Improving Small Drone Detection Through Multi-Scale Processing and Data Augmentation

Rayson Laroca[1,2], Marcelo dos Santos[2], and David Menotti[2]

[1]Pontifical Catholic University of Paraná, Curitiba, Brazil
[2]Federal University of Paraná, Curitiba, Brazil

8th WOSDETC Drone-vs-Bird Detection Data Competition @IJCNN25

UNIVERSITÀ DEL SALENTO
L'Ateneo tra i due mari

Fraunhofer
IOSB

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Datasets – DDS (Official Challenge Dataset)

- **77 videos**, with high variability in several aspects:

Datasets – DDS (Official Challenge Dataset)

- **How should the 77 videos be optimally split into training and validation subsets?**

Datasets – DDS (Official Challenge Dataset)

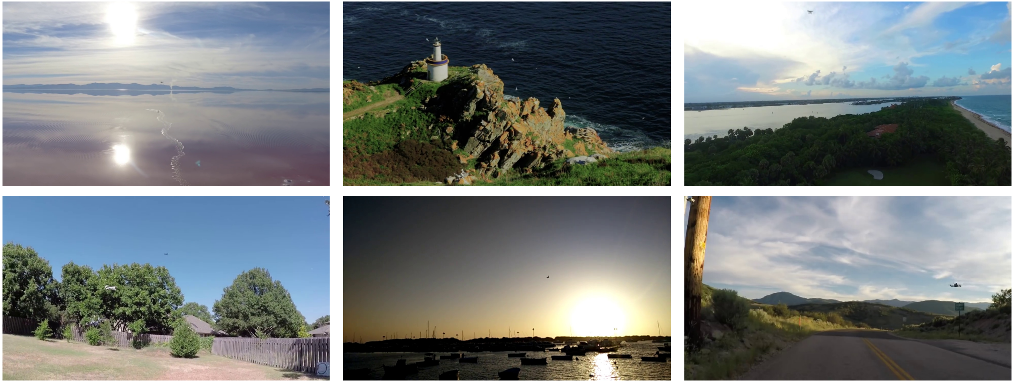
- **How should the 77 videos be optimally split into training and validation subsets?**
  1. We allocated **70 videos for training** and **7 for validation**;
  2. We **manually curated the dataset** to ensure the validation set included videos with diverse drone distances, image resolutions, backgrounds, and environmental conditions;
  3. We performed **preliminary experiments** with different object detectors;
  4. We implemented an **iterative refinement process**:
     - This involved strategically swapping videos between the training and validation sets, guided by empirical observations of model performance while maintaining the intended diversity of the validation set.

Datasets

- **Three additional datasets**:
  - USC-GRAD-STDdb
  - “Dataset2”
  - DUT Anti-UAV
- While these datasets include their own subdivisions for training, testing, and validation, **we utilized all images for <u>training</u> our model**.
  - The validation set consisted exclusively of scenarios from the target competition.

Datasets – USC-GRAD-STDdb

- 115 videos collected from YouTube (25k+ frames)
  - We explored the **2,263 frames** that feature **drones** or **birds**.
  - **Small object instances**, ranging from $\approx 4 \times 4$ to $\approx 16 \times 16$ pixels.

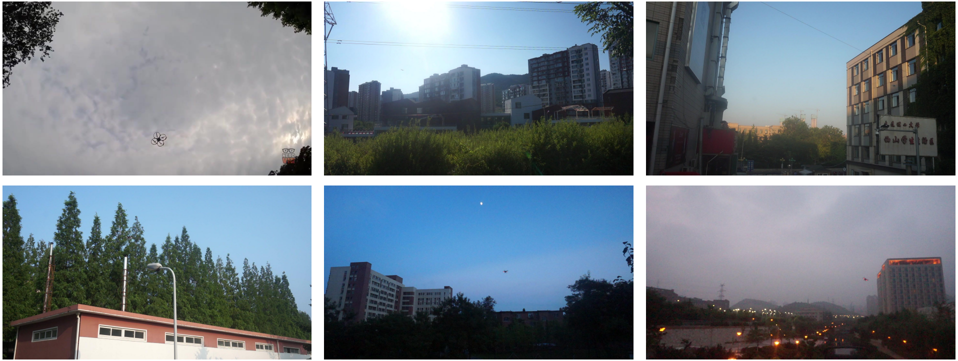


Datasets – Dataset2

- 51 videos depicting birds and 114 videos featuring drones
  - All selected videos have a resolution of **640** $\times$ **512 pixels**.
  - We kept every tenth frame from the videos, resulting in **4,516 frames**.

