

YOLO-based Semantic Feedback for Bird and Drone Detection and Tracking

Yong Song Ji, Song Hun Kim, Tae Myong Jon, Yong Bom Kim

Abstract

Bird and drone detection and tracking is a significant problem in air security. These targets are usually displayed far away with small sizes, making it difficult to obtain sufficient visual identification information.

The latest detectors use multi-scale feature pyramids to deal with each size of the target, but still cannot handle them well when dealing with extremely small objects. Deep features of the lower layers emphasize visual information, while higher-resolution features of the upper layers obtain information which is useful for location search. Designing a well-structured deep neural network to complement these two features is an important issue in object detection.

This paper proposes a novel model based on YOLO26-P2 [1-3] for detecting micro aerial targets. Using the detection ability of the higher-resolution feature of P2/4 layer as its base, this paper proposes a P3-to-P2 static fusion model and evaluates whether it improves the P2 feature. Next, this paper proposes a modified P3-to-P2 semantic feedback model and evaluates its performance by comparing it with the static fusion model. Unlike the static fusion model, the proposed method re-inserts the modified P3 feature into the P2 layer. This method is represented in the following form:

$P2 \rightarrow \text{Modified } P3 \rightarrow \text{Modified } P2$

This paper sets four models – the standard model [4, 5], the high-resolution P2 model, the P3-to-P2 static fusion model and the P3-to-P2 semantic feedback model, and compares each model to evaluate the performance of the proposed models.

Finally, the ByteTrack [21] tracking algorithm was combined with the final detector to implement MOT(multi-object tracking).

The proposed method applied simple changes of the feature layers within YOLO26 without any external image processing.

In this paper, we set up a testing plan considering both the computational load and real-time performance to evaluate the performance of the proposed method.

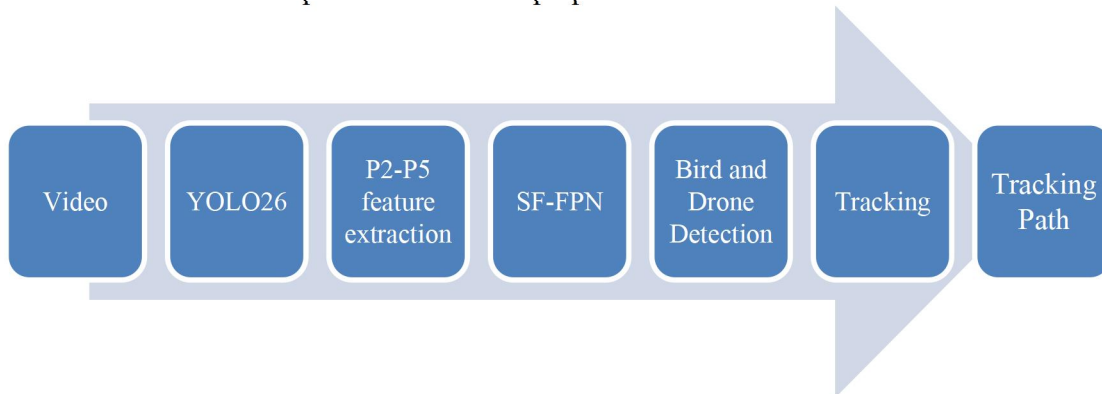


Figure 1: Flowchart

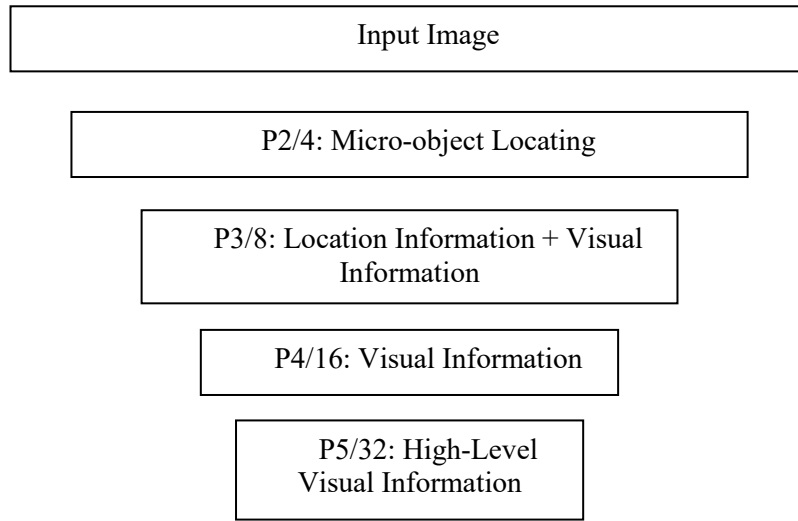


Figure2: Resolution in Each Detection Layer

Keywords: Micro-object detection, Bird and drone detection, YOLO26, Feature pyramid network, Semantic feedback, ByteTrack, Multi-object tracking

1. Introduction

The accuracy of bird and drone detection [11] is a significant issue in aviation surveillance and security fields.

