Nothing to See Here: Background Data and False Positives in UAV Imagery Object Detection

Sarah Williams, Harvey Mudd College

Mentors: Duane Edgington, Danelle Cline

Summer 2026

Keywords: object detection, tiled inference, detection transformers, drones

ABSTRACT

Uncrewed aerial vehicles (UAVs) provide an efficient, cost-effective way to survey a wide range of marine organisms on the ocean's surface, such as jellies, kelp, and marine mammals. However, the large images collected by UAVs are typically cropped tightly around annotated objects before training, leaving object detectors with little exposure to background-only crops. We hypothesized this training/inference mismatch drives increased rates of false positives at inference. We swept the percentage of background-only image crops (0%, 5%, 10%, and 20%) included in RF-DETR object detector training. Increasing background augmentation from 0% to 20% reduced the number of false positives by approximately an order of magnitude across all confidence intervals. However, the effects of increasing background augmentation almost immediately saturated at 5% background-only crops and the trade-off with recall was increasingly less optimal as the percentage of background augmentation was increased. Comparing the performance of a YOLO26 model to the RF-DETR models, we found that YOLO26 is more effective at suppressing false positives but generally has lower recall than RF-DETR.

INTRODUCTION

Uncrewed aerial vehicles (UAVs) have been increasingly used for environmental and population monitoring in place of traditional in situ survey methods, offering greater efficiency, precision, and spatial coverage while reducing habitat disturbance (Hodgson et al., 2016). Paired with deep learning methods, UAVs have been used for a range of marine monitoring tasks, such as detecting sea cucumbers on coral reef flats (Li et al., 2021) and identifying marine litter (Bartolo et al., 2026).

Across these applications, object detectors are trained on images cropped around annotated objects rather than full-size UAV images. However, when images are large with sparse objects, object detectors may misfire on the background of the image if their training data does not include negative examples, i.e. background-only cropped images. Such false positive rates are commonly mitigated with background augmentation in the training data but there is little guidance on how much background data should be included in the training dataset. For example, some model frameworks suggest 0-10%, while the COCO (Common Objects in COntext) dataset contains only approximately 1% background-only images, but these suggestions are only common practices rather than evidence-based findings (Ultralytics 2026). Additionally, our UAV images of the ocean's surface contain substantially more background (i.e. empty water) than common benchmarks. Thus, we quantified the effects of training RF-DETR models (Robinson et al. 2026) with 0%, 5%, 10%, and 20% background-only cropped images. We also compare the performance of a YOLO26 Medium model (Jocher et al. 2026) without any background augmentation to the RF-DETR models.

MATERIALS AND METHODS

DATA COLLECTION

MBARI has conducted sustained aerial surveys of the ocean's surface at six sites around Monterey Bay (Fig 1.1). Ocean surface imagery is collected with the Trinity Pro at approximately 60m at times chosen to reduce environmental interference, such as sun glint and water reflections. The flight paths cover each site while staying in the line of

sight of the UAV operator. Images captured by the Trinity Pro are 42-megapixel 8k 7952 × 5304 px (Fig 1.1).



Fig 1.1 Left: Monterey Bay with six flight sites marked with red pins. Right: Example of one of the Trinity Pro's flight paths at one of the six survey sites.

DATASET AUGMENTATION & CURATION

Data augmentation includes locating the annotated objects in the images, then cropping around those objects to the specified dimensions and specified overlap (Fig 1.2). For training RF-DETR Medium and YOLO26 models, the image crops are 576 px or 640 px, respectively, with 50% overlap. For the models whose training data included background augmentation, additional crops are added to the dataset where the percent of background-only crops is at most the specified percentage (e.g. 5%, 10%, 20%). These background-only crops were chosen randomly from the full-size images used to create the training dataset.

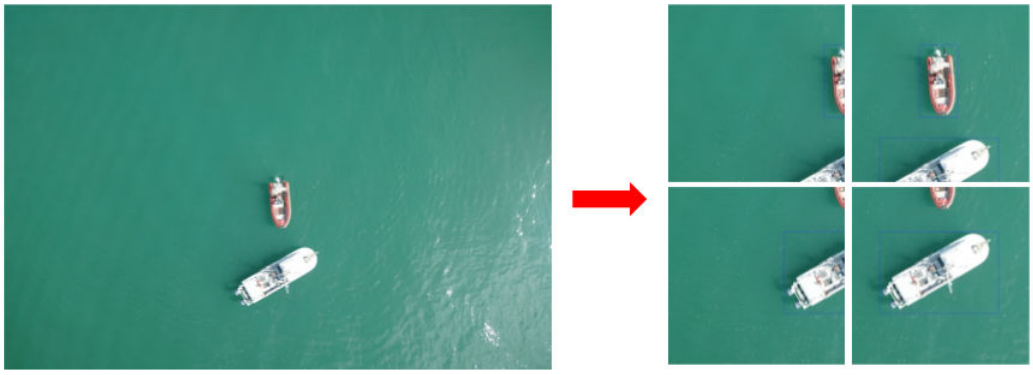


Fig 1.2 Left: original image taken by the Trinity. Right: Image crops used for training object detectors.