Datasets – DUT Anti-UAV

- 10,000 images with 10,109 manually annotated drone positions.
  - **High variability**: resolutions from 240×160 to 5616×3744 pixels, 35+ drone models, and a wide array of backgrounds (sky, clouds, jungles, urban landscapes, farmland, and playgrounds).



Proposed Approach

We started our experiments using the **YOLO11m** model as a baseline.

Proposed Approach

We started our experiments using the **YOLO11m** model as a baseline.

- **1st major challenge** $\rightarrow$ detecting distant drones.
    - When processing high-resolution images, such as 4K or even Quad HD, and resizing them to the model's 640-pixel input, **distant objects became very small**, often approaching or falling below the detection threshold.

Proposed Approach

We started our experiments using the **YOLO11m** model as a baseline.

- **1st major challenge** $\rightarrow$ detecting distant drones.
    - When processing high-resolution images, such as 4K or even Quad HD, and resizing them to the model's 640-pixel input, **distant objects became very small**, often approaching or falling below the detection threshold.
    - We investigated several alternative models. For example:
        - **YOLO11m-p2**, a specialized variant with a finer stride configuration optimized for small object detection;
        - **YOLOv11m** with **doubled input size** (1280×1280 instead of 640×640);

Proposed Approach

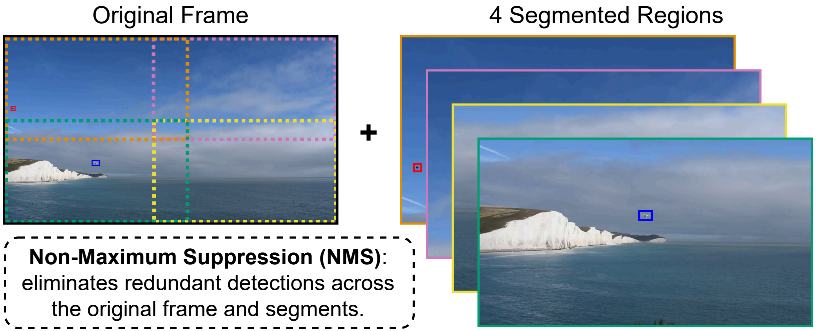
We started our experiments using the **YOLO11m** model as a baseline.

- **1st major challenge** $\rightarrow$ detecting distant drones.
  - When processing high-resolution images, such as 4K or even Quad HD, and resizing them to the model's 640-pixel input, **distant objects became very small**, often approaching or falling below the detection threshold.
  - We investigated several alternative models. For example:
    - **YOLO11m-p2**, a specialized variant with a finer stride configuration optimized for small object detection;
    - **YOLOv11m** with **doubled input size** (1280×1280 instead of 640×640);
  - However, the [relatively small] gains did not justify the added computational cost.

Proposed Approach – Stage 1 – Multi-Scale Processing

Stage 1

Multi-Scale Processing (YOLOv11)



Non-Maximum Suppression (NMS): eliminates redundant detections across the original frame and segments.

Each segment, covering 55% of the original frame's width and height, simulates a **zoom effect** as the model input size (640×640 pixels) remains unchanged, resulting in less downscaling.

Proposed Approach – Drone vs. Bird

The results for distant objects improved considerably, but...

- **2nd major challenge** $\rightarrow$ distinguishing drones from birds.