In general, air targets operating at long distances have limited information about their physical appearance due to the small amount of pixel information they possess. This loss of information can be attributed to factors such as motion blur caused by movement, atmospheric conditions, image compression, illumination changes, and complex backgrounds. Therefore, detecting micro-objects in the air is a more challenging problem than in common target detection tasks.

The latest one-stage detectors [7] extract multi-scale features to handle size variation of the targets.

The feature map of the upper layers is generally higher-resolution and preserve spatial information, making it suitable for indicating the location of the targets. On the other hand, the deep feature map of the lower layers is extracted from a wide receptive field with high abstraction capability, providing strong visual information.

The features of the upper and lower layers differ significantly, which is a problem in micro-object detection tasks. Higher-resolution features are required to preserve spatial properties that help identify them, but do not provide enough visual information to distinguish them from similar background or object types. On the other hand, the deep features emphasize visual information, but suffer from repeated reduction operations leading to loss of detailed spatial information.

This spatial-visual imbalance is a serious problem in object detection tasks, especially when dealing with objects in far distances.

YOLO26-P2 adds P2/4 to the previous P3/8, P4/16 and P5/32 prediction heads to solve this issue. Even though micro-objects occupy more feature maps when the layers get deeper, the recognition problem cannot be solved by simply increasing the resolution of the feature maps.

The higher-resolution P2 feature provides strong spatial information, but weak visual information. The main issue here is that the information exchange between P2 and P3 can improve the performance of micro-object detection. In a word, the proposal is to extend the P3 feature and connect it directly to P2.

The design is based on the principle that the spatial information of P2 should complement the visual information of P3. The problem is how to connect P3 to P2.

Based on the YOLO26-P2 structure, the initial top-down P3 feature is generated before the next bottom-up P2 feature. Next, the P2 feature is propagated upward and combined with P3 to generate an improved P3 feature. It makes the improved P3 feature obtain both the higher-resolution P2 feature and the information which is generated from the multi-scale visual information. Therefore, the information obtained is more useful than the top-down P3 feature.

This paper propose a P3-to-P2 Semantic Feedback Feature Pyramid Network. The proposed method utilizes the modified P3 feature, which is generated through the bottom-up process, and extend and re-insert it into the previous higher-resolution P2. As a result, the improved P2 feature is the combination of spatial information and visual information which is improved through multi-scale processing.

To evaluate the proposed method, three test cases were set up: YOLO26, YOLO26-P2 and P3-to-P2 Static Fusion. It is essential to evaluate these models for distinguishing the semantic feedback model from normal scale feature combination methods.

We make the following contributions:

- 1) We compare micro-object detection models based on YOLO26.
- 2) We propose the static P3-to-P2 fusion structure and evaluate it by inserting visual information directly into the higher-resolution P2 layer.
- 3) We propose a lightweight SF-FPN (Semantic Feedback Feature Pyramid Network).
- 4) We compare the direct static fusion and SF-FPN.
- 5) We improve multi-object tracking by combining the modified detector with the ByteTrack algorithm.

The proposed model is a lightweight variant of YOLO26-P2, which does not require external image processing or additional network modules. It is designed to change how the internal features interact, using the standard operations available in YOLO [14].

The performance of the proposed method is evaluated through tests.

2. Related Work

2.1. Bird and Drone Detection

Bird and drone detection is essential for aviation surveillance [9] and security, so the interest in this field is growing over time. Birds and drones are clearly distinct from each other at close range but appear less distinctly in far distance. These two targets appear small and similar when seen through a complex background such as sky, clouds, buildings, mountains etc.

Object detection based on deep learning [18, 19] have improved the performance significantly compared to traditional object-detection algorithms but suffer from large performance degradation when dealing with micro-objects. This phenomenon is particularly severe when using wide-angle fixed focus cameras. Therefore, this technology requires high stability and accurate location spotting of the targets.

