YOLOv7E: An Attention-Based Improved YOLOv7 for the Detection of Unmanned Aerial Vehicles

Dapinder Kaur[1,2], Neeraj Battish[1], Arnav Bhavsar[3] and Shashi Poddar[1,2]
[1]*CSIR – Central Scientific Instruments Organisation, Sector 30C, Chandigarh 160030, India*
[2]*Academy of Scientific & Innovative Research (AcSIR), Ghaziabad 201002, India*
[3]*IIT Mandi, Himachal Pradesh 175005, India*

Keywords: Deep Learning, UAVs, YOLOv7, Attention Modeling, Air-to-Air Object Detection.

Abstract: The detection of Unmanned Aerial Vehicles (UAVs) is a special case for object detection, specifically in the case of air-to-air scenarios with complex backgrounds. The proliferated use of UAVs in commercial, non-commercial, and defense applications has raised concerns regarding their unauthorized usage and mishandling in certain instances. Deep learning-based architectures developed recently to deal with this challenge could detect UAVs very efficiently in different backgrounds. However, the problem of detecting UAVs in complex background environments need further improvement and has been catered here by incorporating an attention mechanism in the YOLOv7 architecture, which considers channel and spatial attention. The proposed model is trained with the DeTFly dataset, and its performance has been evaluated in terms of detection rate, precision, and mean average precision values. The experimental results present the effectiveness of the proposed YOLOv7E architecture for detecting UAVs in aerial scenarios.

1 INTRODUCTION

Unmanned aerial vehicles (UAVs) have grown in popularity in recent decades due to their vast applications in agriculture, defense, surveillance, healthcare, etc. With the tremendous growth in commercial UAVs, there is also a parallel growth of malicious UAVs, which can carry explosive payloads, capture audio-visual data from a restricted private area, and enter the non-flying zone. Several incidents have been reported wherein malicious UAVs have entered restricted and no-flying zones, creating panic and security threats to the life and strategic infrastructure. These no-fly zones need to have a mechanism by which they can monitor an unidentified flying object. As a result, the development of anti-drone systems is gaining pace worldwide, and the problem of real-time drone detection is becoming more relevant (Seidaliyeva et al., 2020). Vision-based UAV detection is one of the primary elements for any anti-drone system and requires highly accurate drone detection in different scenarios. The problem of UAV detection varies depending on the perspective from which the images are captured, from the ground or from another aerial platform. The detection of UAVs from another aerial platform is far more complex as compared to the images captured from static cameras mounted on the ground to capture UAVs flying in the air due to the diverse views, angles, motion, and complex backgrounds.

With the advent of deep learning techniques, the traditional image processing-based techniques for drone detection have been replaced with more efficient architectures. Several deep learning architectures have been designed, including the region-based convolutional neural network (R-CNN), faster R-CNN, YOLOv3, and YOLOv5. However, these models do not handle key challenges such as moving cameras and detecting distant drones, and their performance is still not at par. Some of the other frameworks use a combination of visible & acoustic data (Jamil et al., 2020), visual & thermal data (Y. Wang et al., 2019), and visible, thermal, & acoustic data (Svanstrom et al., 2021) for the detection of drones.

In the deep learning architectures, several attention-based methods have gained importance and are used for various applications, including object detection (Li et al., 2022). Attention-based methods focus on specific input sections by assigning weights to the input elements. There are two different kinds of attention mechanisms: channel attention and spatial

Kaur, D., Battish, N., Bhavsar, A. and Poddar, S.
YOLOv7E: An Attention-Based Improved YOLOv7 for the Detection of Unmanned Aerial Vehicles.
DOI: 10.5220/0012391500003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 344-351
ISBN: 978-989-758-684-2; ISSN: 2184-4313

attention. Channel attention concentrates on global features, whereas spatial attention focuses on local features (Y. Zhang et al., 2019). Further, spatial attention identifies the key features from the feature maps, and channel attention, on the other hand, increases the feature maps for model learning. Utilizing the inter-spatial relationship between components, spatial attention creates a map that focuses on the locations of essential parts, acting as a supplement to the channel attention (Woo et al., 2018). In this way, channel and spatial attention boost the performance of several different deep learning architectures (Khan et al., 2020).

