Multispectral UAV Fusion for Native Tree Species Recognition in the Cerrado

Pedro Fernandes de Oliveira[1][0009-0000-3649-0580], Ana Sofia S. Silvestre[1][0000-0002-7219-986X], Enzo Z. Celentano[1][0009-0008-7101-2128], Leonardo L. Bergamini[2][0000-0001-8505-5059], Frederico S. C. Takahashi[2][0000-0001-9410-7885], and Díbio L. Borges[1][0000--0002-4868-0629]

[1] University of Brasília, Department of Computer Science, Brasília, DF, Brazil
[2] IBGE Ecological Reserve, Brazilian Institute of Geography and Statistics, Brasília, DF, Brazil

Abstract. The Brazilian Cerrado, one of the world's most biodiverse and threatened savannas, urgently requires scalable tools for identifying native tree species at the individual level. This work presents the first deep learning benchmark for this task, introducing a curated dataset of 1,620 UAV multispectral images covering 16 native Cerrado species, fully annotated by field biologists at an Ecological Reserve. We propose a Gram-Schmidt fusion pipeline that condenses RGB, Red Edge, and NIR bands into a pseudo-RGB representation, enabling unmodified object detectors to exploit near-infrared information. Eight models were evaluated across four modalities (fused, RGB, NDRE, NDVI). The fused modality consistently outperformed all others, with RT-DETR-L and YOLOv5s achieving 85.2% mAP@50. Per-species analysis revealed that near-infrared reflectance is critical for rare and morphologically ambiguous taxa, yielding gains exceeding 70 percentage points over RGB-only inputs for several species. These results demonstrate that multispectral fusion is not merely beneficial but essential for fine-grained Cerrado species discrimination, establishing a practical foundation for automated biodiversity monitoring in this ecologically critical biome. Code and data available at github.com/pedrofernandss/cerrado_tree_species_id.

Keywords: Native Tree Species Recognition · Cerrado · Multispectral Imagery · Deep Learning · Image Fusion · Remote Sensing

1 Introduction

The Cerrado is the world's most diverse tropical savanna and a biodiversity hotspot [11] , yet nearly 50% of its original vegetation has been lost to agricultural expansion and deforestation [1]. Automated monitoring tools are urgently needed for conservation efforts [4, 13].

Tree detection is a key strategy for ecosystem monitoring. Remote sensing approaches vary widely in cost and capability. While LiDAR-derived data such as Canopy Height Models offer rich 3D information, global LiDAR coverage remains scarce and prohibitively expensive [19]. Multispectral UAVs provide a scalable, cost-effective alternative for large-scale vegetation analysis [14, 20].

Despite the Cerrado's ecological importance and accelerated degradation, no deep learning benchmark exists for native tree species identification using UAV imagery — likely due to the lack of annotated multispectral datasets and the morphological similarity among many species.

This work presents the first deep learning benchmark for native tree species detection in the Cerrado using UAV-acquired multispectral imagery, evaluating eight object detection architectures across four imaging modalities. The main contributions are as follows:

- The first deep learning pipeline specifically designed for native Cerrado tree species detection from UAV multispectral imagery.
- A curated dataset of 1,620 multispectral images covering 16 native species, fully annotated by field biologists at the IBGE Ecological Reserve.
- A Gram-Schmidt-based fusion of RGB and multispectral bands that substantially outperforms RGB-only approaches, enabling reliable detection even for morphologically ambiguous and data-scarce taxa.

2 Related Work

Tree species classification from UAV imagery has been extensively explored using RGB and multispectral sensors, yet most studies focus on temperate or subtropical forests. Huang et al. [21] applied CNNs in subtropical forests, addressing challenges such as intra-class variability and inter-class similarity — issues also present in the Cerrado.

Multispectral and hyperspectral data have shown clear advantages for tree identification. Fassnacht et al. [8] emphasized the role of spectral diversity in vegetation mapping more broadly. To identify urban trees, Ventura et al. [19] used multispectral imagery and NDVI derived from Red and NIR bands, demonstrating improvements over RGB-only approaches — suggesting even greater potential in biodiverse biomes like the Cerrado. Yu et al. [20] evaluated a single-sensor solution using multispectral airborne laser scanning and found that multispectral intensity features provide significantly more information for species classification than point cloud structure alone.

Despite these advances, no prior work has systematically benchmarked modern object detectors (e.g., YOLOv5, RT-DETR) on multispectral UAV data for tropical savanna tree species, nor proposed a fusion method that preserves NIR information without modifying existing architectures. The proposed Gram-Schmidt fusion addresses this gap by enabling any off-the-shelf RGB detector to exploit NIR information — bypassing the need for architectural redesign while achieving substantial gains over RGB-only inputs for Cerrado native species.

3 Materials and Methods

3.1 Study Area, Drone, and Cameras

Study Area This study was conducted at the IBGE Ecological Reserve (15°56'31"S, 47°52'47"W), located in the Federal District of Brazil. The reserve covers approximately 1,300 hectares with a perimeter of 14 km and features typical Cerrado sensu stricto vegetation. The region has a tropical seasonal climate (Aw, Köppen-Geiger), with an average annual temperature of 22°C and a distinct dry season from May to September.