2.2. Micro-object detection

Primarily, micro-object detection [15, 17] requires high-resolution input images. As the images propagates through DNNs [10, 12], repeated reduction decreases the resolution of the feature map. In this process, relatively large objects maintain their own information, but micro-objects lose important structure information [16].

The FPN (feature pyramid network) solves this problem by combining information with multi-resolution. The higher-resolution layers at the top preserve spatial information, while the deep layers at the bottom provide strong visual information. Therefore, optimal combination of the feature layers for combining these two pieces of information is an important issue.

When dealing with micro-objects, the accuracy of location spotting the targets can be improved by using higher-resolution prediction layers such as P2. However, since there is insufficient contextual information to perform reliable classification in higher-resolution features, micro-object detection requires effective balance between target spotting and classifying.

2.3. Multi-scale feature fusion

Multi-scale feature fusion [6] refers to the combination of information from multiple layers, which was proposed for detecting targets at various sizes. When extracting features through DNNs, the top-down method generally propagates visual information to the direction of higher-resolution feature map. On the other hand, the bottom-up method propagates detailed location information to the deep layers.

Direct fusion of the features like concatenating or adding can improve the diversity of the features but does not guarantee that additional fusion would automatically improve the detection performance. The choice of how to mix features is a crucial problem for getting useful features.

In this paper, it was tested to determine whether inserting P3 feature into P2 would be useful and what effect inserting the improved P3 feature has on the performance.

2.4. Multi-object tracking

Object tracking relies on detection in successive dynamic frames. In bird and drone surveillance system[8], object tracking [13, 20] aims to estimate the movement trajectory over time and to analyze target motion.

ByteTrack is one of the tracking algorithms for multi-object tracking that considers both the high and low confidence detection candidates by examining their relationship with each other over time.

The accuracy of tracking depends on the accuracy of detection. If detection fails or becomes inaccurate, the tracking path may be disrupted and the target's identity may switch. Therefore, improving the performance of detection is a safeguard for improving the stability of tracking.

3. Motivation of Proposal and Assumptions

3.1. The balance between spatial information and visual information

The proposal is based on the imbalance between spatial and visual information.

The feature map of P2 provides the feature with higher-resolution and it enables the micro-object occupy many pixels, providing suitable condition for detection. But, P2 contains relatively weak visual information. On the other hand, P3 has limited spatial features but strong visual information.

P2: Strong spatial feature + Limited visual information

P3: Limited spatial feature + Strong visual information

Therefore, in order to maintain the balance between spatial and visual information, it is important to combine P2 and P3 effectively.

3.2. Inserting higher-resolution P2 layer

Assumption1: The higher-resolution feature map of P2 can preserve more spatial information than YOLO26 and YOLO26-P2, thus improving the efficiency of location spotting in object detection.

This assumption is evaluated through comparisons between YOLO26 and YOLO26-P2.

3.3. P3-to-P2 static fusion

The second problem is whether the visual information from P3 can enhance the higher-resolution feature of P2.

The static fusion feature is extracted using the following method.

$F2_{static} = H(Cat(F2, F3_u))$, where

F2: The feature of P2

F3: Top-down P3 feature

$F3_u = U(F3)$

U: Up-sampling

Cat: Feature fusion

H: Trainable fusion

The purpose of static fusion is to combine the strong space feature of P2 with the visual information of P3.

Assumption2: P3-to-P2 static fusion combines the higher-resolution spatial information and visual information to improve the performance of micro-object detection.

This assumption is evaluated by comparing YOLO26-P2 and YOLO26-P2 static fusion.

3.4. Improved semantic feedback

The top-down P3 feature used in the static fusion is extracted before P2. If we propagate the P2 feature upward and combine it with P3, this feature will include both the higher-resolution P2 feature and multi-scale visual information.

Initial P2 $\rightarrow$ bottom-up fusion $\rightarrow$ Improved P3

Then, is the improved P3 feature useful for improving the P2 feature, compared with the previous top-down P3 feature? The proposed SF-FPN utilizes the improved P3 feature.

The workflow is as follows:

Initial P2 $\rightarrow$ Bottom-up fusion $\rightarrow$ Improved P3 $\rightarrow$ Semantic Feedback $\rightarrow$ Improved P2

$F2_e = H(Cat(F2, F3_{u,r}))$, where

$F3_{u,r} = U(F3_r)$

F3r: Improved feature of P3

It forms the feedback relation of P2 $\rightarrow$ P3 $\rightarrow$ P2.

