Late Fusion

A visual and a thermal detector run independently on the RGB and thermal frame of each pair. A Visual Quality Estimator (VQE) and a Thermal Quality Estimator (TQE) assess the frames and produce weights that rescale the detector confidences before WBF (Fig. 1).

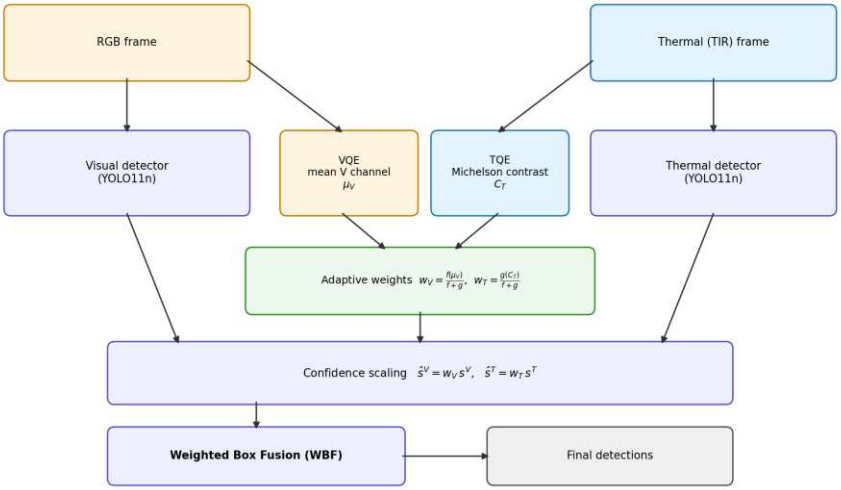

Fig. 1 ALF framework. Detector confidences are rescaled by quality-derived weights before Weighted Box Fusion.

VQE. The mean of the HSV V channel of an $H \times W$ RGB frame (V scaled to [0, 1]) is

$$\mu v = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} V(i,j)$$

Frames that are too dark or overexposed receive low visual quality,

$$f(\mu_V) = \begin{cases} Q_{V,low} & if\ \mu_V < \tau_{dark}\ or\ \mu_V > \tau_{bright} \\ Q_{V,high} & otherwise \end{cases}$$

TQE. For the 8-bit grey thermal frame scaled to [0, 1] (no per-frame stretching), the global Michelson contrast is

$$C_T = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}$$

A low C_T is hypothesised to accompany thermal crossover,

$$g(C_T) = \begin{cases} Q_{T,low} & if\ C_T < \tau_{cross} \\ Q_{T,high} & if\ C_T \geq \tau_{cross} \end{cases}$$

Weights and fusion. The weights are normalised to sum to one,

$$w_V = \frac{f(\mu_V)}{f(\mu_V) + g(C_T)}, \quad w_T = \frac{g(C_T)}{f(\mu_V) + g(C_T)}$$

Each detection confidence is then rescaled,

$$\hat{s}_i^V = w_V \cdot s_i^V, \ \hat{s}_j^T = w_T \cdot s_j^T$$

and WBF merges the boxes of a cluster C with confidence-weighted coordinate averaging,

$$\hat{b} = \frac{\sum_{k \in C} \hat{s}_k \cdot b_k}{\sum_{k \in C} \hat{s}_k}$$

We used $\tau_{dark} = 0.2$, $\tau_{bright} = 0.85$, $\tau_{cross} = 0.25$, $Q_{V,low} = Q_{T,low} = 0.3$ and $Q_{V,high} = Q_{T,high} = 0.8$; they were set from preliminary observations before the evaluation and were not tuned on test data. The four cases yield (w_V, w_T) = (0.50, 0.50) for adequate lighting and contrast, (0.73, 0.27) for suspected crossover, (0.27, 0.73) for poor lighting, and (0.50, 0.50) when both are flagged. WBF used an IoU threshold of 0.55 and score averaging. Because the weights sum to one, a condition-dependent pair also changes the overall scale of the fused confidences. As a post hoc diagnostic, introduced after inspecting the results, we also evaluate a max-normalised variant with $w_V/\max(w_V, w_T)$ and $w_T/\max(w_V, w_T)$, which keeps the dominant modality's score scale.

