Minor Object Recognition from Drone Image Sequence

Duy-Linh Nguyen[0000−0001−6184−4133], Xuan-Thuy Vo[0000−0002−7411−0697], Adri Priadana[0000−0002−1553−7631], and Kang-Hyun Jo[0000−0002−4937−7082]

Department of Electrical, Electronic and Computer Engineering, University of Ulsan, Ulsan 44610, South Korea
ndlinh301@mail.ulsan.ac.kr, xthuy@islab.ulsan.ac.kr, priadana@mail.ulsan.ac.kr, acejo@ulsan.ac.kr

Abstract. Object detected in drone imagery is an interesting topic in the Computer Vision field. This work was widely applied in traffic analysis and control, rescue systems, smart agriculture, etc. However, many challenges exist in developing and optimizing applications because of object density, multi-scale objects, and blur motion. To partly solve the above problems, this research focuses on improving the performance of the YOLOv5m network based on the advantages of the Bi-directional Feature Pyramid Network (BiFPN), Transformer, and Convolutional Block Attention Module (CBAM). The experiments achieve 68.6% and 42.6% of mAP on the proposed datasets (ISLab-Drone) and VisDrone 2021, respectively. That demonstrates the outperformance of the network comparable to other networks under the same testing conditions.

Keywords: Convolutional neural network (CNN) · BiFPN · CBAM · UAV imagery · YOLOv5m · ViT.

1 Introduction

For a long time, researchers have focused on object detection from the air using unmanned aerial vehicles (UAVs). AUVs can collect ground images from different altitudes and speeds. These techniques are widely deployed in forest protection [5], wildlife protection [9], and surveillance systems [4]. In particular, with the rapid development of smart cities, the support tools for the operation, monitoring, and protection process become more and more necessary. The image processing and analysis techniques are also required to accommodate the mobility, compactness, and power limitations of UAVs. An ideal choice for these applications is object detection based on the YOLO [14,15,16,7,8] network family. Most objects detected in drone imagery methods face some common problems, such as object density, multi-scale objects, and blur motion. Object density caused by the overlap of objects hinders the detection of hidden objects. UAVs capture images with variable altitudes, resulting in images that also vary in scale from tiny, small, and medium to large. The flight speed of UAVs leads to the image having a certain blurry quality, which affects the locating and detecting

of objects in the image. From the above observations, this paper proposes an improved method for the YOLOv5m network used in detecting drone-captured objects. Research focuses on optimizing the backbone and neck networks based on the Bi-directional Feature Pyramid Network (BiFPN) [18], Transformer [2], and Convolutional Block Attention Module (CBAM) [20]. Besides, this work also provides a drone-captured image dataset with diverse scenarios for object detection tasks.

The main contribution enlists as follows:
1 - Redesigns the backbone and neck networks of YOLOv5m architecture with a combination of BiFPN, Transformer, and CBAM.
2 - Adds one more detection head with new anchor sets to improve the tiny object detection task.
3 - Proposes a drone-captured image dataset used in the Computer Vision field.
The rest of the paper is organized as follows: Section 2 introduces the related works to object detected in drone imagery tasks. Section 3 explains the proposed method. Section 4 analyzes and proves the experimental results. Section 5 summarizes the important issue and future orientation.

2 Related work

2.1 Deep neural network-based method

These methods leverage the advantage of deep learning neural networks (DNNs) to consider the differences between foreground and background to cluster and detect objects in drone-captured scenarios. The work [21] proposed a framework that combines clustering and detection by sequentially searching the clustered areas and detecting drone-captured objects belonging to these regions. Observing the issues of [21], the method [1] added an efficient self-adaptive region to build the global-local detection network to improve the accuracy of high-density and large-scale object detection. Another approach [23] used a region estimation network to find the high-density drone-captured objects in diverse areas. For drone-captured vehicle image challenges, the study [24] aligns the feature between different viewpoints, backgrounds, illumination, and weather in the domain adaptation. DNN-based methods achieved high accuracy in detecting objects but the computational cost is still a huge problem.