In this paper, the main contribution is incorporating an attention-based backbone into the YOLOv7 architecture for air-to-air (A2A) UAV detection and investigating its performance on publicly available datasets. The contributions of this work are as follows:

- YOLOv7E: an improved YOLOv7 for micro-UAV detection in several complex environments of the A2A object detection scenario.
- Feature enhancement by adding an EPSA attention module in the backbone of the YOLOv7 network, which can extract and enhance the multi-scale feature representation and adaptively recalibrate the channel-wise attention.
- Testing and analysis of the proposed YOLOv7E for A2A UAV detection tasks to compare with YOLOv7 with a variety of metrics.

Intending to detect UAV from a UAV, this paper further discusses the literature on the problem of UAV detection followed by a methodology section, in which firstly, the base model YOLOv7 is discussed to understand the current architecture of the model, the attention-based approach to improve this architecture is deliberated in detail, followed by the proposed YOLOv7E framework. The experimentation section describes and ensures the proposed model's effectiveness on the task of aerial object detection, and finally, the paper is concluded.

2 RELATED WORK

The problem of UAV detection in the A2A scenario is relatively less explored using computer vision approaches, and has, therefore, a limited number of research articles in this direction. The literature review here focuses on vision-based UAV to UAV detection using deep learning architectures alone. The large-scale A2A micro-UAV dataset generated by (Zheng et al., 2021), named the DetFly dataset, has over 13000 images of micro-UAVs captured by another UAV. For the UAV detection task, different deep learning architecture has been trained where grid RCNN achieved the overall highest average precision (AP) and performed well in challenging situations like intense/ weak lighting and motion blurring.

You Only Look Once (YOLO), a convolutional neural network-based model, has been popular because of its fast and precise performance in the literature. (Dadboud et al., 2021) proposed a YOLOv5-based air-to-air object detection and used the DetFly and other UAV datasets for experimentation. The YOLOv5 model was compared with the Faster R-CNN and FPN models, which achieved better detection results. Furthermore, the improved version of YOLO, that is, YOLOR was proposed by (Kizilay & Aydin, 2022), which utilized the lower layers of information called attribute information and improved the performance of UAV detection in A2A scenarios with highest mAP. (Leong et al., 2021), utilized the YOLOv3-tiny model to detect UAVs and estimated their velocity using a filtering framework. Furthermore, (Gonzalez et al., 2021) used YOLOv3 tiny for short and long-range-UAV detection in the A2A environment. Another deep learning model that utilized the backbone of YOLOv5, i.e., CSP-Darknet53 with video swin model was proposed by (Sangam et al., 2022). It utilized spatiotemporal swin transform and generated attention-based features to improve the performance of UAV detection in UAV-to-UAV detection problems.

The above literature survey indicates that deep learning-based architectures facilitate the problem of UAV-to-UAV detection and contribute by offering fast and robust performance. In comparison to two-stage detectors, single-stage detectors, like YOLO-based models, showcase their effectiveness in terms of precision and accuracy. However, the performance is inadequate when the scenario is unclear, including dynamic and complex backgrounds. Furthermore, the recent development in the YOLO-based models claims their better performance for several object detection tasks. In order to conduct a more detailed analysis, this work utilizes YOLOv7 and proposes an attention-based model named YOLOv7E to improve the performance of UAV detection.

As discussed earlier, attention-based models recalibrate the weights of features to improve performance. Like, YOLOv7 is enhance further using a convolutional block attention module that adds multi-scaling for object detection (J. Chen et al.,

2022). This module integrates channel and spatial attention but cannot establish a channel dependency for long ranges. To deal with this, further a global attention module was introduced in YOLOv7 (K. Liu et al., 2023), where the channel attention module is a 3-dimensional module, which amplifies the inter-dimensional dependencies using a multi-layer perceptron. In this, the spatial attention module has two convolution layers to extract the significant regions from the image. However, it increases the complexity of the model, so a simple attention module could be the best fit for YOLOv7 due to its own complex architecture. In this work, a lightweight YOLOv7E architecture is proposed that incorporates Efficient Pyramid Squeeze Attention (EPSA) module in the YOLOv7 architecture to reduce the overall complexity and improve the detection performance.