3.2 Data

Frames were recorded in Kajang, Selangor, with two smartphones mounted side by side on a hand-held stick: a Xiaomi Mi MIX 3 (RGB) and a 7.5–14 μm spectral range Seek Thermal X module on a Honor Magic 6 Pro (TIR). The cameras are not co-located, so fields of view differ. Recordings covered dawn (07:00–07:30), midday (12:00–13:00), dusk (19:00–19:30) and night (21:00–22:00) in 13th June 2026. The **training set** has 290 pairs from seven sessions at a different scene (monkeys on overhead cables, top of roof, tree), 864×480 pixels, annotated separately per modality because the frames are not aligned. The **test set** has 120 pairs from 13 sessions at a different scene (monkeys on overhead cables, concrete fence, top of roof, tree), 30 per time window; each pair was cropped to a common field of view and resized to 575×320 pixels, so one set of boxes serves both modalities (207 instances, one class, *Macaca fascicularis*). Boxes were drawn in CVAT and Roboflow.

3.3 Detectors and Evaluation

Two YOLO11n detectors [12], one per modality, were initialised from COCO weights and trained for 100 epochs (batch 16, initial learning rate 0.01, input 575 px, default Ultralytics augmentation) with one class. *Source-only* detectors were trained on the 290 training pairs. For the main results we used *session-grouped 3-fold cross-validation*: the 13 test sessions were split into three folds of 40 pairs balanced across time windows; for each fold both detectors were retrained on the training pairs plus the test sessions of the other two folds, using the final-epoch checkpoint, and applied to the held-out fold. The three held-out predictions were pooled, so every test pair is predicted by detectors that never saw its session. Frames within a session are highly correlated, which is why random frame-level splits would be optimistic. The "best static" baseline picks, for each held-out fold, the visual weight (grid 0–1, step 0.05) that maximises mAP@0.5 on the other folds' out-of-fold predictions.

We report mAP@0.5, mAP@0.5:0.95 (COCO-style 101-point AP [17]), and precision and recall at the F1-optimal confidence threshold. Results are *pooled*: all pairs of a window (or of the test set) are ranked together, as with a single global threshold. Intervals use a stratified session-level bootstrap [18] (1000 resamples of sessions within each window), and paired differences use the same resamples.

Ablations fix one estimator at its high-quality value. Oracle weights and fixed-weight sweeps use test labels and are descriptive bounds only.

4 Results

4.1 Training Scene and Overall Performance

Source-only detectors transferred poorly to the new scene (Table 1): 3.0% mAP@0.5 for the visual and 41.7% for the thermal detector. Adding other test-scene sessions raised these to 25.3% and 61.9%, and fusion from about 40% to about 70%. This comparison measures the benefit of in-scene data, since the folds share a scene. With cross-validated detectors, equal-weight fusion reached 70.8%, the best static weight 68.7% and ALF 69.2%. ALF had the highest precision and recall at the F1-optimal threshold but a lower mAP@0.5 than equal weights, and its mAP@0.5:0.95 (35.3%) was marginally below that of the thermal detector (35.8%).

Table 1 Pooled performance on the 120 test pairs (%). CV: session-grouped cross-validation; brackets: 95% session-level bootstrap interval.

Method	Source-only mAP@0.5	CV mAP@0.5 [95% CI]	mAP@0.5:0.95	P	R
Visual-only	3.0	25.3 [8.4, 45.3]	9.2	70.6	23.2
Thermal-only	41.7	61.9 [47.0, 77.6]	35.8	68.2	56.0
Fixed WBF (0.5/0.5)	40.9	70.8 [62.1, 83.2]	36.5	73.1	68.1
Fixed WBF (best static)	–	68.7 [60.3, 82.4]	34.7	69.6	69.6
ALF	39.6	69.2 [64.4, 79.0]	35.3	74.1	70.5

Paired bootstrap comparisons show that ALF was significantly better than the visual detector (+43.9 points, $p = 0.001$) but not significantly better than the thermal detector (+7.2; 95% CI −5.6 to 22.3; $p = 0.19$), and indistinguishable from equal weights (−1.7; CI −5.2 to 3.8; $p = 0.52$) and from the best static weight (+0.5; CI −4.0 to 5.5; $p = 0.83$). We therefore do not claim that ALF outperforms fixed-weight fusion in this dataset.

4.2 Reliability Across Times of Day

The reliability of the two modalities reversed with time of day (Table 2). At dawn and dusk the thermal detector reached 94.4% against 10.0% for visual; at night 61.1% against 1.1%; at midday visual reached 82.4% against 17.0% for thermal. The weak midday thermal result is consistent with thermal crossover, but we did not measure scene temperatures. Sweeping a fixed visual weight within each window (Fig. 2) shows a sharp dependence: at dawn and dusk mAP@0.5 stayed near 97% and 89% up to $w_V \approx 0.5$ and collapsed beyond about 0.6; at night it fell monotonically from 62.8% at $w_V = 0$; at midday it rose from about 15% at $w_V = 0$ to about 82% at $w_V \approx 0.6$–0.7. No single weight is best for all windows.