Drone and Sensors We used the DJI Mavic 3M multispectral drone [6], a quadcopter equipped with two imaging systems: (1) a 20 MP RGB camera, and (2) a four-band multispectral sensor capturing green (550 nm ± 16 nm), red (650 nm ± 16 nm), red edge (730 nm ± 16 nm), and near-infrared (NIR, 860 nm ± 26 nm) bands. Radiometric calibration was performed using a reflectance panel before each flight mission.

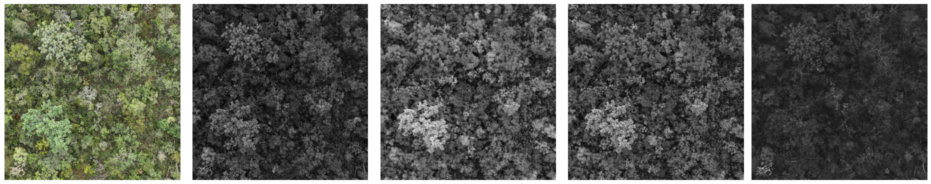
Data Acquisition Flights were conducted on three dates (August 28, 2024; January 27, 2025; January 28, 2025) between 10 AM and 2 PM under clear sky conditions to minimize shadow and illumination variability. Flight altitude was 40 m above ground level, with longitudinal and lateral overlap set to 75% and 70%, respectively. A total of 324 individual tree crowns were randomly sampled across the reserve, covering 16 native Cerrado tree species identified and annotated in situ by field biologists. Each tree crown image contains five channels: one RGB composite and four individual spectral bands (green, red, red edge, NIR), resulting in a total of 1,620 images.

3.2 Image Preprocessing

Vignetting Correction The initial step involved correcting the vignetting effect. Applying vignetting correction is crucial for subsequent image processing steps, ensuring more uniform illumination across the image and preventing brightness variations from negatively impacting analysis [22]. The parameters for the drone model were extracted from the XML metadata embedded in the image files by the cameras. We applied the vignetting correction to the input image $I(x, y)$ as described in Eq. (1).

$$I_{\text{corr}}(x, y) = I(x, y) \times \left(\sum_{i=0}^{5} k[i] \cdot r^{6-i} + 1.0 \right) \tag{1}$$

Here, k represents the polynomial coefficients for vignetting correction (from XML metadata), and r is the Euclidean distance from pixel (x, y) to the image center.



(a) RGB (b) Green Band (c) NIR Band (d) RE Band (e) Red Band

Fig. 1: Examples showing aligned RGB and multispectral images.

Geometric Distortion Correction Lens distortion was corrected using manufacturer-provided intrinsic parameters (focal lengths f_x, f_y and principal point c_x, c_y), together with radial (k_1, k_2, k_3) and tangential (p_1, p_2) distortion coefficients embedded in the XML metadata.

Inter-band Image Alignment Alignment of multi-channel images proceeds in two steps. First, a metadata-based homography (Eq. 2) corrects coarse perspective misalignments from camera geometry. Second, fine alignment uses SIFT feature detection [9] and RANSAC estimation, followed by Enhanced Correlation Coefficient (ECC) refinement [7]. Figure 1 shows the final aligned bands.

$$dst(x,y) = \left(\frac{H_{11}x + H_{12}y + H_{13}}{H_{31}x + H_{32}y + H_{33}}, \frac{H_{21}x + H_{22}y + H_{23}}{H_{31}x + H_{32}y + H_{33}} \right) \tag{2}$$

Normalization and Image Resizing To ensure consistency across the red, red edge, NIR (Near-Infrared), and green spectral channels, which have values ranging from 0 to 65535, the channels were normalized to match the 8-bit range of the RGB channels, which range from 0 to 255.

Additionally, all RGB images were resized from 5280×3956 to 1024×1024 to meet the requirements of image alignment algorithms, which depend on uniformly sized inputs for accurate feature matching and transformation across images.

3.3 Vegetation Index Calculation

Given the unique environmental characteristics of the Cerrado biome, our training pipeline incorporates two robust vegetation indices to evaluate the feasibility of identifying tree species based on their health and internal physiological states: the Normalized Difference Vegetation Index (NDVI) [16] and the Normalized Difference Red Edge Index (NDRE) [2].

NDVI is one of the most established indices for identifying active vegetation by exploiting the contrast between strong chlorophyll absorption in the red band and high reflectance in the near-infrared (NIR) spectrum. As such, NDVI provides an effective proxy for vegetation presence, canopy density, and seasonal dynamics, being particularly useful for large-scale vegetation monitoring. The index is computed as follows:

$$NDVI = \frac{NIR - RED}{NIR + RED} \tag{3}$$

Despite its effectiveness, NDVI is known to suffer from saturation effects in areas with dense vegetation and reduced sensitivity to early physiological stress. This limitation is especially relevant in heterogeneous biomes such as the Cerrado, where vegetation structure and phenology vary substantially across space and time. To address this limitation, we also employ the NDRE, which replaces the red band with the red-edge band, a narrow spectral region located between the red and NIR wavelengths.