3 METHODOLOGY

The main focus of the proposed work is to improve the performance of the UAV detection framework using an attention-based mechanism in YOLOv7 architecture. YOLO-based single-stage detectors consider the object detection problem as a regression problem and do not include the region proposal stage, making them faster than the two-stage detectors like RCNN, fast RCNN, Faster RCNN, etc. YOLOv7 is the recent addition to the YOLO models, and literature claims its better performance than other existing YOLO models and two-stage detectors (C.-Y. Wang et al., 2022). This section deliberates on the existing YOLOv7 architecture, the attention based EPSA approach, and the proposed YOLOv7E framework.

3.1 YOLOv7

YOLOv7 is a recent version of YOLO-based architectures for object detection problems and outperforms previous versions in terms of speed and accuracy. This model includes an extended efficient layer aggregation network (E-ELAN), model scaling, RepConvN, and coarse to fine lead head guided label assigner (C.-Y. Wang et al., 2022). Here, E-ELAN focuses on improving the network's backbone by continuously elevating its learning capabilities. It introduces three different operations in the architecture: (i) expansion, (ii) shuffling, and (iii) merging to increase the cardinality of features without changing its gradient propagation path. Group convolutions are used in each computational block with the same group and channel multiplier parameters for expansion. Shuffling is performed on the computed feature maps and is divided into different groups using the same defined parameters. These feature maps are then concatenated to generate the same number of channels as the original. Finally, the merging operation is carried out by adding all the groups to their feature maps and enhancing the feature maps as a result. The feature maps extracted from these computational blocks are passed through the SPPCSPC (Spatial Pyramid Pooling and Cross Stage Partial Channel) module which increases the model's receptive field.

Furthermore, a compound model scaling is proposed in YOLOv7 due to its concatenation-based architecture. As model scaling is used to adjust different attributes of the model and to generate it on different scales, it decreases or increases the computational blocks in the case of concatenation-based architecture. Hence, to deal with this issue, a compound scaling first scales the depth factor and computes changes in the output channels. Then, it uses the same amount of change for width factor scaling and maintains the properties of the model to retain its optimal structure. Other than this, it was identified that the identity connection of the RepConv eradicates the residuals, so YOLOv7 replaces it with RepConvN, which does not include any identity connections and provides diverse gradients for distinct feature maps.

These extracted features aggregate the information only on three different scales using the neck module in a PANet architecture (S. Liu et al., 2018) and performed the final detections with the detection head. In object detection networks based on deep learning, some detection head utilizes deep supervision, while others work without it. With deep supervision, there are two types of heads: (i) the lead head, responsible for final output, and (ii) the auxiliary head, to assist training. Additionally, label assigners are used to assign soft labels using prediction results with ground truth labels. The issue of assigning soft labels to the lead or auxiliary head was unresolved until YOLOv7. In YOLOv7, a new label assigner is proposed which guides both the lead and auxiliary head based on the prediction of lead head, called a lead head guided label assigner. Besides, it is refined using coarse and fine labels for optimizing the final predictions and improving the overall ability of the model.

Though YOLOv7 is an advanced version that proposes E-ELAN in the computational blocks to enhance the feature maps, it is still struggling to detect small objects, specifically in complex backgrounds. In addition, its several applications for

different object detection tasks (Yang et al., 2022) and its improvements based on attention mechanisms (Zhao et al., 2023) provides space for further improvement.

3.2 Attention-Based Approach

Attention-based methods are widely used in computer vision applications, including object detection, classification, segmentation, and localization. Channel and spatial attention are the two methods that can improve the performance of deep neural networks. The most common method of channel attention is the squeeze and excitation (SE) module (Devassy & Antony, 2023), which assigns the weights to generate informative outcomes. It first squeezes the input using global average pooling (GAP) to encode the global information and then uses excitation to recalibrate the channel-wise relationship. It does not provide any importance to the spatial information, hence losing the feature information. In order to further improve the model, channel and spatial information were combined, and modules such as bottleneck attention module (X. Chen et al., 2023) and convolution block attention module (Jiang & Yin, 2023) were proposed. However, they fail to establish multi-scale and long-range channel dependencies and impose a burden on the models due to their heavy computations. An Efficient Pyramid Squeeze Attention (EPSA) module was proposed in (H. Zhang et al., 2021) to provide a low-cost, high-performance solution, as shown in Figure 1.