Proposed Approach – Drone vs. Bird

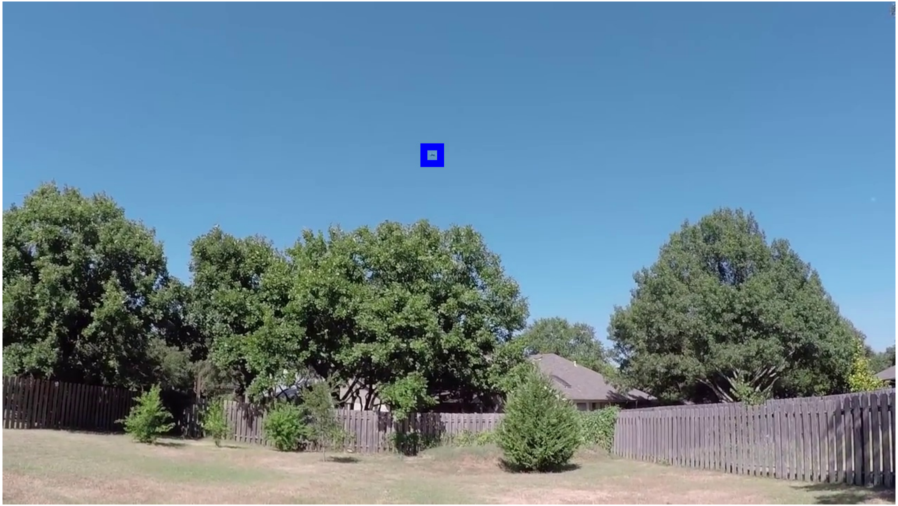
The results for distant objects improved considerably, but...

- **2nd major challenge** $\rightarrow$ distinguishing drones from birds.
  - We leveraged **a copy-paste data augmentation technique** to improve the training set with additional drone and bird instances.
    - This involved randomly placing cropped and scaled instances into new locations, **ensuring they did not overlap with existing instances**.
    - The images used for pasting were collected from both **the training set and various online sources**, all featuring transparent background.