The red-edge region is particularly sensitive to variations in chlorophyll concentration and can penetrate deeper into leaf layers, allowing for improved characterization of vegetation condition in denser canopies and during different phenological stages [18]. NDRE is therefore better suited for detecting subtle physiological changes that may not yet be reflected in canopy structure. It is defined as:

$$NDRE = \frac{NIR - RE}{NIR + RE} \tag{4}$$

NDVI and NDRE were computed as single-channel inputs for comparative experiments (Section 4), serving as baseline vegetation indices against our fused pseudo-RGB approach.

3.4 Multispectral Image Fusion via Gram-Schmidt Pansharpening

To enable compatibility with off-the-shelf RGB object detectors while preserving the rich spectral information from multispectral bands, we reduced the input dimensionality from 7 to 3 channels using a Gram-Schmidt (GS) spectral sharpening approach [5].

The GS fusion method consists of four steps: (1) synthesis of a low-resolution intensity band from the multispectral data; (2) Gram-Schmidt orthogonalization of the multispectral bands using this intensity band as the reference; (3) substitution of the orthogonalized intensity band with a high-resolution version; and (4) inversion of the Gram-Schmidt transformation to obtain the fused image.

Step 1 – Synthetic intensity band. We first construct a pseudo-RGB composite using three representative bands: red (R), green (G), and near-infrared (NIR). This R-G-NIR configuration was chosen because, in our preliminary experiments, it produced the most reliable results by combining visible light information (R, G) with vegetation-sensitive NIR data. The synthetic intensity band I_{syn} is computed as the average of these three bands:

$$I_{syn}(x, y) = \frac{1}{3} \left(\mathrm{R}(x, y) + \mathrm{G}(x, y) + \mathrm{NIR}(x, y) \right) \tag{5}$$

Step 2 – Gram-Schmidt orthogonalization. The n multispectral bands (originally 7: R, G, B, green, red, red edge, NIR) are orthogonalized with respect to I_{syn}. For each band B_i, the Gram-Schmidt process subtracts the component correlated with I_{syn}:

$$\mathrm{GS}_i = B_i - \frac{\langle B_i, I_{syn} \rangle}{\langle I_{syn}, I_{syn} \rangle} \cdot I_{syn} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. This produces a set of orthogonalized components $\{\mathrm{GS}_1, \ldots, \mathrm{GS}_n\}$ that are uncorrelated with I_{syn}.

Step 3 – Intensity substitution. The synthetic intensity band I_{syn} is replaced by a high-resolution intensity band. In our case, we use the same I_{syn} computed from the original R-G-NIR bands, as we lack a separate panchromatic band. This step preserves the high-resolution spatial structure of the pseudo-RGB composite while maintaining the spectral relationships captured by the orthogonalized components.

Step 4 – Inverse Gram-Schmidt transformation. The fused image is reconstructed by inverting the orthogonalization process. The new multispectral bands $\hat{B}_i$ are obtained as:

$$\hat{B}_i = \mathrm{GS}_i + \frac{\langle B_i, I_{syn} \rangle}{\langle I_{syn}, I_{syn} \rangle} \cdot I_{syn} \tag{7}$$

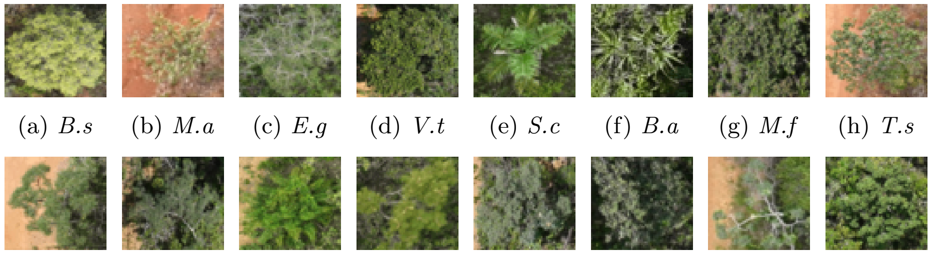
The final three-channel fused image is formed by selecting the bands corresponding to R, G, and NIR from the reconstructed $\hat{B}_i$, yielding a compact representation compatible with any standard RGB object detector.

3.5 Species Labeling and Class Distribution

Biologists from the Ecological Reserve identified 16 native Cerrado tree species with sufficient samples for supervised deep learning. All identifications were cross-validated by two specialists.

The dataset comprises 1,620 labeled tree images distributed across 16 species after preprocessing. Class frequencies range from 45 to 178 images per species (mean = 101.3, median = 94.5), reflecting natural ecological heterogeneity. Rare species such as *Butia archeri* presented particular challenges, with fewer than 60 samples. We divided the dataset into training (80%), validation (10%), and test (10%) sets using stratified sampling [17] to preserve class proportions due to the nature of the images, in which multiple species could coexist in the same picture. This stratification is summarized in Table 1.

