Neck Ablation Study of FPN and BiFPN in Clustered Object Detection

[1]Siddh Vyas, [2]Dr. Neelam Jain
[12]Department of Computer Science SVKM's Mithibai College of Arts
[12]Chauhan Institute of Science Amrutben Jivanlal College of Commerce And Economics (Autonomous), Mumbai, India

***Abstract*—Photographed images of unmanned aerial vehicles often have high densities of spatial clusters of objects (pedestrians and vehicles) at a point. The individual objects are difficult to detect in these scenes because of the variation in scale, occlusion and complicated backgrounds. In most aerial activity monitoring projects, it is more beneficial to determine dense areas of activity, as opposed to finding the individual object instances. This paper focuses on a cluster level detection formulation of aerial images. The model does not detect any individual objects but rather detects spatial clusters indicating dense clusters of objects around the object. The clusters are represented by only one bounding box known as a single detection class. This helps in examining the effect of the Feature Pyramid Network structures on the performance of cluster detection models. Two EfficientDet detection architectures are compared; a single-pass pyramid design that estimates a classic Feature Pyramid Network, and a stacked Bidirectional Feature Pyramid Network design. The other elements of the detection system are fixed to allow control of an architectural ablation. It takes place to test the hypothesis with experiments based on manually re-annotated imagery based on the VisDrone aerial dataset. COCO-style metrics of performance on models are based on conceptually estimation of performance through the computational complexity and floating point operations. The findings will conclude on providing quantitative improvements in cluster level aerial detection with respect to iterative bidirectional multi-scale feature fusion.**

***Index Terms*—Aerial Object Detection, Cluster Detection, Feature Pyramid Network (FPN), Bidirectional Feature Pyramid Network (BiFPN), EfficientDet, Multi-scale Feature Fusion, VisDrone Dataset.**

I. Introduction

Drones are used in increasing applications of drone-based imaging systems in diverse applications including monitoring activities in cities, road patrols, amazed disaster response and patrols. These systems have high resolution aerial images that preserve intricate scenes with countless minor objects in diverse environmental conditions. Recent progress in deep learning has greatly boosted the object detection capabilities of natural images. But one of the peculiarities of aerial imagery is the introduction of individual difficulties that have not previously been faced in standard data. Objects tend to be photographed in small scaled, greatly overlapping and can be found in thick spatial patterns. There is also the factor of occlusion, motion blur and perspective distortion that makes proper detection even more difficult.

Traditional object detection models are designed to recognize and localize the instances of objects. Although this formulation performs well in general, in the case of aerial images, it is no longer trustworthy because the density of objects and the size of these surfaces is extreme. An even greater number of real-world applications do not seek to identify objects in particular, but to discover areas of great activity. An example is that a detection of a group of vehicles can make more actionable information than individual vehicle detection. This fuels the necessity of a different formulation which dwells upon defining spatial clusters of objects but not a single object.

In this study, a formulation of aerial imagery cluster-level detection is proposed. The model does not detect any individual objects but rather detects spatial clusters indicating dense clusters of objects around the object. A cluster is represented by a single bounding box and a single class label, and the problem represents a class-agnostic that is single-class detection task. In order to research the effectiveness of this formulation, the study assesses the effect of varying layers of features pyramids on the performance of cluster detection.

The main aim of this research works is to examine how feature pyramid neck designs influence cluster-based detection of aerial images. It is done by a controlled architecture ablation study of two configurations in the EfficientDet framework :Single-pass feature pyramid With a configuration similar to a typical Feature Pyramid Network (FPN) and A stacked Bidirectional Feature Pyramid Network (BiFPN) enabling iterative multi-scale feature fusion. The study concentrates on observation at cluster level of aerial images on a single-class formulation. In the analysis, feature pyramid neck architectures are only evaluated in a controlled experimental environment.

II. LITERATURE REVIEW

Detection of objects has tremendously changed due to the introduction of deep learning methods. The traditional techniques employed handcrafted features, but the recent techniques are the localization and extraction of features using convolutional neural networks. Extensive surveys, including but not limited to, [1, 2] reflect the shift towards deep-learning-based models, such as region-based and single-stage detection models.

Aerial imagery presents unique challenges compared to ground-level vision tasks. The objects are smaller in size, tightly packed and there are big differences in scale and orientation. The research by the authors of [?, 3] includes a comprehensive analysis of the areas in which deep learning is being deployed powerfully in aerial object detection, analyzing such concerns as collisions, sparse image quality, and under-representation. Studies by Elhagry et al. [4] specifically point to the effect of scale change and imbalance in the classes on detection performance in aerial scene. These difficulties drive the desire of better representation of features and multi-scale methods of processing.