Table 2 mAP@0.5 (%) by time window (dawn and dusk pooled) and pooled over all pairs.

Method	Dawn/Dusk	Midday	Night	All
Visual-only	10.0	**82.4**	1.1	25.3
Thermal-only	**94.4**	17.0	**61.1**	61.9
Fixed WBF (0.5/0.5)	91.5	77.3	54.7	**70.8**
Fixed WBF (best static)	89.6	77.3	53.6	68.7
ALF (= VQE only)	88.4	77.3	58.0	69.2

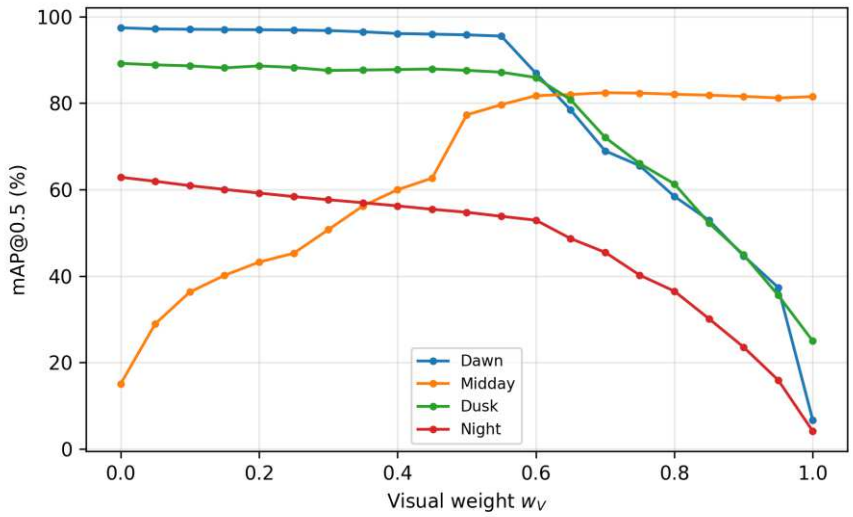

Fig. 2 mAP@0.5 versus visual weight w_V ($w_T = 1 - w_V$) within each time window (test labels; descriptive).

4.3 Quality Estimators and Ablation

The VQE flagged the RGB frame as poorly lit in all dusk and night frames, in 10% of dawn frames and in none at midday (mean w_V: 0.27 at dusk and night, 0.48 at dawn, 0.50 at midday). The TQE never fired: C_T had window means of 0.71–0.80 and no frame fell below $\tau_{cross} = 0.25$. The midday frames, where the thermal detector performed worst, had a mean contrast of 0.71, similar to other windows, so this global max/min statistic did not discriminate the failing condition. Accordingly, the TQE-only configuration equalled equal-weight fusion and the VQE-only configuration equalled full ALF. The VQE improved night performance (54.7% to 58.0%) but reduced the pooled dawn/dusk result (91.5% to 88.4%); the pooled effect (−1.7 points) is not distinguishable from zero.

4.4 Per-Window Versus Pooled Performance, and Headroom

Within each window separately (Table 3), ALF was numerically equal to or better than equal weights in all four (mean 79.8% versus 78.8%), yet its pooled mAP@0.5 was lower (69.2% versus 70.8%). This is consistent with a calibration effect: weights multiply detector confidences, so condition-dependent weights shift the confidence scale across conditions, and a single global ranking penalises this. The post hoc max-normalised variant left per-window results unchanged and raised the pooled value to 71.6%; as it was devised after seeing the results and the gap lies within the intervals above, we treat this as exploratory. Oracle per-window weights (test labels) reach a mean of 83.0%, 4.2 points

above equal weights, of which ALF recovered about 1.0 point; the largest remaining gaps are at midday (+5.1 over ALF) and night (+4.8).

Table 3 mAP@0.5 (%) within each window, mean of the four windows, and pooled. Oracle visual weights: 0.0 (dawn), 0.7 (midday), 0.0 (dusk), 0.0 (night).