Figure 2 shows representative RGB samples. The limited dataset size stems from field labeling difficulty in the IBGE reserve, where dense vegetation and closed canopies hinder access and visibility. Target species occur only in specific areas, further reducing labeled samples. To our knowledge, no prior *in situ* identification of Cerrado tree species from UAV imagery exists in the literature.



(a) *B.s* (b) *M.a* (c) *E.g* (d) *V.t* (e) *S.c* (f) *B.a* (g) *M.f* (h) *T.s*

(i) *S.a* (j) *P.p* (k) *S.f* (l) *M.c* (m) *D.mi* (n) *E.p* (o) *D.ma* (p) *C.b*

Fig. 2: Representative RGB images of the 16 native Cerrado tree species. Species abbreviations correspond to Table 1. Images were acquired at 40m altitude.

3.6 Object Detection Models

We evaluated eight model variants: YOLOv5n, YOLOv5s, YOLOv8n, YOLOv8s, YOLO11n, YOLO11s, RT-DETR-L, and RF-DETR. This selection spans three YOLO generations at two scales (nano and small), plus two transformer-based real-time detectors, enabling assessment of how architectural complexity interacts with multispectral input modalities for detecting Cerrado tree species under complex canopy conditions.

Each model was trained separately on four input modalities: (1) standard RGB, (2) fused pseudo-RGB (R-G-NIR via Gram-Schmidt), (3) single-channel NDVI, and (4) single-channel NDRE. For single-channel inputs, we replicated the channel three times to match each architecture's expected input dimensions.

The YOLO family (v5, v8, v11) was selected for its state-of-the-art inference speed, real-time detection capability, and widespread adoption in forestry monitoring. Real-Time Detection Transformer (RT-DETR) [10] was the first transformer-based object detector to achieve real-time performance with state-of-the-art accuracy, combining a CNN backbone with a transformer decoder (32.9M parameters). RF-DETR [15] is a real-time transformer architecture built on a DINOv2 vision transformer (ViT) backbone with 30.5M parameters.

3.7 Training Protocol

To evaluate model robustness and stability, each architecture was trained across different input modalities using three distinct random seeds. This resulted in four independent experimental setups per architecture: (i) standard RGB, (ii) the proposed fused imagery, (iii) isolated NDVI, and (iv) isolated NDRE.

Due to computational limitations all models were trained with an optimized input resolution of 640x640 pixels and a batch size of 12. YOLO models were trained for 100 epochs, and the transformers models for 50 epochs, as preliminary experiments showed convergence within 50 epochs for transformers. All models used the AdamW optimizer with an initial learning rate of 1e-3, weight decay of 5e-4, and a cosine annealing scheduler. YOLO models employed default augmentation parameters from Ultralytics YOLOv8. Transformer models used a batch size of 8 (due to memory constraints) with gradient accumulation to reach effective batch size of 12. Input resolution of 640×640 balances computational efficiency with spatial detail retention, as preliminary tests showed minimal gains above 640×640 with significant memory overhead. Batch size 12 was the maximum supported by the GPU for transformer models.

The dataset contains 324 original images, of which 256 were assigned to the training split. These 256 images contain a total of 1,071 labeled tree instances (bounding boxes) by which data augmentation techniquesexpanded the training set to 768 images while preserving the original label distribution. Each training image underwent random horizontal flipping, ±15° rotation, and random zoom crops between 0% and 15%. These augmentations tripled the training set from 256 to 768 images

Table 1: Labels stratification for each species.

Tree species	**# total**	**# training**	**# validation**	**# test**
Dalbergia miscolobium (D.m)	181	143	16	22
Tachigali subvelutina (T.s)	145	114	14	17
Caryocar brasiliense (C.b)	134	110	10	14
Blepharocalyx salicifolius (B.s)	131	106	15	10
Miconia ferruginata (M.f)	102	82	6	14
Vochysia thyrsoidea (V.t)	88	73	5	10
Syagrus comosa (S.c)	84	67	7	10
Mimosa claussenii (M.c)	84	61	10	13
Pterodon pubescens (P.p)	83	71	5	7
Didymopanax macrocarpus (D.m)	81	63	9	9
Eriotheca pubescens (E.p)	58	46	6	6
Syagrus flexuosa (S.f)	48	35	5	8
Stryphnodendron adstringens (S.a)	47	37	4	6
Enterolobium gummiferum (E.g)	35	27	5	3
Butia archeri (B.a)	33	20	7	6
Miconia albicans (M.a)	24	16	4	4
Total	**1.358**	**1.071**	**128**	**159**

All selected architectures were initialized with random weights to provide a consistent and unbiased baseline. By training these models from scratch, we aim to isolate the impact of the proposed multispectral fusion from any prior knowledge embedded in standard datasets, thereby evaluating the models' intrinsic capacity to process specialized environmental data.