The approach to solving the issue of identifying objects across scales was proposed by Feature Pyramid Networks (FPN) [?]. FPN is able to combine both low-level spatial features and highlevel semantic features, thereby facilitating the detection of objects of different sizes. EfficientDet is a further improvement of this concept [5] that adds a scalable architecture, a Bidirectional Feature Pyramid Network (BiFPN). BiFPN increases the feature fusion through twoway sharing of information across the various scales and weighted aggregation of features.

Various architectures were suggested to use on aerial object detection. LR-CNN to the aerial vehicle detection task is enhanced by adding the regional proposal, which is local-aware, to better detect vehicles in the imagery [6]. Likewise, the SFTN [7] aims at making the detection efficient and accurate in air conditions. Aerial object models like ReDet [8] uses rotation-awareness to handle changes in object orientation by use of rotation equals representations. More modern methods include NATCA YOLO [9], which is a lightweight object detector in first place, enabling better performance by improving its architecture and refining the features of the main object, and ST-YOLO [10], that is a set of strategies aimed at enhancing the performance of a small object detector, developed on the architecture of the original one, and reinforcing it with refined features of the subject of interest, to begin with. Among the most popular benchmarks that are used in aerial object detection, there is the VisDrone dataset as of now [11]. It offers a massive dataset of drone-aerial images in a variety of settings, such as crossroads of various cities, highways, and residential neighborhoods. The data set is fully annotated with instance-level data on several types of objects. Nevertheless, the majority of available studies employ such annotations to detect the objects of multiple classes.

Although great progress had been made in detecting aerial objects, the contemporary methods focus primarily on detection on instances. Little effort has been made on alternative formulations that are based on identifying clusters of objects and not individual objects. Identifying dense regions of activity is in some way more applicable than objectidentification, which is a potential solution offered by cluster-level

detection to this situation. Nevertheless, scanty literature investigates the performance of modern feature pyramid architectures in such a formulation. Moreover, although BiFPN was confirmed to improve its performance in standard detection tasks, its performance on cluster-level aerial detection has not been thoroughly studied. That is the reason why this study fills these gaps with the introduction of the cluster-based detection formulation and a controlled comparison between FPN-style and BiFPN-based feature aggregation strategies.

III. RESEARCH METHODOLOGY

3.1 Overview

This study deals with the effects of the feature pyramid neck designs on cluster level detection in an aerial image. It is a controlled architectural ablation study where there is a variation in the feature aggregation constituent of the detection pipeline when all the other constituents are fixed. The detection system has been developed on EfficientDet and we test two configurations; a baseline which is the one that approximates a Feature Pyramid Network (FPN) and a Bidirectional Feature Pyramid Network (BiFPN). This experimental design allows determining that any differences in performance may be assigned to the feature fusion mechanism itself.

3.2 Problem Formulation

Conventional object detectors seek to determine and locate individual examples of objects that fall into a set of established semantic categories. Nonetheless, aerial imagery poses a range of issues such as minute object size, high density of object, and serious occlusions, which decrease the accuracy of instance-level forecasts. To overcome these shortcomings, detection problem is re-defined as a class-agnostic cluster detection problem. Instead of detecting individual objects, the model identifies spatial regions containing dense groupings of objects, referred to as clusters.

3.3 Dataset and Annotation

The dataset that will be utilized in the current research will be based on the VisDrone dataset, described in the case of the drone-mounted cameras and their capture of aerial images that belong to various urban settings [11]. The objects on the dataset are highly diverse in terms of density, viewpoint, lighting, and complex layers of the scene. Examples of scenes are road intersections, car parks, construction sites, crowds of people where cars and people tend to be clustered in large spatial groupings. The dataset is appropriate in assessing detection models in realistic aerial settings because of this diversity. But VisDrone annotations are originally formulated as an instance-level, multi-class object detector which cannot be simply applied to the cluster detection formulation used in the present study.

Figure 1: Cluster-level annotation in aerial imagery, where multiple objects are grouped into a single bounding box.

All the images undergo a process of manual re-annotating in order to bring the dataset in line with the goals of the cluster detection. A cluster can be defined as an area of space grouping of objects that seem to be relatively close together in the image. The purpose of each cluster is to be annotated by one bounding box that covers the entire portion of objects in the cluster. It simplifies the work of annotation but at the same time, maintains geographical data associated with the dense area. The Roboflow platform is used to create annotations in a Pascal VOC form. Although this procedure certainly adds subjectivity to the process, it allows to create a dataset that is directly adapted to the cluster detection problem.