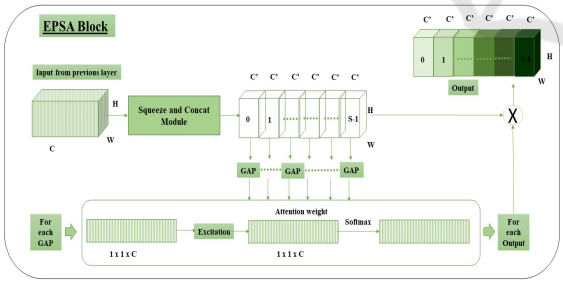

Figure 1: EPSA block.

This EPSA block consists of a squeeze and concat module, which extracts spatial information and obtains rich positional information by processing the input parallelly in multiple scales. Further, the channel attention of these multi-scale feature maps is extracted by the SE module, where each channel is passed through a GAP and excitation module. This fuses the different scale context information and produces pixel-level attention. Additionally, the softmax function is used to obtain the location information on the space and attention weights in the channels. These recalibrated weights are multiplied with the corresponding scales' feature maps to obtain local and global context information.

3.3 Proposed YOLOv7E

This work aims to effectively detect A2A objects (UAVs), even in a complex environment. Hence, it needs to learn the channel and spatial features, and their intrinsic relationships in the images. With this objective, an enhanced version of YOLOv7, called YOLOv7E, is proposed that uses an attention module to extract both the local and global context information by focussing on spatial and channel attention. Here, an EPSA attention block is introduced in the YOLOv7 backbone, which has a low-cost and high-performance pyramid squeeze attention (PSA) module that processes the input tensor at multiple scales and integrates the information for input feature maps. The spatial information from each channel feature map is extracted on different scales, precisely giving context features of all the neighboring scales. Further, the channel-wise attention weights are extracted for multi-scale feature maps, which help the network model to analyze and extract the relevant information. Additionally, to recalibrate the attention weights, a SoftMax operation is employed. The EPSA block in the proposed YOLOv7E backbone learns richer multi-scale feature representation and adaptively recalibrates the cross-dimension channel-wise attention. As shown in Figure 2, the EPSA block is added by replacing few convolutional layers of the YOLOv7.

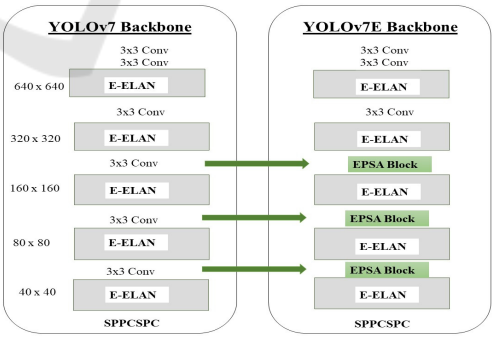

Figure 2: YOLOv7 and YOLOv7E backbone.

In YOLO-based models, feature extraction has a vital role in improving the detection accuracy, and therefore, by adding an EPSA block, multi-scale coarse spatial information is extracted with a long-range channel dependency. As shown in Figure 2, three EPSA blocks are added by replacing the 3 x 3

convolution block of the YOLOv7 backbone. This block processes the features of the E-ELAN module and enhances their representation without destroying the original feature map, providing more contextual information. These blocks are placed in the middle layers of the backbone to maintain both the low-level and the high-level information, which are aggregated in the neck module of the YOLO model.

The neck module of YOLOv7 uses a path aggregation network (PANeT) (C.-Y. Wang et al., 2022), a feature pyramid network (FPN) based structure, and has an additional bottom-up path aggregation. In YOLOv7E, the EPSA module integrates the multi-scale spatial information and cross-channel attention for each feature group and obtains the local and global channel interaction information, which is aggregated with the feature maps using element-wise sum, same as the PANeT. This multi-scale spatial information and the cross-channel attention adds robustness to the YOLOv7E backbone and improve its efficiency. The detection head of the model remains the same, which detects the object and localizes it on the image.