3.8 Evaluation metrics

The evaluation of model performance requires two conditions for a correct prediction: the *Intersection over Union* (IoU), calculated as the ratio of intersection to union area between predicted and ground truth objects (equation 8), and the accurate classification of the predicted object. Upon meeting both conditions, the corresponding reference object is considered, and precision and recall are calculated (equations 9 and 10), where TP, FP, and FN represent true positives, false positives, and false negatives, respectively, based on the match to the ground truth and the correctness of the prediction.

$$IoU = \frac{\text{Intersection Area}}{\text{Union Area}} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

3.9 Calculation of mAP@50

The mean Average Precision (mAP) is a standard evaluation metric for object detection that summarizes the precision-recall curve for each class. For mAP@50, a predicted bounding box is considered correct if its IoU with a ground-truth box is at least 0.5. Each class's Average Precision (AP) is computed as the area under the precision-recall curve. Mathematically, AP for a given class can be approximated as:

$$AP = \int_0^1 p(r)\, dr \approx \sum_{n=1}^{N} p(r_n) \Delta r_n, \tag{11}$$

where $p(r)$ is the precision at recall r, N is the number of recall levels, and Δr_n represents the change in recall between successive points. The mAP@50 is then calculated by taking the mean of the AP values over all C classes:

$$mAP@50 = \frac{1}{C} \sum_{c=1}^{C} AP_c. \tag{12}$$

4 Results and Discussion

4.1 Inference Results

Table 2 summarises mAP@50 results for all model and modality combinations across three seeds. The fused modality consistently achieved the highest mAP@50 across all architectures, confirming that near-infrared information provides discriminative features not captured by visible-spectrum imagery alone. The best overall results were obtained by RT-DETR-L and YOLOv5s, both reaching 85.2% mAP@50 under the fused modality. Notably, YOLOv5s outperformed its more recent successors (YOLOv8s and YOLO11s) under the fused modality, suggesting that architectural complexity does not guarantee improved performance on specialized multispectral data.

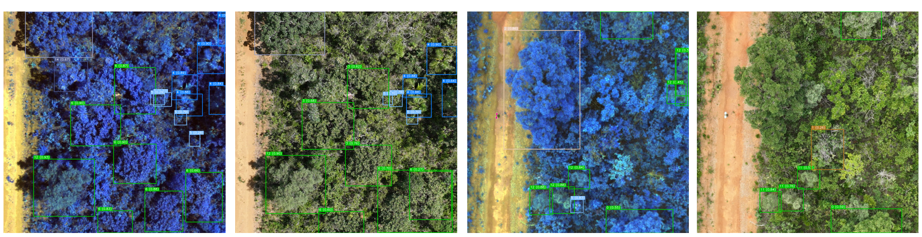
Despite sharing the top position, these models represent markedly different trade-offs: YOLOv5s with its 7.5M parameters is a lightweight, single-stage detector suitable for resource-constrained environments, while RT-DETR-L is a transformer-based architecture with 32.9M parameters and substantially higher computational cost. Depending on operational requirements, we believe that YOLOv5s may be preferred without sacrificing detection performance.

The RF-DETR Nano, trained under the same from-scratch protocol, achieved substantially lower mAP@50 across all modalities (21.0% for fusion), revealing a fundamental difference between CNN-based and Vision Transformer-based architectures. Unlike YOLO and RT-DETR-L, whose CNN backbones learn effective representations from limited data, RF-DETR's DINOv2 ViT backbone was designed for transfer learning. Without pretrained weights, its large parameter count cannot be adequately optimized on the small Cerrado dataset, resulting in underfitting. We intentionally trained all models from random weights to ensure fair comparison; future work should explore fine-tuning RF-DETR with remote sensing pretraining.

Table 2: mAP@50 (%) per model, reported as mean ± std over training seeds. **Bold** indicates the best modality per model.

Input	**v5n**	**v5s**	**v8n**	**v8s**	**v11n**	**v11s**	**rtdetr-l**	**rfdetr-n**
Fused	**76.7±2.2**	**85.2±1.6**	**80.0±1.5**	**83.8±1.1**	**81.3±2.4**	**84.8±2.6**	**85.2±1.1**	**21.0±1.5**
RGB	62.3±4.7	66.8±4.5	64.8±2.3	62.2±3.3	63.8±4.0	77.4±2.0	46.2±2.8	20.5±1.2
NDRE	49.9±3.2	59.0±5.5	54.2±2.2	59.3±2.5	49.5±5.1	58.6±0.7	51.7±4.5	6.3±1.3
NDVI	55.8±0.7	63.7±3.0	46.0±13.6	64.0±2.6	59.2±4.4	59.0±2.7	46.9±1.5	11.8±0.9

RGB-only models achieved intermediate performance, ranking second in most architectures. Although visible-spectrum imagery is considerably easier and cheaper to acquire than multispectral data, the consistent gap relative to the fused modality averaging 19% across all models demonstrates the added value of near-infrared information for fine-grained species discrimination in the Cerrado. These results indicate that the proposed fusion not only achieves higher accuracy but also greater training robustness in low-data scenarios. Fig. 3 illustrates representative detections under both modalities.