Assumption3: The improved feature of P2 from the bottom-up process provides more rich visual information for improving P2 than static fusion using the top-down P3 feature.

This assumption is evaluated by comparing YOLO26-P2-static-fusion and YOLO26-P2-feedback.

4. Proposed Methodology

4.1. YOLO26-P2

The backbone generates the feature maps at various resolutions, while the head constructs the deep top-down features and generate the higher-resolution feature of P2.

In this model, the P2 feature is obtained as follows:

$F2 = H2(Cat(U(F3), B2))$

Here, B2 is the backbone feature of the corresponding resolution.

4.2. P3-to-P2 static fusion

The static fusion model uses the top-down P3 feature.

This feature is concatenated with the original P2 feature after being expanded.

$F3_{u,t} = U(F3_t)$

$F2_c = Cat(F2, F3_{u,t})$

The fusion block generates $F2_{static} = H(F2_c)$
This model represents the direct visual information insertion into P2.

4.3. The proposed SF-FPN

SF-FPN uses the improved P3 feature which is generated through the bottom-up feature fusion.
First, P2 is used in the bottom-up feature process.
The improved P3 feature is generated as follows:
 $F3_r = Hr(Cat(D(F2), F3_t))$
Here, D means down-sampling.
The improved feature expands afterward.
 $F3_{u,r} = U(F3_r)$
Finally, $F2_e = H(Cat(F2, F3_{u,r}))$
The improved feature $F2_e$ is used in the P2 detection.

4.4. The principle of semantic feedback

The main difference between static fusion and SF-FPN is the source of feature.

Static fusion: $F2_{static} = H(Cat(F2, U(F3_t)))$

SF-FPN: $F2_e = H(Cat(F2, U(F3_r)))$

Therefore, the static fusion is the top-down $P3 \rightarrow P2$, and the SF-FPN is $P2 \rightarrow \text{improved } P3 \rightarrow P2$

The proposed method does not simply mean the additional fusion only. It indicates that the improved visual information is more useful than the visual information which is directly extracted through the top-down process.

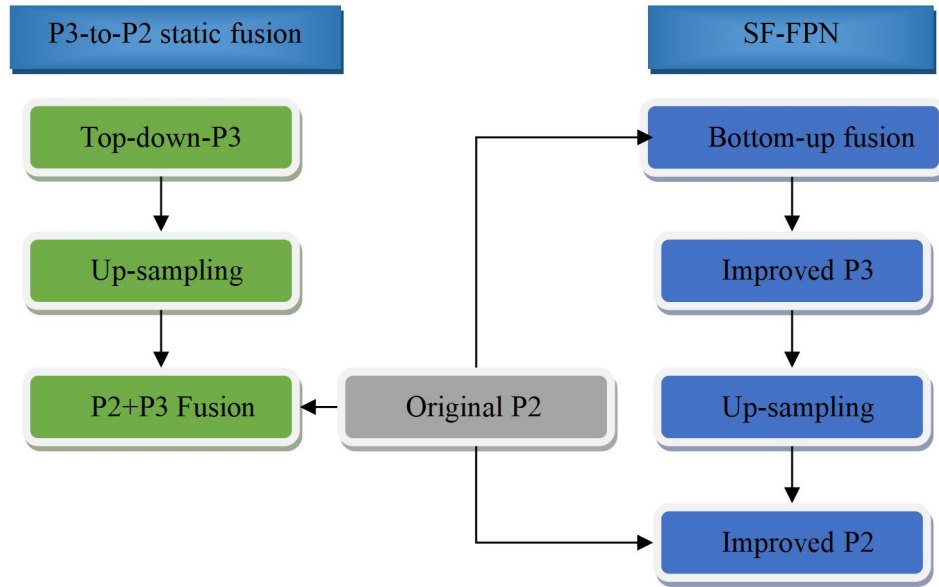


Figure3: Comparison between static fusion and SF-FPN

The final detection utilizes the features of the following layers:
Improved P2 + Improved P3 + P4 + P5

4.5. Characteristics of calculation

The proposed SF-FPN applies fusion calculation and feature concatenation calculation, and it utilizes a lightweight C2k2 fusion block, which means that the proposed method requires less computational load.