2.2 Convolutional neural network-based method

In recent years, the strong development of convolutional neural networks (CNNs) in the objects detected in drone imagery topic has attracted the attention of many researchers. The research [17] evaluated different backbone architectures, prediction heads, and model pruning techniques to select a better combination in a fast object detection network. TPH-YOLOv5 [27] combined the ideas of the transformer detection head and original detection head in the YOLOv5 network architecture to improve the accuracy in detecting large-size variation objects and

high-density objects. Inheriting previous work, the work in [25] work proposed a cross-layer asymmetric transformer (CA-Trans) to replace the additional prediction head in TPH-YOLOv5 for more efficiency in tiny object detection. The outstanding advantage of CNN-based methods is achieving ideal object detection accuracy with flexible integration of other techniques such as Transformer or Attention algorithms.

3 Methodology

Fig. 1 presents in detail the proposed object detected in drone imagery network. This method is improved from the original YOLOv5m architecture [7] with three components: Backbone, Neck, and detection head.

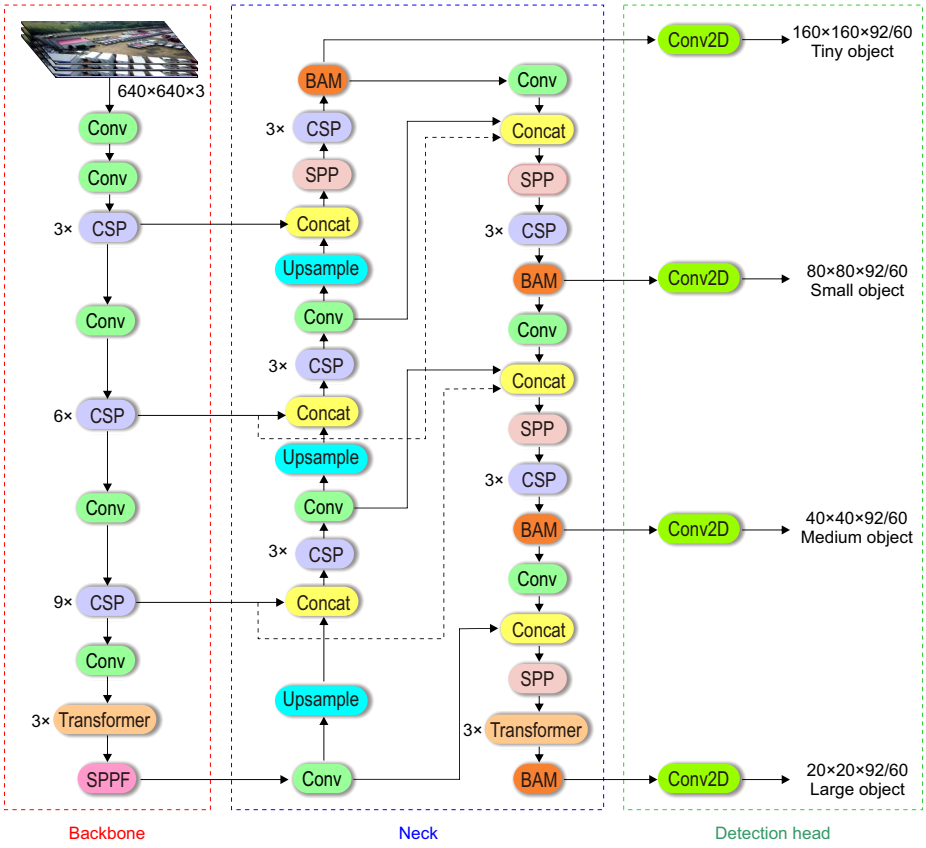

Fig. 1. The proposed object detected in drone imagery network. Numbers 92/60 is detector output coefficients for ISLab-Drone and VisDrone 2021 datasets, respectively.