(a) Fused (b) RGB (c) Fused (d) RGB

Fig. 3: Sample detections of native Cerrado tree species in RTDETR-L.

Collectively, single-index modalities (NDRE, NDVI) and RGB-only imagery proved insufficient for reliable species monitoring in the Cerrado, an ecosystem characterised by high floristic diversity and low inter-species spectral separability. The proposed fusion consistently emerged as the most effective strategy, suggesting that near-infrared reflectance captures discriminative information that neither visible colour nor narrowband vegetation indices alone can provide. Future work should investigate whether this fusion can serve as a practical substitute for dedicated vegetation indices in applications such as vegetation health assessment, biomass estimation, phenological monitoring, and stress detection — potentially simplifying data acquisition pipelines without sacrificing analytical capacity.

4.2 Per-species Performance

To further investigate the model behaviour, Table 3 presents the RT-DETR-L architecture, selected for its state-of-the-art performance and low standard deviation between the seeds, in a per-species analysis. The analysis exposes two structural challenges inherent to automated biodiversity monitoring: long-tail data distributions and low inter-species visual separability. The proposed spectral fusion substantially mitigates both, as detailed below.

Table 3: Per-class metrics for RT-DETR-L on *fused* images, averaged over three seeds. Best per metric in **bold**.

Species	Precision (%)	Recall (%)	F1-score (%)	mAP@50 (%)
E. gum.	80.1	**100.0**	88.8	**99.5**
M. fer.	97.5	92.8	**95.0**	99.0
P. pub.	85.8	87.4	86.5	96.9
S. com.	77.0	**100.0**	86.9	94.1
V. thy.	93.9	83.3	88.3	93.4
S. ads.	**99.8**	76.2	86.3	92.2
D. mis.	96.5	80.4	87.6	89.7
B. sal.	85.2	76.4	80.5	83.8
E. pub.	77.9	83.3	80.4	83.7
M. cla.	97.3	63.3	76.3	80.6
T. sub.	90.0	75.0	81.7	80.6
C. bra.	88.9	77.9	83.0	78.2
D. mac.	89.4	63.8	74.2	75.9
M. alb.	74.4	66.7	70.1	75.2
S. fle.	86.8	61.1	71.4	73.8
B. arc.	91.7	63.8	74.8	66.7

Our dataset aims to reflects the natural imbalance between trees species quantity, with instance counts ranging from 24 (*Miconia albicans*) to 181 (*Dalbergia miscolobium*). Despite this constraint, 12 of 16 species surpass 75% mAP@50 under the fused modality, and five exceed 90%. This robustness under data scarcity is largely enabled by the NIR channel: comparing RT-DETR-L under fused and RGB-only inputs as show in Table 4 reveals that the gains are most pronounced precisely for the rarest and most visually ambiguous species. *Mimosa claussenii* improves from 2.4% to 80.6%, *Stryphnodendron adstringens* from 8.4% to 92.2%, and *Tachigali subvelutina* from 6.0% to 80.6%, gains exceeding 70 percentage points — improvements that would be unattainable. The perfect recall for *E. gummiferum* and *S. comosa* should be interpreted with caution given their small test set sizes (3 and 14 samples, respectively). These results confirm that near-infrared reflectance encodes discriminative spectral signatures that the visible spectrum simply cannot provide for fine-grained Cerrado species identification.

The three lowest-performing species under the fused modality share a combination of data scarcity and morphological ambiguity that the fusion alone cannot fully resolve: *Butia archeri* (66.7%), *Syagrus flexuosa* (73.8%), and *Miconia albicans* (75.2%). *B. archeri*, *S. flexuosa*, and *S. comosa* are all Cerrado palms with similar pinnate crown architecture, a confusion documented in botanical literature: *B. archeri* was historically classified within *Syagrus* due to their morphological resemblance [3], and *S. flexuosa* is reported to be misidentified as *S. comosa* by taxonomic specialists [12]. Our findings suggest that NIR reflectance captures phenotypic differences that even specialist botanists struggle to distinguish visually.

For two species (*B. salicifolius* and *D. macrocarpus*), RGB-only surprisingly outperformed the fused modality. Both species have distinctive bark and leaf textures that may be more discriminative in visible wavelengths, suggesting that optimal modality selection may be species-dependent. Yet even for these challenging cases, the fused modality delivers meaningful improvements over RGB: *B. archeri* rises from 16.0% to 66.7% and *S. flexuosa* from 36.2% to 73.8% suggesting that NIR partially captures structural differences invisible to the human eye.

Table 4: Per-species mAP@50 (%) for RT-DETR-L under fused and RGB, averaged over three seeds. Best modality per species in **bold**.