The YOLOv7 was not explicitly designed to address complex situations; however, this proposed YOLOv7E has the additional capabilities of EPSA to extract the multi-scale enhanced representations that add the ability in the proposed model to obtain the information of the objects even in complex environments and improve its performance.

4 EXPERIMENTATION AND ANALYSIS

The performance of the proposed YOLOv7E architecture is benchmarked with the YOLOv7 based on its detection performance, precision, and mAP. The experiments are carried out here on a window-based system with NVIDIA RTX A6000 GPU and 128 GB RAM capabilities. Moreover, python is used for implementation with different libraries. The other details related to parameters and experimentation are deliberated below.

4.1 Dataset

A2A object detection remains a relatively less explored research field and hence, the dataset availability is also limited. Among the available datasets, the DetFly dataset (Zheng et al., 2021) is the most recent available A2A UAV detection dataset, which contains 13271 images of micro-UAVs in different lighting conditions, background sceneries, relative distances, and viewing angles. The images are of high resolution and have a size 3840 x 2160 pixels. Experts annotate all the images, and their annotations are also available for research. Further, the dataset is evaluated for UAV detection using different object detection methods where Grid RCNN achieved a maximum average precision. The sample dataset images have different backgrounds, lighting conditions, view angles, etc. In the dataset, images are available in equal proportions for each type of view and background. Furthermore, the images of UAVs in direct sunlight environment from MIDGARD dataset (Walter et al., 2020) were also utilized for the analysis of YOLOv7E performance.

4.2 Training Parameters

The proposed model and base model are trained with 70% data, and the rest, 30% data, is used for testing purposes. The dataset of different views and scenarios is divided into equal proportions for both training and test data. The other training parameters, including its input size, optimizer, epochs, etc., are kept identical for both models and are given in the following table:

Table 1: Training Parameters.

Parameter	Values
Input Size	640
Batch Size	16
Optimizer	SGD
Learning rate	1e-2
Momentum	0.93
Weight Decay	5e-4
Epoch	500, 1000

4.3 Detection Results

The visual results of the detection using YOLOv7, and its improved version are presented in Figure 3. It shows that the base model missed some target UAVs in highly complex scenarios, whereas the proposed YOLOv7E detects them accurately. To quantify these detection results, the detection rate is also computed, and it is found that the proposed YOLOv7E has a relatively higher detection rate of 93% as compared to 92% for YOLOv7 model using DeTFly dataset. The detection rate is 89.8% and 93.9% using MIDGARD dataset with YOLOv7 and YOLOv7E, respectively. YOLOv7 has false detections where it detects random points from the image frame as a UAV, as shown in Figure 4. It is found that YOLOv7 has 1225 false detections, whereas this count is

reduced to 36% by the proposed YOLOv7E in DetFly dataset. Moreover, with MIDGARD data the false detection rate reduced by 21% using YOLOv7E where YOLOv7 has 443 false detections.

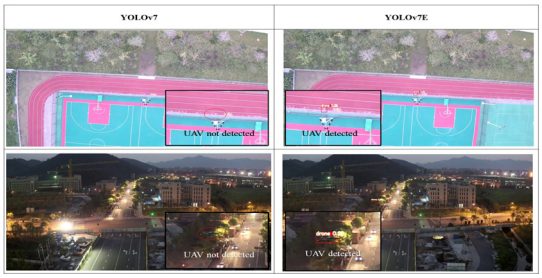

Figure 3: Detection Results (presenting missed targets in YOLOv7).

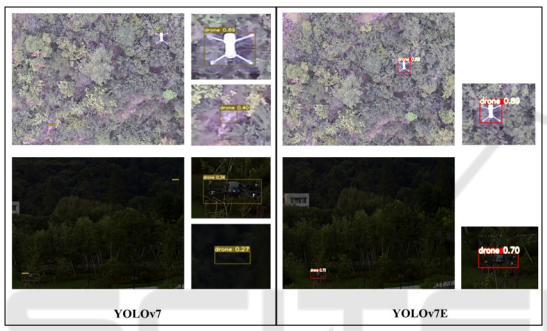

Figure 4: False detection by YOLOv7.