3.1 Proposed network architecture

Backbone module: The backbone network plays a very important role in extracting features for the entire network. Based on the existing architecture of YOLOv5, this work evaluates and replaces several components to reduce computational complexity and network parameters while still ensuring feature extraction capabilities. Specifically, the Focus module is replaced by a simpler architecture, named the Conv block. This block includes one 2D convolution (Conv2D), one batch normalization (BN), and one Sigmoid Linear Unit (SiLU) activation function. The design of the Conv block is shown in Fig. 2 (a). The body of the backbone network still uses a combination of Conv blocks and Cross Stage Partial modules (CSP) with ratios of 3, 6, and 9. Fig. 2 (b) describes the architecture of the CSP module. The end of the backbone network adds three Transformer blocks and replaces the Spatial Pyramid Pooling (SPP) module with the Spatial Pyramid Pooling Fast (SPPF) module. The SPPF module applies all of the max pooling (MaxPooling) layers with the same kernel size (k = 5). The architecture of SPP and SPPF modules are depicted in Fig. 2 (c) and Figs. 2(d). This block is inspired by the Vision Transformer (ViT) [2]

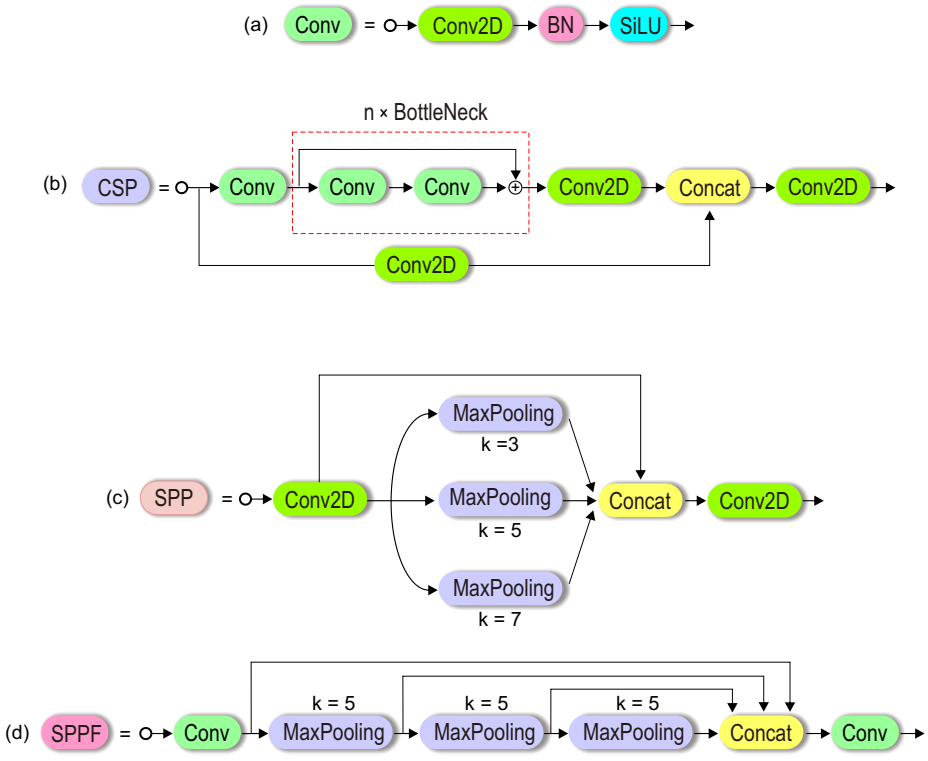

Fig. 2. The architecture of basic modules.

which is used to capture global information and rich contextual information [27]. The structure of the Transformer blocks is depicted in Fig. 3. Each Transformer block is built from two sub-blocks, the Multi-head Attention (MA) layer and the Multilayer Perceptron (MLP) layer (fully connected layers). Besides, Residual connections are used between two sub-blocks. Therefore, Transformer bock also increases the ability to extract rich local information. Based on Neck's existing architecture in YOLOv5m, this work adds a CBAM at the end of each level in the multi-level feature map (small, medium, and large). On the other hand, the Neck also appends a feature map level to detect extremely small objects (Tiny). In total, this module has four feature map levels.