Species	**Fused**	**RGB**	***Δ***
E. gum.	**99.5**	83.3	+16.2
M. fer.	**99.0**	31.5	+67.5
P. pub.	**96.9**	14.6	+82.3
S. com.	94.1	**94.9**	–0.8
V. thy.	**93.4**	85.9	+7.5
S. ads.	**92.2**	8.4	+83.8
D. mis.	**89.7**	18.3	+71.4
B. sal.	83.8	**96.0**	–12.2
E. pub.	**83.7**	69.9	+13.8
M. cla.	**80.6**	2.4	+78.2
T. sub.	**80.6**	6.0	+74.6
C. bra.	**78.2**	76.5	+1.7
D. mac.	75.9	**84.7**	–8.8
M. alb.	**75.2**	15.0	+60.2
S. fle.	**73.8**	36.2	+37.6
B. arc.	**66.7**	16.0	+50.7

The Precision-Recall profiles in Fig. 4 further reveal that detection errors carry asymmetric conservation consequences. Species in the high-precision, low-recall quadrant - *S. adstringens* (P = 99.8%, R = 76.2%) and *M. claussenii* (P = 97.3%, R = 63.3%) - are systematically undercounted despite being reliably identified when detected. In conservation monitoring, systematic undercounting of rare taxa translates directly into underestimated population sizes and potentially flawed management decisions. Adopting species-specific confidence thresholds in operational deployments, rather than a single global threshold, could rebalance this trade-off in favour of recall for the species that matter most.

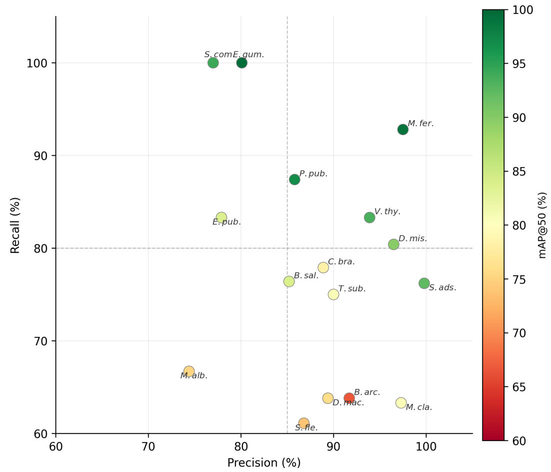


Fig. 4: Precision–Recall profile of RT-DETR-L under fused modality. The dashed diagonal indicates F1-score = 0.8. Species in the high-precision, low-recall quadrant (shaded) are systematically undercounted, with implications for conservation monitoring.

5 Conclusion

This work presented an approach combining multichannel UAV imagery with deep learning to address the urgent need for automated biodiversity monitoring in the Brazilian Cerrado. Our preprocessing pipeline, employing Gram-Schmidt fusion, condenses RGB and multispectral bands (green, red, red edge, NIR) into a three-channel pseudo-RGB format, enabling state-of-the-art object detection models to exploit near-infrared information without architectural modifications. The dataset collected within the IBGE Ecological Reserve, although limited in size, is one of the first of its kind, with all annotations certified by field biologists.

Across eight model architectures, the fused modality consistently outperformed RGB-only and single-index (NDVI, NDRE) inputs. RT-DETR-L and YOLOv5s achieved the highest mAP@50 of 85.2%, with the latter offering a compelling accuracy-efficiency trade-off (7.5M parameters vs. 32.9M for RT-DETR-L). Per-species results further showed that spectral fusion yields its largest gains for the rarest and most morphologically ambiguous taxa, demonstrating that near-infrared reflectance captures discriminative information that the visible spectrum cannot provide — with gains exceeding 70 percentage points for several species. Vegetation indices alone proved insufficient for reliable species discrimination, reinforcing that spectral fusion, not single-band indices, is required for fine-grained identification in biodiverse ecosystems.

Several limitations should be acknowledged. Our dataset is limited to a single site (IBGE Ecological Reserve) and a narrow seasonal window, leaving generalization to other Cerrado regions and phenological stages as open questions. All flights were conducted under clear sky conditions; performance under variable illumination remains untested. Additionally, models were trained from random weights; transfer learning from remote sensing pretraining may further improve performance, particularly for transformer-based architectures.

Future work should prioritize: (i) expansion of the dataset to additional Cerrado sites and seasonal cycles; (ii) evaluation of alternative fusion methods (e.g., attention-based or learnable spectral weighting); (iii) temporal analysis using time-series UAV data to capture phenological variation; and (iv) operational deployment of YOLOv5s with species-specific confidence thresholds, as suggested in this research.

With nearly 50% of the Cerrado's original vegetation already lost, scalable monitoring tools are not merely academic exercises but conservation imperatives. This work establishes multispectral fusion as a practical and accessible foundation for fine-grained tree species detection in one of the world's most threatened savannas.

Acknowledgments. CNPq (Brazilian Research Council) partially funded this study. We thank Isaac A. Silva (UnB) and the staff of the IBGE Ecological Reserve for field support, especially Marina L. F. Resende and Betania T.P.M. Góes.