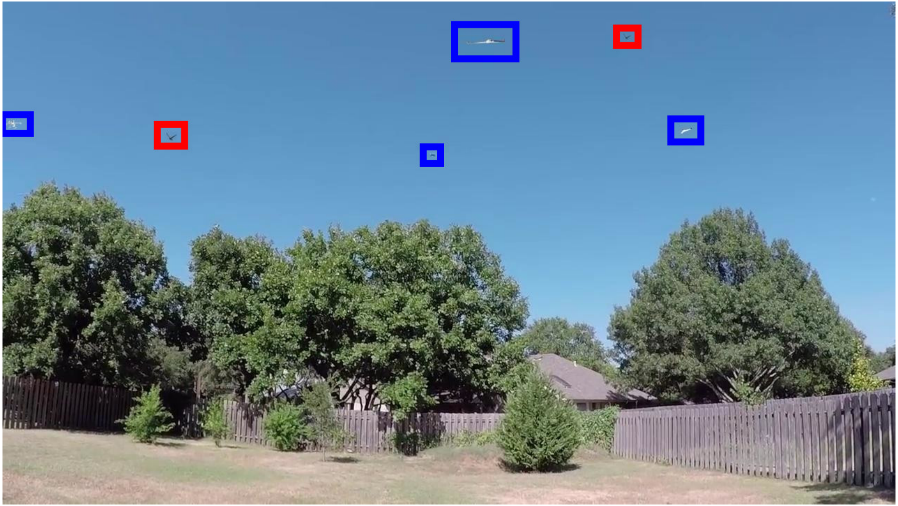



Proposed Approach – Data Augmentation

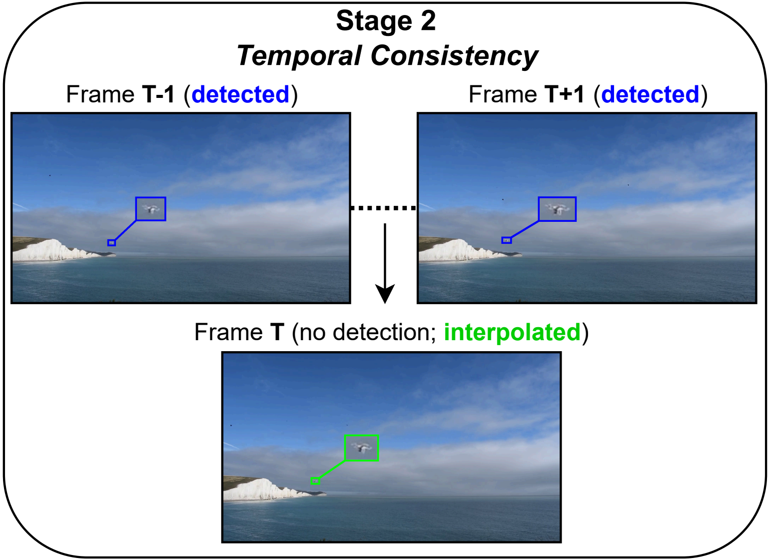
- **Original:**



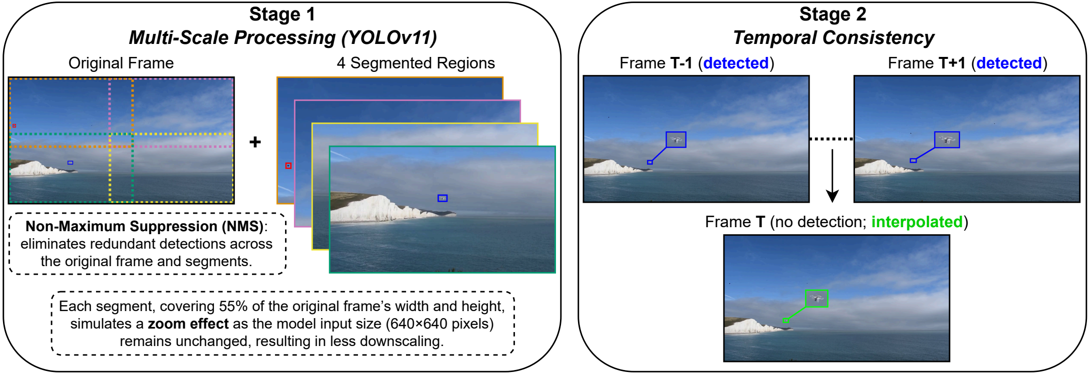
- **Augmented:**



Proposed Approach – Stage 2



Proposed Approach



Results (on the validation set)

Video Name	mAP_{50}	mAP_{50} †
dji_mavick_mountain	**0.9891**	0.6431
2019_10_16_C0003_3633_inspire	**0.9421**	0.9219
parrot_disco_distant_cross_3	**0.8684**	0.5550
GOPR5843_002	**0.7175**	0.3371
swarm_dji_phantom4_2	**0.7077**	0.6566
dji_phantom_4_hillside_cross	0.4992	**0.7406**
gopro_002	**0.4491**	0.0121
Average	0.7390	0.5523

† using the simplified variant that processes only whole images.

Results (Qualitative)

Results (Qualitative)

Conclusions

- We utilized the medium-sized **YOLOv11** model for drone detection.
  - The input image is processed both **as a whole and in segmented parts**;
  - This strategy boosted detection performance, especially for **distant drones**.
- We employed extensive **data augmentation**.
  - A **copy-paste technique** to increase the number of drone and bird instances in the training images.
- A **post-processing stage** was incorporated to mitigate missed detections by leveraging **temporal information**.

INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS
IJCNN2025
30 JUNE - 5 JULY 2025 | ROME, ITALY
INTERNATIONAL NEURAL NETWORK SOCIETY

Thank you!

https://raysonlaroca.github.io/supp/drone-vs-bird/

CAPES
ARAUCARIA
CIÊNCIA, TECNOLOGIA E INOVAÇÃO PARA O PARANÁ
PARANÁ
GOVERNO DO ESTADO
SECRETARIA DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
CNPq
PUCPR
GRUPO MARISTA
UFPR
UNIVERSIDADE FEDERAL DO PARANÁ
NVIDIA