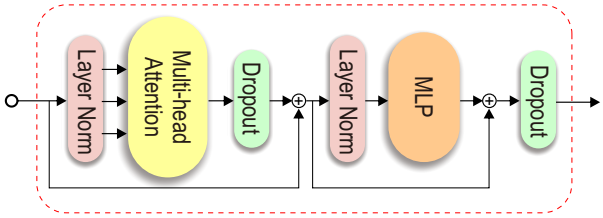

Fig. 3. The Transformer block.

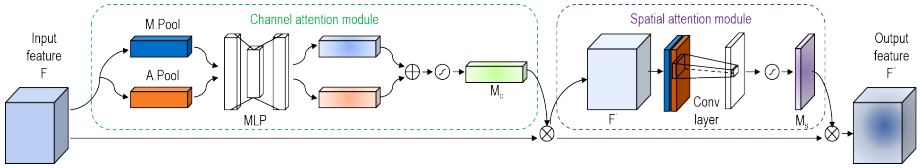

Fig. 4. The architecture of the CBAM module.

Neck module: The Neck module is designed based on the combination of the Path Aggregation Network (PAN) [12] and the Bi-directional Feature Pyramid Network (Bi-FPN) [19]. These two architectures support each other to synthesize the current feature map with the feature maps in previous stages to enrich the information for the feature maps in the next stages. Based on Neck's existing architecture in YOLOv5m, this work stacks the SPP module and CBAM at the end of each level in the multi-level feature map (small, medium, and large). Especially in the last level, three transformer blocks are stacked between the SPP and CBAM modules to enrich the useful information for large-size object detection. On the other hand, the Neck also adds a feature map level to detect extremely small objects (tiny). In total, this module generates four feature map levels. The architecture of the CBAM module is presented in Fig. 4.

Detection head module: The detection head module leverages three feature map levels of the YOLOv5m architecture from the Neck module, including $80 \times 80 \times 256$, $40 \times 40 \times 512$, and $20 \times 20 \times 1024$. Besides, this study adds one more detection head at $160 \times 160 \times 128$ feature map level to increase tiny object detection ability. The number of anchor boxes is set at four and their sizes are redesigned to be suitable for the objects in the ISLab-Drone and VisDrone 2021 datasets. The details of each detection head and the anchor size are described in Table 1.

Table 1. Detection heads and anchors.

Head	Input	Anchors	Ouput	Object
1 (Added)	$160 \times 160 \times 128$	(7, 9), (9, 17), (17, 15), (13, 27)	$160 \times 160 \times 92/60$	Tiny
2	$80 \times 80 \times 256$	(21, 28), (36, 18), (23, 47), (35, 33)	$80 \times 80 \times 92/60$	Small
3	$40 \times 40 \times 512$	(58, 29), (43, 60), (82, 46), (66, 88)	$40 \times 40 \times 92/60$	Medium
4	$20 \times 20 \times 1024$	(133, 77), (111, 135), (206, 137), (197, 290)	$20 \times 20 \times 92/60$	Large

3.2 Loss function

The loss function is defined as follows:

$$Loss = \lambda_{box}\mathcal{L}_{box} + \lambda_{obj}\mathcal{L}_{obj} + \lambda_{cls}\mathcal{L}_{cls}, \tag{1}$$

in which, $\mathcal{L}_{box}$ is the bounding box regression loss using CIoU loss, $\mathcal{L}_{obj}$ is the object confidence score loss using Binary Cross Entropy loss, and L_{cls} is the classes loss also using Binary Cross Entropy loss to calculate. λ_{box}, λ_{obj}, and λ_{cls} are balancing parameters.