From the above results, the performance of the proposed YOLOv7E using an attention-based mechanism is found to be better in detecting UAVs in complex and diverse background environments. The visual results with MIDGARD datasets are available at: https://github.com/dapinderk-2408/YOLOv7E

4.4 Quantitative Performance Analysis

The performance of the proposed YOLOv7E is also computed in terms of average precision (AP), recall, mean average precision ($mAP^{0.5}$), and $mAP^{0.5\text{-}0.95}$ where $mAP^{0.5}$ is the mean AP with Intersection over Union (IoU) threshold 0.5 and $mAP^{0.5\text{-}0.95}$ takes the average values with the IoU thresholds from 0.5 to 0.95 with step size 0.05. Here, IoU is the ratio of area overlapped between the predicted box (P) and its ground truth (G) and is defined by:

$$IoU = \frac{P \cap G}{P \cup G} \tag{1}$$

The mAP with IoU threshold measures the correctness of the predicted bounding boxes. Moreover, precision and recall are defined based on the positive and negative predictions. The results given in the Table 2 presents the average precision, recall and mAP values on the overall dataset for both the YOLOv7 and YOLOv7E models.

Table 2: Performance Analysis.

Epochs	Model	Precision	Recall	$mAP^{0.5}$	$mAP^{0.5\text{-}0.95}$
500	YOLOv7	0.92	0.75	0.84	0.47
	YOLOv7E	**0.98**	0.89	0.92	0.59
1000	YOLOv7	0.97	0.86	0.91	0.60
	YOLOv7E	0.98	**0.89**	**0.92**	**0.61**

Table 2 signifies the performance of YOLO models based on different performance metrics. The proposed model performed better than the existing base model as the attention mechanism increases the model learning capabilities by providing weights to the features, thus improving performance. For performance evaluation, the models are trained with 500 and 1000 epochs. The loss and other parameters computed during training with 1000 epochs are shown in Figure 5.

YOLOv7 is a complex model, so it takes more time to learn. As shown in Figure 5, the performance of the proposed YOLOv7E in terms of box loss, objectness loss score, precision, recall, $mAP^{0.5}$, and $mAP^{0.5\text{-}0.95}$ improves continuously after 500 epochs and is better than YOLOv7. For instance, box loss and object loss converged more in YOLOv7E as compared to YOLOv7. Similarly, precision, recall, $mAP^{0.5}$, and $mAP^{0.5\text{-}0.95}$ value increases and is higher than YOLOv7. Here, the box loss depicts the algorithm's performance in locating an object's centre and how well an object is covered by the anticipated bounding box and is calculated as:

$$L_{box} = mean(L_{CIoU}). \tag{2}$$

Here L_{CIoU} is the complete IoU loss that uses different geometric factors such as aspect ratio, overlap area and the central point distance, and is computed as:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(p, p^G)}{c^2} + \alpha v \tag{3}$$

p and p^G are the central points of the predicted bounding box (P) and ground truth (G), respectively. ρ(.) presented the Euclidean distance between the central points, c is the diagonal length of the box, v & α represents the discrepancy measure of width-to-height ratio.

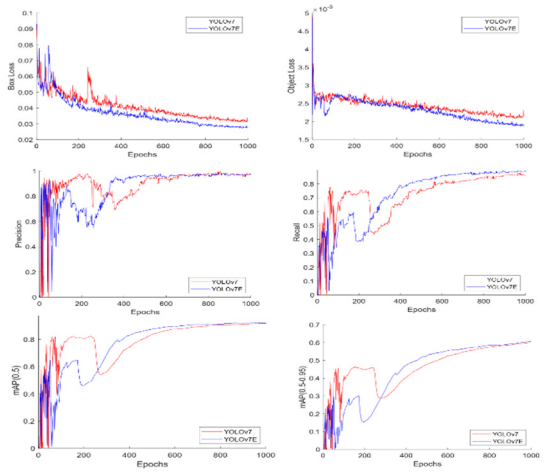

Figure 5: Performance Analysis of the proposed YOLOv7E (1000 epochs).