After discovering that the existing internal test set was under-annotated, we used a preliminary object detector to screen the images in the test set for annotation completeness, followed by a human reviewer checking that each preserved image was correctly and completely annotated. The resulting benchmark contains 98 images, all 7952 × 5304 px, and 523 boxes of a single object class.

MODEL TRAINING

All RF-DETR models (Medium, 576 px image input) were trained on the NVIDIA DGX Spark for 20 epochs on identical training data, except for the amount of background-only crops. All RF-DETR models were trained on the same 93,425 object crops, augmented with 0, 4,917, 10,380, and 23,356 background-only crops so that background forms at most 0%, 5%, 10% or 20% of the training set.

The YOLO26 model (Medium) was trained in Google Colab on an A100 GPU for a total of 300 epochs across two training sessions (150 epochs per session). The training data for the YOLO26 model was cropped from the same images as the RF-DETR training datasets, though object crops were cropped to 640 px (native YOLO input) and no background crops were included.

INFERENCE PIPELINES

For our inference pipeline, we combined ONNX Runtime with SAHI (Slicing Aided Hyper Inference). Thus, during inference, the object detectors were not given full-size UAV images, but smaller tiles (matching the size of the images the models were trained on) for better performance (Akyon et al. 2022).

METRICS

While Intersection over Union (IoU) and mean average precision (mAP) are commonly used to quantify a model's performance, we observed that such metrics did not accurately represent our model's performance. In our data, tight boxes are drawn around tiny objects, meaning that small deviations in predicted box position or size, within the range of ordinary annotation noise, can shift IoU from a confident match to zero overlap, causing correct detections to be labeled as false positives or false negatives under a fixed IoU threshold. Because mAP is computed from IoU-thresholded matches, the sensitivity of IoU impacts mAP as well, and object detector performance is evaluated to be artificially poor on small objects. Instead, we used Normalized Wasserstein Distance (NWD) to evaluate model performance. NWD models each bounding box as a 2D Gaussian distribution and computes similarity via the Wasserstein distance between predicted and ground-truth distributions rather than requiring pixel overlap, and thus is a more robust measure that does not collapse to zero for near-miss detections on small objects (Wang et al., 2021; Xu et al., 2022).

RESULTS

ADDING BACKGROUND CROPS SUPPRESSES FALSE POSITIVES

The inclusion of any proportion of background images to the training dataset suppresses false positives in RF-DETR object detectors. Across all confidence thresholds, we observed a decrease in false positives by as much as an order of magnitude, as a result of background augmentation (Fig 2a). However, we observed that the positive impacts of background augmentation quickly saturated at 5% background-crops. The F1 score peaks

at 5% background (Fig 2b), indicating that precision and recall are most balanced by the 5% background-only crops model. Similarly, the precision-recall curve of the 5% background-only crops model bows more towards the top-right corner of the plot, compared to the other models, demonstrating better model performance (Fig 2c).

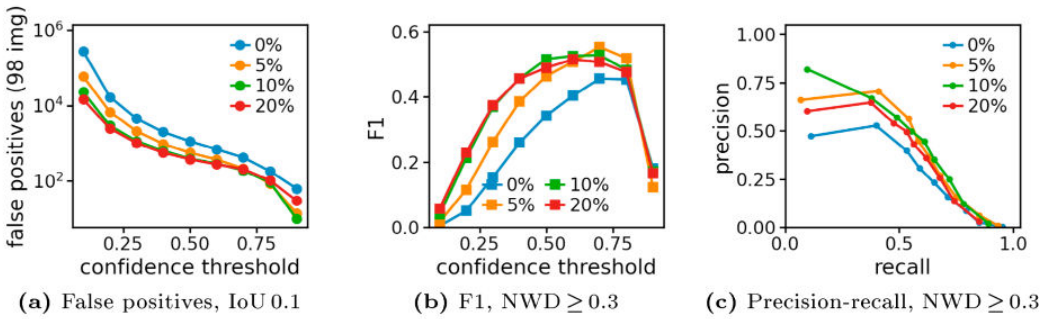


Fig 2. Performance of RF-DETR models with 0%, 5%, 10%, and 20% background crops shown in blue, orange, green, and red, respectively. (a) Number of false positives of each RF-DETR model when evaluated against the 98 image benchmark across varying confidence thresholds. (b) F1 score of each RF-DETR model across varying confidence thresholds with NWD ≥ 0.3. (c) For each of the RF-DETR models, precision versus recall is shown, for NWD ≥ 0.3.