4 Experiments

4.1 Dataset

The experiments in this paper are trained and evaluated on two datasets, ISLab-Drone and VisDrone 2021 [26]. The ISLab-Drone dataset was proposed by the Intelligent Systems Laboratory (ISLab) at the University of Ulsan, South Korea. This dataset includes 10,000 images collected using a UAV under different weather and altitude conditions in Ulsan City and Daegu City, South Korea. It contains 18 categories: tree, person, animal, house, apartment/building, school, office, traffic sign, traffic light, streetlamp/telephone pole, banner, milestone, bridge, tower, car_vechicle, bus_vehicle, truck_vehicle, motorcycle/bike_vehicle. The number of images is divided based on the ratio 5:2.5:2.5 for the training, evaluation, and testing sets. The VisDrone Dataset 2021 is a large-scale benchmark created by the AISKYEYE team at the Lab of Machine Learning and Data Mining, Tianjin University, China. VisDrone 2021 consists of four subsets,

including a training set, validation set, test-dev set, and test-challenge set. These experiments only use the training set, the validation set, and the test-dev set with 6,471, 548, and 1,610 images, respectively. The images are separated into 10 classes: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, motor.

4.2 Experimental setup

This experiment applies the original YOLOv5 [7] as a code base using Python programming language and the Pytorch framework. The training and evaluation processes are conducted on a GeForce GTX 1080Ti GPU. The Adam optimization is used. The learning rate is initially set to 10^2 and the final by 10^5. The momentum began at 0.8 and then increased to 0.937. The training process takes 300 epochs with a batch size of 32. The balancing parameters λ_{cls}=0.5, λ_{box}=0.05, and λ_{obj}=1, respectively. Several data augmentation methods are applied such as mosaic, mix-up, flip left-right, and flip up-down.

4.3 Experimental result

For the ISLab-Drone dataset, this experiment conducts training and evaluation from scratch YOLOv5 (n, s, m, l, x) series, proposed network, and then compares the performance between them. As the results are shown in Table 2, the proposed network achieves 68.6% mean Average Precision (mAP) which demonstrates that the network outperforms the whole of the YOLOv5 series. When compared with the best competitor (YOLOv5l), the proposed network achieves better performance at 2.5% of mAP while the number of parameters and GLOPS are lower nearby twice. Compared to YOLOV5m, the size and computational complexity of the proposed network are light larger than YOLOv5 but the performance is better than YOLOv5m at 6.4% of mAP.

Table 2. The comparison result of the proposed detection network with YOLOv5 series on the test set of the ISLab-Drone dataset.

Model	Parameter	GFLOPs	Weight (MB)	mAP (%)
YOLOv5x	86,287,807	204.1	173.3	65.6
YOLOv5l	46,199,823	107.9	114.3	66.1
YOLOv5m	20,921,631	48.1	42.3	62.2
YOLOv5s	7,058,671	15.9	14.5	63.6
YOLOv5n	1,783,519	4.2	3.9	55.3
Proposed method	27,377,416	67.1	55.8	68.6

In the case of the VisDrone dataset, this experiment compares the proposed network with recent work under the same conditions. As a result, Table 3 shows that the proposed network still outperforms existing studies and better best competitor by 0.5% of mAP. However, when compared to the latest version of

YOLO (YOLOv8m), the proposed network is 2.4% of mAP worse. This poses many challenges for continuously improving the proposed network in the future. The proposed network can present a better ability in object detection tasks. However, this experiment also issues several problems when detecting objects with very small scale and overlapping objects. The qualitative results of the proposed network on the test set of the ISLab-Drone and the test-dev-set of the VisDrone datasets are shown in Fig. 5.

Table 3. The comparison result of the proposed method with other networks on the test-dev set of the VisDrone dataset. The symbol "†" denotes the re-trained models.