Table 1. Comparison of Models

Model	Number of parameters	FLOPs	Inference time
YOLO26	2,505,360	5.8G	50ms
YOLO26-P2	2,516,768	7.5G	80ms
Static Fusion	47,350,656	140.7G	700ms
Semantic Feedback	47,350,656	140.7G	710ms

5. Detection and Tracking

5.1. Brid and drone detection

The detector gets images or videos as input and predict the object’s location, confidence and its kind.

The proposed detection model is designed to demonstrate strong detection performance using the multi-scale features to preserve visual information.

5.2. Tracking

This paper applied the ByteTrack tracking algorithm for its high speed and stability.

Therefore, the ByteTrack tracking engine gets the detection result as its input and generates the tracking path for each target based on the relatedness of the detected objects in subsequent frames.

5.3. Importance of detection quality

The stability of detection is crucial to the performance of tracking. Unstable detection of small objects can lead to the fragmentation or loss of track path and confusion of identification information. Therefore, improving object detection quality is a responsibility for enhancing tracking performance.

Tracking performance can be evaluated by MOTA, IDF1, HOTA and changes in identification information, track breaks etc.

6. Experimentation

6.1. Dataset

The dataset used for the experiment consists of images of birds, drones and backgrounds. To ensure generalizability, the dataset includes various conditions such as target size, distance, lighting, weather, background complexity, direction of targets, blur, etc.

Table 2. Dataset Details

Total Number of Images	Train	Validation	Test	Birds	Drones
14182	12450	876	856	4349	4388

6.2. Target size classification

The proposed method deals with micro-objects, thus it is important to analyze the size of the targets.

We have classified target size into four categories according to the size and resolution of the target.

Micro-objects: target size $< 32 \times 32$

Small objects: $32 \times 32 < \text{target size} < 96 \times 96$
Medium objects: $96 \times 96 < \text{target size} < 128 \times 128$
Large Objects: $\text{target size} > 128 \times 128$

Table 3. Distribution of target size

	Micro	Small	Medium	Large
Bird	69.8%	19.7%	3.8%	6.7%
Drone	43.4%	56.4%	0.1%	0.4%

6.3. Training

Training is performed with NVIDIA H100 for 5 days.

All test cases were set with the same training parameters, including input image resolution, data subdivision strategy, number of epochs, batch size, optimizer, learning-rate strategy, and data augmentation etc.

Table 4. Training parameters

Resolution of input image	Epochs	Batch size
640 * 640	100	24

Precision, recall, mAP@0.5:0.95, and class-wise accuracy are evaluated.

6.4. Experiment cases

①. Case A: Standard YOLO26

In case A, we have evaluated the performance of micro-object detection with the standard YOLO26 which has P3/8, P4/16 and P5/32 layers. The purpose is to establish a baseline performance.

②. Case B: YOLO26-P2

In case B, YOLO26-P2 has P2/4, P3/8, P4/16, P5/32 layers..

The purpose is to evaluate whether the higher-resolution detection with P2 improves the performance.

③. Case C: Static Fusion

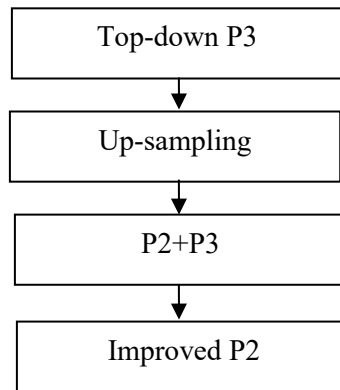


Figure4: Process flow of static fusion

The purpose is to evaluate whether the static fusion improves the performance of P2.

④. Case D: SF-FPN

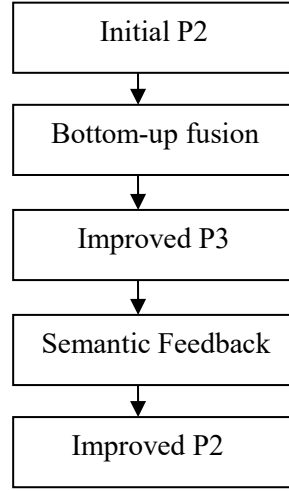


Figure5: Process flow of SF-FPN

The purpose is to evaluate whether the SF-FPN improves performance over static fusion.