Further, objectness is a probability measure for the presence of an object in a suggested area of interest. An item is probably present in the image window if the objectivity is high. Further, in YOLO based models, the objectness loss score is computed using binary cross entropy (BCE) loss of the predicted objectness probability and CIoU of the matched target. The performance comparison given in Table 2 indicates the performance improvement with 1000 epochs. The performance of YOLOv7E is compared to YOLOv7 based on inference time as shown in Table 3.

Table 3: Inference Time per frame (in sec).

Dataset	YOLOv7	YOLOv7E
DeTFly	0.20	0.23
MIDGARD	0.034	0.036

Based on these results, it is found that the computational cost of YOLOV7E model is similar to YOLOv7, with improved performance ensuring effectiveness in the proposed field.

4.5 Comparison with Existing A2A UAV Detection Models

There are few detection models developed for A2A UAV detection problem in the past. The comparison of the proposed YOLOv7E with the existing A2A detection models (Zheng et al., 2021) based on AP is presented in the Table 4:

Table 4: Performance Comparison (DeTFly Dataset).

Model	Input size	Iterations/ epochs*	AP
YOLOv3	[416,416]	7000	72.3
SSD512	[512,512]	46564	78.7
FPN	[600,600]	49993	78.7
Cascade RCNN	[640,640]	6652	79.4
Grid RCNN	[600,600]	46564	82.4
YOLOv7	[640,640]	300*	84.17
YOLOv7E	[640,640]	300*	**91.2**

The above results indicate the effectiveness of the proposed YOLOv7E in terms of AP as it achieves highest AP in comparison to the other existing models.

5 CONCLUSION

The problem of UAV detection from other UAVs is a complex task, and this work proposes a YOLOv7E: an attention-based YOLOv7 model to achieve better performance than the existing YOLOv7 architecture. Attention-based approaches help to maintain both the spatial and channel information to maintain the object information in the deep layer architectures. In this work, a lightweight EPSA block that extracts the multi-scale spatial information with the essential features across dimensions in the channel attention is added to the backbone of the network. This addition helps in extracting the multi-scale coarse spatial information with a long-range channel dependency and maintains the complexity of the model. The proposed YOLOv7E is tested using the DeTFly dataset to analyze its performance for UAV detection in A2A complex scenarios. The performance in terms of detection rate, false detection, precision, recall, and mAP is improved by a proposed YOLOv7E compared to the base model of YOLOv7. Hence, the proposed method ensures the model's effectiveness for UAV detection tasks in A2A scenarios. Further, it can be tested for other object detection tasks to evaluate its efficiency.

ACKNOWLEDGEMENTS

The author would like to thank TiHAN-IITH for their support through project funding.

REFERENCES

Chen, J., Liu, H., Zhang, Y., Zhang, D., Ouyang, H., & Chen, X. (2022). A Multiscale Lightweight and

Efficient Model Based on YOLOv7: Applied to Citrus Orchard. *Plants*, *11*(23), 3260. https://doi.org/10.3390/plants11233260

Chen, X., Ma, W., Gao, W., & Fan, X. (2023). BAFNet: Bottleneck Attention Based Fusion Network for Sleep Apnea Detection. *IEEE Journal of Biomedical and Health Informatics*, 1–12. https://doi.org/10.1109/JBHI.2023.3278657

Dadboud, F., Patel, V., Mehta, V., Bolic, M., & Mantegh, I. (2021). Single-Stage UAV Detection and Classification with YOLOV5: Mosaic Data Augmentation and PANet. *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8. https://doi.org/10.1109/AVSS52988.2021.9663841

Devassy, B. R., & Antony, J. K. (2023). Histopathological image classification using CNN with squeeze and excitation networks based on hybrid squeezing. *Signal, Image and Video Processing*, *17*(7), 3613–3621. https://doi.org/10.1007/s11760-023-02587-y

Gonzalez, F., Caballero, R., Perez-Grau, F. J., & Viguria, A. (2021). Vision-based UAV Detection for Air-to-Air Neutralization. *2021 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, 236–241. https://doi.org/10.1109/SSRR53300.2021.9597861