Method	Dawn	Midday	Dusk	Night	Mean	Pooled
Fixed WBF (0.5/0.5)	95.8	77.3	87.6	54.7	78.8	70.8
ALF	95.9	77.3	88.0	58.0	79.8	69.2
ALF, max-normalised (post hoc)	95.8	77.3	88.0	58.0	79.8	71.6
Oracle w_V per window	97.4	82.4	89.2	62.8	83.0	74.5

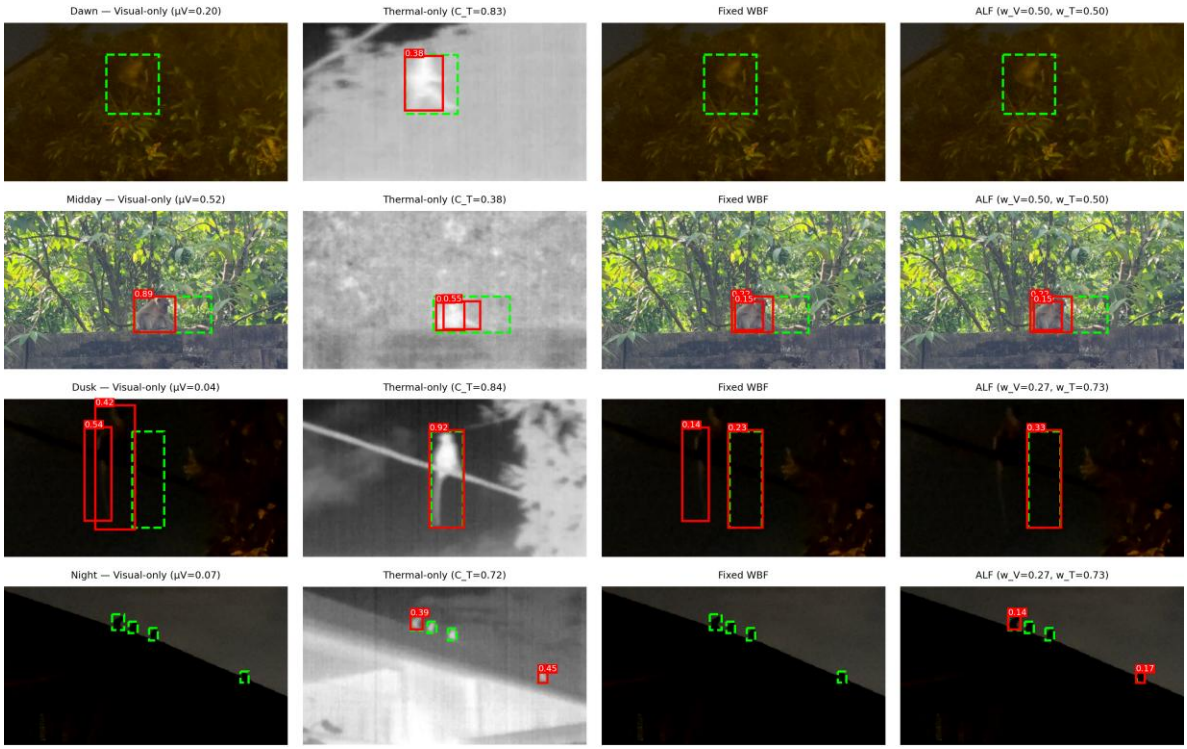

Fig. 3 Representative test frames per window: visual-only, thermal-only, fixed-weight fusion and ALF detections (red) and ground truth (green dashed).

5 Discussion

When does quality-aware fusion help? Fusion itself was valuable: equal weights raised pooled mAP@0.5 by about 9 points over the best single detector (point estimates; the paired interval for ALF versus thermal was wide). The modalities were strongly complementary, with a sharp transition between weights that help and harm. A quality-aware rule helped only where its cue tracked the failing modality: the brightness cue correctly down-weighted RGB at dusk and night and gave a small gain at night, but could not flag RGB failure at dawn, when frames were not dark (mean μ_V = 0.33) yet monkeys were nearly invisible, and offered no signal at midday. The thermal-contrast cue never fired, leaving the midday window, where visual was far better than thermal, at equal weights. Finally, weights that multiply confidences alter score scales across conditions, which can offset per-window

gains in a pooled ranking. Equal-weight WBF was a strong baseline, within 2 points of the oracle at dawn and dusk; a sensible default is equal-weight WBF, with adaptive weighting reserved for cases where a reliable cue for the failing modality exists and score scales are controlled.

Domain gap. Detectors trained on one scene failed on another (3.0% for visual), and in-scene sessions added 20–30 points. Quality-aware fusion cannot repair a detector that does not generalise, and evaluation on training sessions overestimates field performance.