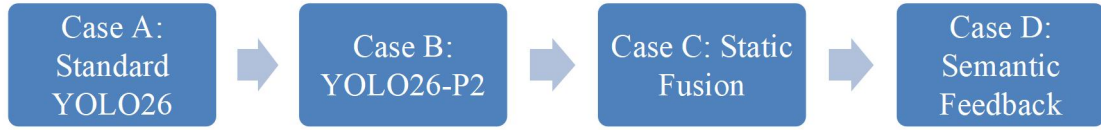


Figure6: Experiment Cases

7. Result Analysis

Table 5. Overall Performance Evaluation of Different Proposed Models

Model	Precision	Recall	mAP50	mAP50-95
Case A(YOLO26)	0.80629	0.81648	0.85721	0.5502
Case B(YOLO26-P2)	0.8175	0.84753	0.88056	0.57987
Case C(Static Fusion)	0.85256	0.87458	0.90413	0.61219
Case D(Semantic Feedback)	0.82774	0.89375	0.89754	0.61433

Table 6. Performance Evaluation of Micro-object Detection

Model	Precision	Recall	mAP50	mAP50-95
Case A(YOLO26)	0.755	0.731	0.744	0.321
Case B(YOLO26-P2)	0.67	0.748	0.718	0.337
Case C(Static Fusion)	0.747	0.77	0.74	0.33
Case D(Semantic Feedback)	0.749	0.753	0.759	0.378



Figure7: Bird and Drone Detection Result

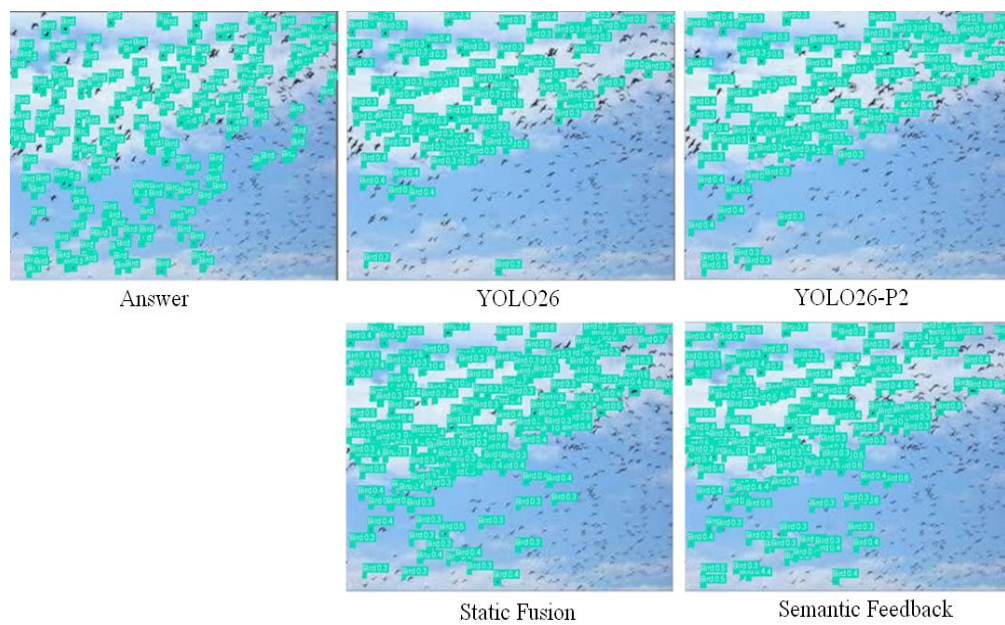


Figure8: Detection Results of Proposed Models



Figure9: Bird and Drone Detection Failure

8. Limitations

The proposed method outperforms the previous methods in terms of accuracy but also had several limitations.

Firstly, if the visual information of the target has severe damage caused by atmospheric conditions, blurring, motion blur, and noise introduced by image compression, the detection performance will be significantly degraded.

Secondly, compared to the standard YOLO26, this proposed model has a larger computational load.

Thirdly, the distribution of target sizes or the quality of images used in training will affect the detection performance.

9. Conclusion

This paper propose the study on the fusion of higher-resolution features and visual information based on YOLO26 for micro-object detection.

The paper proposes the SF-FPN (Semantic Feedback-Feature Pyramid Network) model based on the principle that micro-object detection requires detailed spatial and visual information.

The proposed network structure shows significant improvement in micro-object detection compared to the previous networks by comparing P2, which preserves high-resolution spatial information, with P3-to-P2 static fusion model.