Model	Parameter	GFLOPs	Weight (MB)	mAP (%)
YOLOv5m†	20,889,303	48.1	42.3	29.6
YOLOv8m†	25,845,550	78.8	52.0	45.0
HawkNet [11]	N/A	N/A	N/A	25.6
ClusDet [21]	N/A	N/A	N/A	28.4
DMNet[10]	N/A	N/A	N/A	29.4
Method in [23]	N/A	N/A	N/A	30.3
DSHNet [22]	N/A	N/A	N/A	30.3
CDMNet [3]	N/A	N/A	N/A	31.9
GLSAN [1]	N/A	N/A	N/A	32.5
DCRFF [13]	N/A	N/A	N/A	35.0
UFPMP-Net [6]	N/A	N/A	N/A	39.2
TPH-YOLOv5++ [25]	N/A	N/A	N/A	41.4
TPH-YOLOv5 [27]	N/A	N/A	N/A	42.1
Proposed method	27,331,208	66.9	55.7	42.6

4.4 Ablation study

To evaluate the influence of each module in the proposed network, this study also conducts several ablation studies. By replacing the proposed modules with the original YOLOv5m network architecture, training, and evaluating on the test data set of the ISLab-Drone dataset. The results in Table 4 show that the Bi-FPN network plays an important role in enriching information for feature maps. Besides, Transformer and CBAM modules support the information capture process from local to global. That is why this work chooses the perfect combination of Bi-FPN, CBAM, and Transformer modules to improve the YOLOv5m network from 62.2% to 68.6% of mAP.

5 Conclusion

This paper conducted a technique to improve the original YOLOv5m architecture for objects detected in drone imagery. Based on YOLOv5m, the proposed network contains three parts: backbone, neck, and head modules. The backbone

Fig. 5. The qualitative results of the proposed network on the test set of the ISLab-Drone and the test-dev-set of the VisDrone datasets.

Table 4. Ablation studies with different proposed networks on the test set of the ISLab-Drone dataset.

Module	Proposed network			
Transformer	✓			✓
Bi-FPN		✓		✓
BAM			✓	✓
SPPF	✓	✓	✓	✓
SPP	✓	✓	✓	✓
Parameter	24,494,767	26,954,255	25,330,744	27,377,416
Weight (MB)	55.3	55.5	54.9	55.8
GFLOPs	66.5	66.9	64.7	67.1
mAP(%)	58.8	65.2	51.2	68.6

is redesigned using simple architectures and Transformer modules. The neck is redesigned with the bi-FPN network, CBAM, and Transformer. A new head for tiny object detection is added to the detection head module and resized the anchors to fit the detection tasks. The experimental result presented the outstanding performance of the proposed network. This study will be further developed with tiny and overlapping object detection integrated with the idea in the latest YOLOv8 architecture for the future.

Acknowledgement

This result was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-003)

References

1. Deng, S., Li, S., Xie, K., Song, W., Liao, X., Hao, A., Qin, H.: A global-local self-adaptive network for drone-view object detection. IEEE Transactions on Image Processing **30**, 1556–1569 (2021). https://doi.org/10.1109/TIP.2020.3045636
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. CoRR **abs/2010.11929** (2020), https://arxiv.org/abs/2010.11929
3. Duan, C., Wei, Z., Zhang, C., Qu, S., Wang, H.: Coarse-grained density map guided object detection in aerial images. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2789–2798 (2021). https://doi.org/10.1109/ICCVW54120.2021.00313
4. Gu, J., Su, T., Wang, Q., Du, X., Guizani, M.: Multiple moving targets surveillance based on a cooperative network for multi-uav. IEEE Communications Magazine **56**(4), 82–89 (2018). https://doi.org/10.1109/MCOM.2018.1700422