Limitations. The data are small (120 test pairs, 13 sessions, one site and species), so intervals are wide and many differences between fusion variants are not significant. The cross-validation shares a scene across folds and thus measures cross-session, not cross-site, generalisation, and the best-static selection has mild stacking-type leakage. The cameras are not co-located and test frames were cropped to share one set of boxes, so residual misalignment may affect box-level fusion. The ALF parameters were fixed in advance and not tuned. The oracle analysis and the max-normalised variant are post hoc, and we did not measure scene temperatures, so attributing the midday thermal failure to crossover is an inference. Only YOLO11n was evaluated. Future work should use cues closer to the failure modes (local thermal contrast, histogram saturation, detector-confidence statistics), learn the weighting from validation data, calibrate scores per modality before fusion, use a calibrated co-located rig, and test other sites and species.

6 Conclusion

Using paired RGB–thermal frames of macaques across the day, we found that (i) detectors trained on one scene transferred poorly, and in-scene data raised mAP@0.5 by 20–30 points; (ii) modality reliability reversed with time of day, and equal-weight fusion was a strong, robust baseline (70.8% pooled); and (iii) a lightweight brightness-and-contrast ALF rule was numerically equal or better than equal weights within every window but did not improve the pooled result (69.2%; −1.7 points, 95% CI −5.2 to 3.8), because its thermal cue never fired and its condition-dependent weights distorted pooled score calibration. Adaptive weighting has about 4 points of headroom, but realising it requires cues that track the failing modality and calibrated confidence scores.

Declarations

This research received no funding.

Conflict of interest: No conflict of interest in this research

Ethics approval: No ethics approval was needed in this research, as the monkeys commonly appear in residential areas at the research site.

Data and code availability https://github.com/ellail-mardi/monkey-detection/

Author contributions

I.L. Mohd Yasar: Conceptualization, Methodology, Data curation, Writing – original draft

M.I.F. Musaddin: Writing – review & editing, Validation, Data curation

References

[1] Md-Zain, B. M., Ruslin, F., & Idris, W. M. R. (2014). Human-Macaque Conflict at the Main Campus of Universiti Kebangsaan Malaysia. Pertanika Journal of Tropical Agricultural Science, 37(1), 73-85 (2014)

[2] Burke, C., Rashman, M., Wich, S., Symons, A., Theron, C., Longmore, S.: Optimizing observing strategies for monitoring animals using drone-mounted thermal infrared cameras. Int. J. Remote Sens. 40(2), 439–467 (2019)

[3] Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image Vis. Comput. 107, 104117 (2021)

[4] Dionne-Pierre, C., Foucher, S., Théau, J., Lemaître, J., Charbonneau, P., Brousseau, M., Varin, M.: Fusion or confusion? Assessing the impact of visible-thermal image fusion for automated wildlife detection. Remote Sens. Ecol. Conserv. (2025). 0: 1 - 12. https://doi.org/10.1002/rse2.70089

[5] Li, C., Song, D., Tong, R., Tang, M.: Illumination-aware Faster R-CNN for robust multispectral pedestrian detection. Pattern Recognit. 85, 161–171 (2019)

[6] Roberts, D.R., Bahn, V., Ciuti, S., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8), 913–929 (2017)

[7] Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: European Conference on Computer Vision (ECCV) (2018)

[8] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems 28 (2015)

[9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proc. IEEE CVPR, pp. 779–788 (2016)

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot multibox detector. In: ECCV 2016, pp. 21–37 (2016)

[11] Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proc. Natl. Acad. Sci. USA 115(25), E5716–E5725 (2018)

[12] Jocher, G., Qiu, J.: Ultralytics YOLO11 (version 11.0.0). https://github.com/ultralytics/ultralytics (2024)

[13] Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR), vol. 3, pp. 850–855 (2006)

[14] Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS – improving object detection with one line of code. In: Proc. IEEE ICCV, pp. 5561–5569 (2017)

[15] Hwang, S., Park, J., Kim, N., Choi, Y., Kweon, I.S.: Multispectral pedestrian detection: benchmark dataset and baseline. In: Proc. IEEE CVPR, pp. 1037–1045 (2015)

[16] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proc. ICML, PMLR 70, pp. 1321–1330 (2017)

[17] Lin, T.-Y., Maire, M., Belongie, S., et al.: Microsoft COCO: common objects in context. In: ECCV 2014, pp. 740–755 (2014)

[18] Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall, New York (1993)