3.4 Model Architecture3.4.1 Backbone Network

The backbone network is EfficientNet-D0. It derives hierarchical feature representations at varying spatial resolutions with shallow layers representing finer spatial features to deep layers describing semantic features.

3.4.2 Feature Pyramid Neck

Multiscale features are clumped together at the feature pyramid neck which is the key point of interest in this study.

FPN-style Baseline

The default setting involves the use of one layer of BiFPN, which is close to a standard feature pyramid network. This is a single-pass operation on scales which combines features.

BiFPN Configuration

In the BiFPN set up, two layers of BiFPN are stacked, and this allows feature fusion on both directions. This enables top-down and bottom-up propagation of information and thus resulting in an iterative process to refine the feature representations.

BiFPN also adds learnable feature weights and thus the network can put more emphasis on more informative levels of features.

3.4.3 Detection Head

The detection head is made up of two subnetworks:

- Classification Head: Predicts the probability of the cluster with the help of Focal Loss.
- Regression Head: Out prediction in Smooth L1 loss of bounding box coordinates.

The subnetworks are both computationalally efficient with depthwise separable convolutions.

3.5 Training Procedure

The resizing of all images to the constant resolution of 512 × 512 pixels is carried out, and then the images are processed by the model. Besides resizing, geometrical data augmentation methods are used to enhance resistivity to spatial changes which are typically frequent in aerial imagery.

The augmentations that are employed include:

1 Horizontal Flip

2 Vertical Flip

3 Rotation (90° increments): clockwise, counter-clockwise, and upside-down

These augmentations add the effective diversity of the dataset by producing several variants of orientation of an image. This fact is especially critical in aerial imagery where the objects orientation cannot be fixed and can vacuate greatly across the scenes.

Standard ImageNet normalization is applied to all input images.

3.7 Evaluation Pipeline and Evaluation Metrics

The evaluation pipeline consists of:

- Conversion of Pascal VOC annotations to COCO format
- Model inference
- Conversion of predictions to COCO format
- Evaluation using COCO metrics

Primary Metric

The main measure of evaluation is:

$$mAP@[0.50 : 0.95]$$

Secondary Metric

[1] Precision: Ratio of correct detections

[2] Recall: Detection coverage

[3] F1 Score: Balance between precision and recall

Computational Cost

Computational complexity is accomplished with respect to FLOPs. The calculation of FLOPs is done by calculating the fvcore library with a dummy input size 512 × 512.

IV. RESULTS AND DISCUSSION

The experiments in this report were done in the form of controlled-ablation in which the only part affected was the neck architecture. This block is a short, reproducible summary of the training and evaluation configuration that was applied in the Colab experiments (raw data of logs of the experiments and notebooks).

- Dataset: VisDrone-derived cluster annotations (6018 images; Pascal VOC for training, converted to COCO for eval-uation).
- Model: EfficientDet-D0 (EfficientNet-D0 backbone), single-class detection ("cluster").
- Input: 512 × 512 pixels.
- Optimizer: AdamW.
- Learning rate: 1×10^{-4}.
- Weight decay: 1×10^{-4}.
- Seed: 42.
- Training: evaluated at 20 epochs (both variants).
- Controlled variable: numbifpnlayers (1 = FPN-style baseline; 2 = BiFPN).
- Evaluation: COCO mAP (mAP@[0.50:0.95]) and AR@[1,10,100]; FLOPs measured with fvcore FlopCountAnaly-sis.

The table below reports the principal accuracy and computational costs measures calculated based on the evaluation artifacts generated on Google Colab (T4). These values are extracted out of the experiment logs, and assessment results that are inscribed in the experiment scenario. Complete provenance and logs can be found in the file with the experiment extract that is availed along with this project.

Table 1: Accuracy-cost comparison evaluated on VisDrone-derived cluster annotations (6018 images).

Model	mAP@[0.50:0.95]	AR@100	FLOPs (GFLOPs)
FPN (EfficientDet-D0, num bifpn layers=1)	0.000	0.009	2.291
BiFPN (EfficientDet-D0, num bifpn layers=2)	0.000	0.012	2.291

Interpretation

At such an EfficientDet-D0 scale and parameter of dataset/annotation the controlled ablation exhibits no improvement in the strict metric of COCO mAP of BiFPN over the FPN-like baseline. BiFPN did not have a significant difference with with respect to AR@100; it was slightly higher (0.012 vs 0.009) indicating that there was a slightly higher level of coarse-area sensitivity but not accurate localization.