References

1. Alencar, A., Z. Shimbo, J., Lenti, F., et al.: Mapping three decades of changes in the brazilian savanna native vegetation using landsat data processed in the google earth engine platform. Remote Sensing **12**(6) (2020), https://www.mdpi.com/2072-4292/12/6/924
2. Barnes, E., Clarke, T., Richards, S., Colaizzi, P., Haberland, J., Kostrzewski, M., Waller, P., Choi, C., Riley, E., Thompson, T., et al.: Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In: Proceedings of the fifth international conference on precision agriculture, Bloomington, MN, USA. vol. 1619 (2000)
3. Carvalho, L.F.L., Francino, D.M.T., Nunes, E.L.P., Soffiatti, P., Sant'Anna-Santos, B.F.: Splitting *Butia archeri* (Arecaceae) of Brazilian central highlands: new taxonomical and conservation insights for *Butia*. Plant Systematics and Evolution **310**, 24 (2024). https://doi.org/10.1007/s00606-024-01908-z
4. da Costa, F.G., da Costa, V.S., da Silva Martins, I., et al.: Knowing cerrado: field classes and their importance for environmental knowledge and preservation. Research, Society and Development **9**(10), e1589108201 (Sep 2020). https://doi.org/10.33448/rsd-v9i10.8201
5. Dalla Mura, M., Vivone, G., Restaino, R., Addesso, P., Chanussot, J.: Global and local gram-schmidt methods for hyperspectral pansharpening. In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 37–40. IEEE (2015)
6. DJI Corp: Mavic 3m: Multispectral surveying drone. https://ag.dji.com/pt-br/mavic-3-m (2023)
7. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization **30**(10) (2008)

8. Fassnacht, F.E., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L.T., Straub, C., Ghosh, A.: Review of studies on tree species classification from remotely sensed data. Remote Sensing of Environment **186**, 64–87 (2016). https://doi.org/10.1016/j.rse.2016.08.013
9. Karami, E., Prasad, S., Shehata, M.: Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images. arXiv preprint arXiv:1710.02726 (2017). https://doi.org/10.48550/arXiv.1710.02726
10. Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., Liu, Y.: RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. arXiv preprint arXiv:2407.17140 (2024). https://doi.org/10.48550/arXiv.2407.17140
11. Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J.: Biodiversity hotspots for conservation priorities. Nature **403**(6772), 853–858 (feb 2000). https://doi.org/10.1038/35002501
12. Noblick, L.R.: *Syagrus cearensis*, a twin-stemmed new palm from Brazil. Palms **48**, 109–176 (2004)
13. Ouattara, T.A., Sokeng, V.C.J., Zo-Bi, I.C., Kouamé, K.F., Grinand, C., Vaudry, R.: Detection of forest tree losses in côte d'ivoire using drone aerial images. Drones **6**(4) (2022). https://doi.org/10.3390/drones6040083
14. Park, H.G., Yun, J.P., Kim, M.Y., et al.: Multichannel object detection for detecting suspected trees with pine wilt disease using multispectral drone imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 8350–8358 (2021). https://doi.org/10.1109/JSTARS.2021.3102218
15. Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., Peri, N.: Rf-detr: Neural architecture search for real-time detection transformers (2025), https://arxiv.org/abs/2511.09554
16. Rouse, J. W., J., Haas, R.H., Schell, J.A., Deering, D.W.: Monitoring vegetation systems in the Great Plains with ERTS. In: Third Earth Resources Technology Satellite-1 Symposium, Volume 1: Technical Presentations, Section A. pp. 309–317. NASA SP-351, NASA (1974), https://ntrs.nasa.gov/citations/19740022614
17. Sechidis, K., Tsoumakas, G., Vlahavas, I.: On the stratification of multi-label data. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) Machine Learning and Knowledge Discovery in Databases. pp. 145–158. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
18. Sims, D.A., Gamon, J.A.: Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sensing of Environment **81**(2), 337–354 (2002). https://doi.org/10.1016/S0034-4257(02)00010-X
19. Ventura, J., Pawlak, C., Honsberger, M., et al.: Individual tree detection in large-scale urban environments using high-resolution multispectral imagery. International Journal of Applied Earth Observation and Geoinformation **130**, 103848 (2024). https://doi.org/10.1016/j.jag.2024.103848
20. Yu, X., Hyyppä, J., Litkey, P., et al.: Single-sensor solution to tree species classification using multispectral airborne laser scanning. Remote Sensing **9**(2) (2017). https://doi.org/10.3390/rs9020108
21. Zhao, M.H., Li, X., Zhang, X.X., Zhang, H., Zhao, X.Y.: Mutation mechanism of leaf color in plants: A review. Forests **11**(8) (2020). https://doi.org/10.3390/f11080851
22. Zheng, Y., Yu, J., Kang, S.B., Lin, S., Kambhamettu, C.: Single-image vignetting correction using radial gradient symmetry. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587413