DOES IT WORK IN PRODUCTION? QUALITATIVE EVALUATION

We visualized the performance of the RF-DETR models with 5%, 10% and 20% background crops on a singular image from the test set (Fig 3). The image in this example is primarily background (empty water) and contains one small object (a buoy). Increasing the proportion of background images in training data steadily improves model performance on this image; with 20% background augmentation, the model prediction matches the ground-truth annotation exactly.

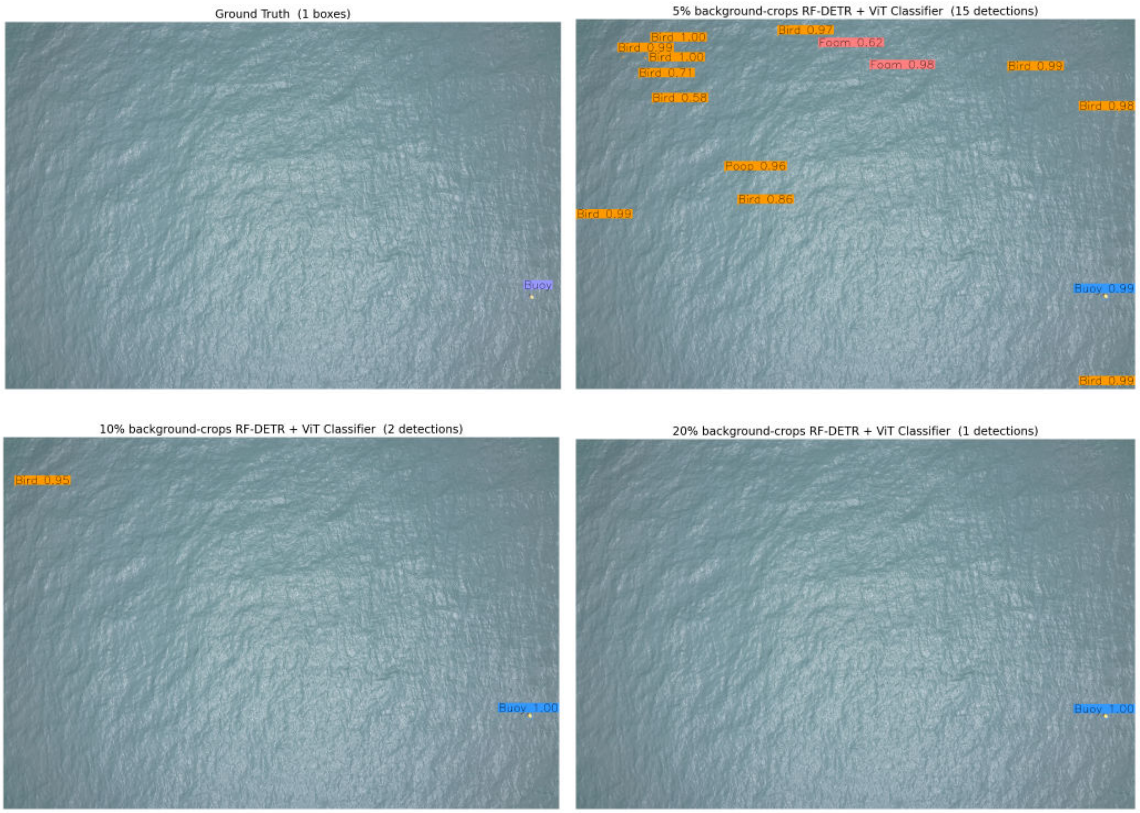


Fig 3. Ground-truth annotation and model predictions (5%, 10%, 20%) on the same image taken by the Trinity Pro, depicting one buoy amidst empty water. Top-left: Ground-truth annotation labels “Buoy”; top-right: inference pipeline with 5% background-crops RF-DETR + Vision Transformer (ViT) predictions; bottom-left: inference pipeline with 10% background-crops RF-DETR + Vision Transformer (ViT) predictions; bottom-right: inference pipeline with 20% background-crops RF-DETR + Vision Transformer (ViT) prediction.

RF-DETR v. YOLO26 COMPARISON

We compared all RF-DETR models (0%, 5%, 10%, and 20% background-only crops) to one YOLO26 model (0% background-only crops) to explore if one model performs better than the other on our data. YOLO26, with no background augmentation, consistently suppressed false positives more effectively than RF-DETR across all confidence thresholds (Fig 4). However, when comparing the precision-recall curves of the five models, we observed that YOLO26’s recall drops off rapidly compared to the RF-DETR models (Fig 4).