V. CONCLUSION

This study was a controlled architectural ablation work conducted to compare the relevance of feature pyramid neck design with regard to the presence of class-agnostic cluster detection on aerial imagery. The paper re-posed the detection challenge as an instance-level recognition task to a cluster-level localization challenge, in which spatially concentrated assemblage of objects is characterized by single bounding boxes.

The main goal was to separate and examine the performance of the neck constituent as a part of the EfficientDet system. The research kept the same configurations of the backbone, detection head, training

procedure, and dataset in the study; therefore, the study was able to identify the performance variation as due only to the feature aggregation strategy.

According to the experimental reports that were done on EfficientDet-D0 and had 20 training epochs, the addition of iterative bidirectional feature fusion in the form of BiFPN to the model did not do by any means increase detection accuracy as compared to the baseline provided by FPN. In particular, no relevant differences appeared in the mAP of [0.50:0.95], which indicated that the extra complexity of BiFPN was not reflected in the enhanced localization performance in the described environment.

Nevertheless, there was a small recall (AR@100) improvement, which suggests that BiFPN can enhance the capacity of the model to detect possible cluster areas, though not with enough accuracy to have any effect on overall mAP. This implies that although bidirectional feature fusion improves feature propagation it can take a more closely supervised or structurally compatible form to be changed into a corresponding accuracy improvement.

From the perspective of computation, it can be noted that both models have similar FLOPs at this scale, further supporting the notion that architectural sophistication does not necessarily improve performance. This highlights the need for data quality, consistency, and task-specific design considerations over architectural improvements.

In general, research indicates that in the context of cluster-level aerial detection, interaction between model architecture, annotation strategy and dataset characteristics is highly important to the success of multi-scale feature aggregation.

VI. FUTURE SCOPE

Although the current work defines the baseline of the cluster-level detection of aerial images, multiple ways may be considered to enhance the performance levels and extend the boundary of the research.

First, the generalization of the model is likely to be enhanced by increasing the dataset size by adding more images and aerial scenes that are more diverse.

Second, it is possible to improve annotation methodology which can be relevant in promoting detection performance to a high degree. Semi-automated clustering methods could further refine the boundary of clusters to increase their consistency.

Third, the existing formulation is to consider all clusters as a represented class. This can be expanded to class-specific cluster detection in the future with clusters being grouped by the dominant types of objects which could be vehicle clusters, pedestrian clusters or mixed-object clusters. This would allow more scene understanding with the advantages of clusterlevel abstraction.

Also, other methods of architecture could be considered. These are anchor-free detection, transformer-based detectors and hybrid models comprising the detection with segmentation to gain better understanding at the region level.

Lastly, future studies can integrate other assessment measures other than mAP and FLOPs e.g. cluster consistency measures or application-specific performance metrics especially in the context of application e.g. traffic monitoring or crowd analysis.

These guidelines offer a way through which the accuracy and applicability of cluster detection models can be enhanced in a real-life aerial imaging situation.

References

[1] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, “A Survey of Deep Learning-Based Object Detection,” *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[2] S. S. Zaidi, M. Riaz, and S. A. Khan, “A Survey of Modern Object Detection Literature Using Deep Learning,” *Electronics*, vol. 10, no. 22, p. 2860, 2021.

[3] D. Cazzato, C. Cimarelli, S. Pini, and R. Cucchiara, "Computer Vision for UAVs: A Survey," *ACM Computing Surveys*, vol. 52, no. 6, pp. 1–36, 2020.

[4] T. Nguyen, A. Nguyen, and H. Tran, "Detecting Objects in Aerial Images Using Deep Learning," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 1–8, 2020.

[5] M. Elhagry *et al.*, "Challenges of Object Detection in Aerial Images and Deep Learning-Based Solutions," *Remote Sensing*, vol. 14, no. 3, pp. 1–20, 2022.

[6] M. Liao, B. Shi, and X. Bai, "Vehicle Detection in Aerial Images Using Local Region-Based CNN," in *Proc. IEEE Conf.*, pp. 1–8, 2020.

[7] W. Chen *et al.*, "SFTN: A Lightweight Network for Efficient Aerial Object Detection," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[8] J. Han, J. Ding, J. Li, and G.-S. Xia, "ReDet: A Rotation-Equivariant Detector for Aerial Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 2786–2795, 2021.

[9] Y. Zhu *et al.*, "NATCA YOLO: Improving Small Object Detection in Aerial Images," *IEEE Access*, 2024.

[10] L. Yan *et al.*, "ST-YOLO: Spatio-Temporal YOLO for UAV-Based Object Detection," *IEEE Transactions*, 2025.

[11] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object′ Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2117–2125, 2017.

[12] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 10781–10790, 2020.

[13] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and Tracking Meet Drones Challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.