5. Hird, J.N., Montaghi, A., McDermid, G.J., Kariyeva, J., Moorman, B.J., Nielsen, S.E., McIntosh, A.C.S.: Use of unmanned aerial vehicles for monitoring recovery of forest vegetation on petroleum well sites. Remote Sensing **9**(5) (2017). https://doi.org/10.3390/rs9050413, https://www.mdpi.com/2072-4292/9/5/413
6. Huang, Y., Chen, J., Huang, D.: Ufpmp-det: Toward accurate and efficient object detection on drone imagery. CoRR **abs/2112.10415** (2021), https://arxiv.org/abs/2112.10415
7. Jocher, G., et al.: ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements (Oct 2020). https://doi.org/10.5281/zenodo.4154370, https://doi.org/10.5281/zenodo.4154370
8. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics yolov8 (2023), https://github.com/ultralytics/ultralytics
9. Kellenberger, B., Volpi, M., Tuia, D.: Fast animal detection in uav images using convolutional neural networks. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 866–869 (2017). https://doi.org/10.1109/IGARSS.2017.8127090
10. Li, C., Yang, T., Zhu, S., Chen, C., Guan, S.: Density map guided object detection in aerial images. CoRR **abs/2004.05520** (2020), https://arxiv.org/abs/2004.05520
11. Lin, H., Zhou, J., Gan, Y., Vong, C.M., Liu, Q.: Novel up-scale feature aggregation for object detection in aerial images. Neurocomputing **411**, 364–374 (2020). https://doi.org/https://doi.org/10.1016/j.neucom.2020.06.011, https://www.sciencedirect.com/science/article/pii/S0925231220309784
12. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. CoRR **abs/1803.01534** (2018), http://arxiv.org/abs/1803.01534
13. Mittal, P., Sharma, A., Singh, R., Dhull, V.: Dilated convolution based rcnn using feature fusion for low-altitude aerial objects. Expert Systems with Applications **199**, 117106 (2022). https://doi.org/https://doi.org/10.1016/j.eswa.2022.117106, https://www.sciencedirect.com/science/article/pii/S0957417422005103
14. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection. CoRR **abs/1506.02640** (2015), http://arxiv.org/abs/1506.02640
15. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. CoRR **abs/1612.08242** (2016), http://arxiv.org/abs/1612.08242
16. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR **abs/1804.02767** (2018), http://arxiv.org/abs/1804.02767
17. Ringwald, T., Sommer, L., Schumann, A., Beyerer, J., Stiefelhagen, R.: Uav-net: A fast aerial vehicle detector for mobile platforms. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 544–552 (2019). https://doi.org/10.1109/CVPRW.2019.00080
18. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. CoRR **abs/1911.09070** (2019), http://arxiv.org/abs/1911.09070
19. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. CoRR **abs/1911.09070** (2019), http://arxiv.org/abs/1911.09070
20. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. CoRR **abs/1807.06521** (2018), http://arxiv.org/abs/1807.06521
21. Yang, F., Fan, H., Chu, P., Blasch, E., Ling, H.: Clustered object detection in aerial images. CoRR **abs/1904.08008** (2019), http://arxiv.org/abs/1904.08008
22. Yu, W., Yang, T., Chen, C.: Towards resolving the challenge of long-tail distribution in uav images for object detection. In: 2021 IEEE Winter Con-

ference on Applications of Computer Vision (WACV). pp. 3257–3266 (2021). https://doi.org/10.1109/WACV48630.2021.00330
23. Zhang, J., Huang, J., Chen, X., Zhang, D.: How to fully exploit the abilities of aerial image detectors. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 1–8 (2019). https://doi.org/10.1109/ICCVW.2019.00007
24. Zhang, R., Newsam, S., Shao, Z., Huang, X., Wang, J., Li, D.: Multi-scale adversarial network for vehicle detection in uav imagery. ISPRS Journal of Photogrammetry and Remote Sensing **180**, 283–295 (2021). https://doi.org/https://doi.org/10.1016/j.isprsjprs.2021.08.002, https://www.sciencedirect.com/science/article/pii/S0924271621002021
25. Zhao, Q., Liu, B., Lyu, S., Wang, C., Zhang, H.: Tph-yolov5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer. Remote Sensing **15**(6) (2023). https://doi.org/10.3390/rs15061687, https://www.mdpi.com/2072-4292/15/6/1687
26. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3119563
27. Zhu, X., Lyu, S., Wang, X., Zhao, Q.: Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. CoRR **abs/2108.11539** (2021), https://arxiv.org/abs/2108.11539