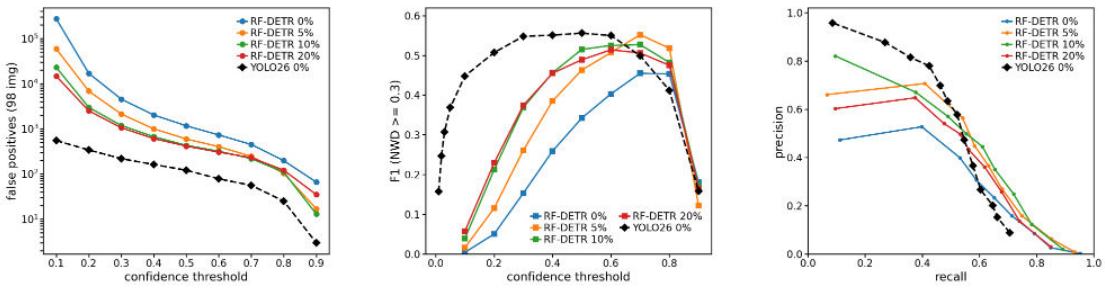


Fig 4. From left to right: false positives (y-axis) and confidence threshold (x-axis) for all RF-DETR and YOLO26 models with 0%, 5%, 10%, and 20% RF-DETR models shown in blue, orange, green and red, respectively, and YOLO26 (no background crops) shown in black; F1 score (y-axis) and confidence threshold (x-axis) shown for all RF-DETR and YOLO26 models; precision-recall curves for all RF-DETR and YOLO26 models.

DISCUSSION

Evaluating the four RF-DETR models' (0%, 5%, 10%, and 20% background crops) performance against our curated testing benchmark showed that the introduction of any amount of background-only crops during training suppressed the number of false positives during inference. However, this effect saturates almost immediately; in other words, continuing to increase the proportion of background crops in training data does not necessarily make the model 'better' since recall begins to decrease as the number of false-positives decreases (Fig 2). The same impact of background augmentation on the number of false positives was observed visually in the production pipeline (Fig 3), showing the real-world functionality of such model improvements. YOLO26 models have better native suppression of false positives compared to RF-DETR models, but at the expense of lower recall (Fig 4). Additionally, NWD is a more robust metric, compared to IoU, for evaluating our object detector performance since objects in our data may be small.

CONCLUSION

Background augmentation in training data suppresses false positive rates in RF-DETR object detectors for UAV marine imagery, though these effects quickly saturate

and have trade-offs with recall. We recommend sweeping background augmentation percentages, since the optimal percentage may depend on the nature and contents of the specific training and testing datasets.

Further work may consider sweeping further background augmentation (40% and more) to explore if the trend of increasing false positive suppression continues. Similarly, background augmentation for YOLO models may also be considered; sweeping background-only crop percentages for YOLO models may be useful to determine the optimal amount of background augmentation for these models.

ACKNOWLEDGEMENTS

My sincerest gratitude to my mentors Duane and Danelle for all of their support throughout this project. I would also like to thank the UAVs team for the data, as well as George, Megan, and the other interns for this internship experience.

The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation. Additional funding is provided by Dalio Philanthropies and the Maxwell/Hanrahan Foundation.

References:

Akyon, F. C., Altinuc, S. O., & Temizel, A. 2022. Slicing aided hyper inference and fine-tuning for small object detection. arXiv. https://doi.org/10.48550/arXiv.2202.06934

Bartolo, M., Hili, G., Seychell, D. *et al.* 2026. Litter detection from aerial imagery: a review of UAV-based approaches and deep learning techniques. *Multimed Tools Appl,* 85(232). https://doi.org/10.1007/s11042-026-21440-1

Hodgson, J., Baylis, S., Mott, R. *et al.* 2016. Precision wildlife monitoring using unmanned aerial vehicles. *Sci Rep,* 6(22574). https://doi.org/10.1038/srep22574

Jocher, G., Qiu, J., Liu, M., Lyu, S., Akyon, F. C., & Kalfaoglu, M. E. 2026. Ultralytics YOLO26: Unified real-time end-to-end vision models. arXiv. https://doi.org/10.48550/arXiv.2606.03748

Li, J. Y. Q., Duce, S., Joyce, K. E., & Xiang, W. 2021. SeeCucumbers: Using Deep Learning and Drone Imagery to Detect Sea Cucumbers on Coral Reef Flats. *Drones*, *5*(2): 28. https://doi.org/10.3390/drones5020028

Robinson, I., Robicheaux, P., Popov, M., Ramanan, D., & Peri, N. 2026. RF-DETR: Neural architecture search for real-time detection transformers. arXiv. https://doi.org/10.48550/arXiv.2511.09554

Ultralytics. 2026. Tips for best YOLOv5 training results. Ultralytics YOLOv5 documentation.

Wang, J., Xu, C., Yang, W., & Yu, L. 2021. A Normalized Gaussian Wasserstein Distance for Tiny Object Detection. arXiv:2110.13389.

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., & Xia, G. S. 2022. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing,* 190:79-93. arXiv:2206.13996.