Jamil, S., Fawad, Rahman, M., Ullah, A., Badnava, S., Forsat, M., & Mirjavadi, S. S. (2020). Malicious UAV Detection Using Integrated Audio and Visual Features for Public Safety Applications. *Sensors*, *20*(14), 3923. https://doi.org/10.3390/s20143923

Jiang, M., & Yin, S. (2023). Facial expression recognition based on convolutional block attention module and multi-feature fusion. *International Journal of Computational Vision and Robotics*, *13*(1), 21. https://doi.org/10.1504/IJCVR.2023.127298

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

Kizilay, E., & Aydin, I. (2022). A YOLOR Based Visual Detection of Amateur Drones. *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 1446–1449. https://doi.org/10.1109/DASA54658.2022.9765252

Leong, W. L., Wang, P., Huang, S., Ma, Z., Yang, H., Sun, J., Zhou, Y., Abdul Hamid, M. R., Srigrarom, S., & Teo, R. (2021). Vision-Based Sense and Avoid with Monocular Vision and Real-Time Object Detection for UAVs. *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, 1345–1354. https://doi.org/10.1109/ICUAS51884.2021.9476746

Li, Z., Wang, Y., Zhang, N., Zhang, Y., Zhao, Z., Xu, D., Ben, G., & Gao, Y. (2022). Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sensing*, *14*(10), 2385. https://doi.org/10.3390/rs14102385

Liu, K., Sun, Q., Sun, D., Yang, M., & Wang, N. (2023). *Underwater target detection based on improved YOLOv7*. http://arxiv.org/abs/2302.06939

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). *Path Aggregation Network for Instance Segmentation*. http://arxiv.org/abs/1803.01534

Sangam, T., Dave, I. R., Sultani, W., & Shah, M. (2022). *TransVisDrone: Spatio-Temporal Transformer for Vision-based Drone-to-Drone Detection in Aerial Videos*.

Seidaliyeva, U., Akhmetov, D., Ilipbayeva, L., & Matson, E. T. (2020). Real-Time and Accurate Drone Detection in a Video with a Static Background. *Sensors*, *20*(14), 3856. https://doi.org/10.3390/s20143856

Svanstrom, F., Englund, C., & Alonso-Fernandez, F. (2021). Real-Time Drone Detection and Tracking With Visible, Thermal and Acoustic Sensors. *2020 25th International Conference on Pattern Recognition (ICPR)*, 7265–7272. https://doi.org/10.1109/ICPR48806.2021.9413241

Walter, V., Vrba, M., & Saska, M. (2020). On training datasets for machine learning-based visual relative localization of micro-scale UAVs. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 10674–10680. https://doi.org/10.1109/ICRA40945.2020.9196947

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. https://doi.org/2207.02696

Wang, Y., Chen, Y., Choi, J., & Kuo, C.-C. J. (2019). Towards Visible and Thermal Drone Monitoring with Convolutional Neural Networks. *APSIPA Transactions on Signal and Information Processing*, *8*(1). https://doi.org/10.1017/ATSIP.2018.30

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). *CBAM: Convolutional Block Attention Module*. https://doi.org/1807.06521

Yang, F., Zhang, X., & Liu, B. (2022). *Video object tracking based on YOLOv7 and DeepSORT*. https://doi.org/2207.12202

Zhang, H., Zu, K., Lu, J., Zou, Y., & Meng, D. (2021). *EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network*. http://arxiv.org/abs/2105.14447

Zhang, Y., Fang, M., & Wang, N. (2019). Channel-spatial attention network for fewshot classification. *PLOS ONE*, *14*(12), e0225426. https://doi.org/10.1371/journal.pone.0225426

Zhao, H., Zhang, H., & Zhao, Y. (2023). YOLOv7-Sea: Object Detection of Maritime UAV Images Based on Improved YOLOv7. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 233–238.

Zheng, Y., Chen, Z., Lv, D., Li, Z., Lan, Z., & Zhao, S. (2021). Air-to-Air Visual Detection of Micro-UAVs: An Experimental Evaluation of Deep Learning. *IEEE Robotics and Automation Letters*, *6*(2), 1020–1027. https://doi.org/10.1109/LRA.2021.3056059.