References

- [1] Peijun WANG and Jinhua ZHAO, "SOD-YOLO: Enhancing YOLO-Based Detection of Small Objects in UAV Imagery," in CVPR, 2025
- [2] Dapinder Kaur, Neeraj Battish , Arnav Bhavsar , and Shashi Poddar, "YOLOBirDrone: Dataset for Bird vs Drone Detection and Classification and a YOLO based enhanced learning architecture, " in CVPR, 2026.
- [3] Farzaneh Dadrass Javan, Farhad Samadzadegan, Mehrnaz Gholamshahi, and Farnaz Ashatari Mahini, "A Modified YOLOv4 Deep Learning Network for Vision-Based UAV Recognition, " in Drones 2022, 6
- [4] Ranjan Sapkota and Manoj Karkee, "ULTRALYTICS YOLO EVOLUTION: AN OVERVIEW OF YOLO26, YOLO11, YOLOV8, AND YOLOV5 OBJECT DETECTORS FOR COMPUTER VISION AND PATTERN RECOGNITION, " in CVPR, 2025
- [5] Juan R. Terven and Diana M. Cordova-Esparza, "A COMPREHENSIVE REVIEW OF YOLO ARCHITECTURES IN COMPUTER VISION: FROM YOLOV1 TO YOLOV8 AND YOLO-NAS," in CVPR 2024
- [6] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. "Object detection in 20 years: A survey." In Proceedings of the IEEE, 111(3):257–276, 2023.
- [7] Xianxu Zhai 1,2, Zhihua Huang 1,2,*, Tao Li 1,2, Hanzheng Liu 1,2 and Siyuan Wang, in "YOLO-Drone: An Optimized YOLOv8 Network for Tiny UAV Object Detection," in Electronics 2023, 12, 3664
- [8] Shi, X, Yang, C, Xie, W, Liang, C, Shi, Z, and Chen, J, "Anti-Drone System with Multiple Surveillance Technologies, Architecture Implementation, and Challenges,". IEEE Commun. Mag. 2018, 56
- [9] Chen, Q.Q, Feng, Z.W, Zhang, G.B, "Dynamic modelling and simulation of anti-UAV tethered-net capture system," J. Natl. Univ. Def. Technol. 2022, 44
- [10] Wu, X, Sahoo, D, Hoi, S.C.H, "Recent advances in deep learning for object detection," Neurocomputing 2020, 396
- [11] Li, H, Yang, J, Mao, Y, Hu, Q, Du, Y, Peng, J, Liu, C, "A UAV detection algorithm combined with lightweight network," In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021; Volume 5

- [12] Emanuele Malagoli and Luca Di Persio, "2D Object Detection: A Survey," in *Mathematics* 2025, 13, 893
- [13] Jianfeng Han, Feijie Sun, Zihan Xu, Lili Song, and Jiandong Fang, "An Enhanced Algorithm Integrating YOLOv11 and ByteTrack for Small-Object Detection and Tracking in Low-Altitude Remote Sensing Imagery," in *Remote Sens.* 2026, 18, 1547
- [14] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., "You Only Look Once: Unified, Real-Time Object Detection," In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016
- [15] Wang, G., Chen, Y., and An, P., "UAV-YOLOv8: A small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios," in *Sensors* 2023, 23, 7190
- [16] Yue, M., Zhang, L., Huang, J., and Zhang, H., "Lightweight and Efficient Tiny-Object Detection Based on Improved YOLOv8n for UAV Aerial Images," in *Drones* 2024, 8, 276
- [17] Han, Y., Wang, C., Luo, H., Wang, H., Chen, Z., Xia, Y., and Yun, L., "LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design," in *Sci. Rep.* 2025, 15, 22627
- [18] Qihan Feng¹, Xinzheng Xu¹ and Zhixiao Wang, "Deep learning-based small object detection: A survey," in *Mathematical Biosciences and Engineering* Volume 20, Issue 4, 6551-6590
- [19] Syed Sahil Abbas Zaidi, Mohammad Samar Ansari, Asra Aslam, Nadia Kanwal, Mamoona Asghar, and Brian Lee, "A Survey of Modern Deep Learning based Object Detection Models," in *CVPR2021*
- [20] J. Berclaz, F., Fleuret, E., Turetken, and P. Fua., "Multiple object tracking using k-shortest paths optimization," in *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819, 2011
- [21] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang, "ByteTrack: Multi-Object Tracking by Associating Every Detection Box" in *CVPR* 2022