Bird Detection and Classification based on Flow Guided Feature Aggregation

Allan Dong, School of AMME, the University of Sydney

Introduction

Frame by frame bird detection is difficult as birds can fly with fast and erratic motion.

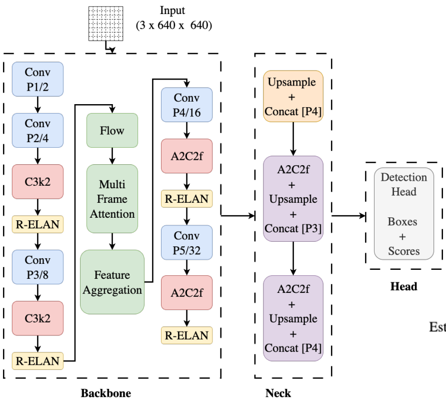
They can appear blurred or only partially visible in individual frames.

Flow Guided Feature Aggregation [1] (FGFA) combines temporal features helping the model retain spatial consistency and detect small or distant birds more reliably.

Figure 1: Common Aus species

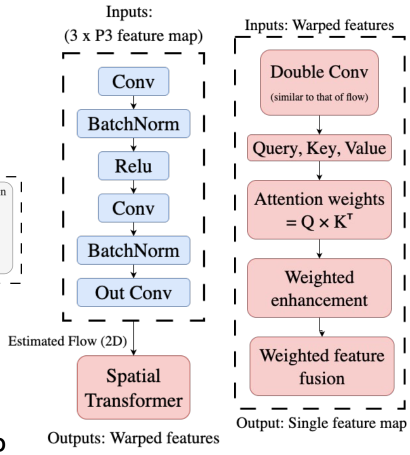
Methods

Figure 2: YOLO_FGFA architecture diagram



- The base architecture is adapted from **YOLO v12**. [2]
- It consists of three main components:
- **Backbone:** A convolutional feature extractor that captures spatial hierarchies from the input image.
- **Neck:** A feature fusion module that combines multi-scale features to enhance object representation across different resolutions.
- **Head:** Responsible for object classification and bounding box regression.

Figure 3: Flow and attention + aggregation diagram



- **Left:** Flow estimator aligns features across frames using learned 2D flow and a spatial transformer. [3]
- **Right:** Attention module computes query, key, value via conv layers.
- Attention weights guide temporal enhancement of features.
- Final output is a single fused feature map for the reference frame.

Literature cited

[1] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-Guided Feature Aggregation for Video Object Detection," *arXiv.org*, 2017.

[2] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," *arXiv.org*, 2025.

[3] T. Asanomi, K. Nishimura, and R. Bise, "Multi-Frame Attention with Feature-Level Warping for Drone Crowd Tracking," 2023

Acknowledgments

I would like to express my sincere gratitude to Dr Zi Wang, my project supervisor, for providing me with this opportunity. I also wish to acknowledge Jianqian for their guidance, support and insightful feedback throughout this research.

Further information

Results – Training

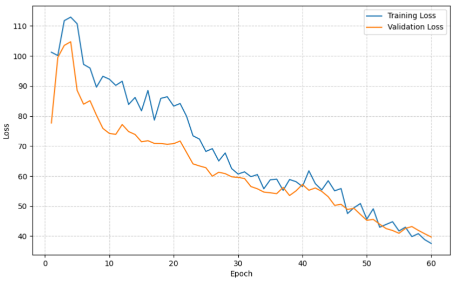


Figure 4: Training and validation loss

Training Parameters:

- Epochs: 60
- Batch size: 8
- Learning rate: 2e-4 → 1e-6 (cosine decay)
- Optimiser: Adam
- Input size: 640 × 640

The graph shows training and validation loss over 60 epochs. Both losses decrease steadily, indicating effective learning. The two losses converge, suggesting good generalisation.

Results – Feature Aggregation

Figure 5: Frames n-1, n and n+1

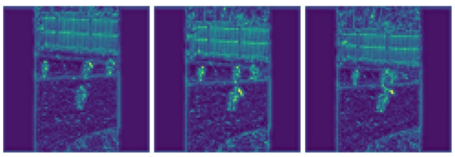


Figure 6: P3 feature maps (256x256)

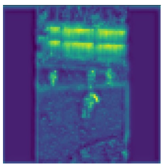


Figure 7: Flow guided, aggregated feature maps

Results - Detection

YOLO-FGFA	YOLO v12

YOLO-FGFA	YOLO v12
• Real Setting • (Bankstown Airport) • 9 Ibis detected, 1 false detection.	• 6 Ibis detected. • Sometimes higher confidence for individual detection.

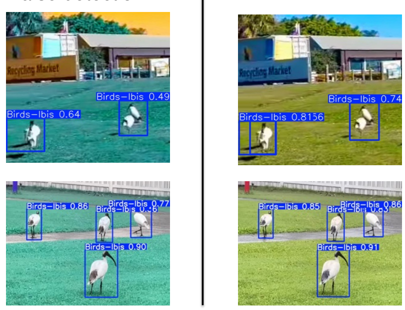


Figure 8: Method Comparison

Conclusion

1. Training was effective and the random initialised weights were trained for feature aggregation.
2. Flow-guided and attention-weighted feature fusion produces more informative feature maps.
3. Overall, FGFA improves occluded object detection, especially in cluttered and low visibility frames.

Future Work

1. The four most common bird species found around Sydney airports are the Cockatoo, Crow, Magpie and Ibis. Training the model on the first three is an imperative next step.
2. Implementation of Optical flow such as RAFT can improve warping accuracy and therefore a sharper final feature map.
3. Integration with Pan-Tilt-Zoom (PTZ) cameras would be necessary for field implementation.