School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ 85281, USA
{nyarava}@asu.edu

CBAM-YOLO FOR WAID DATASET

Nitin Reddy Yarava

ABSTRACT

This study investigates the potential of a novel architecture, CBAM-YOLO, for wildlife detection in drone imagery. Building upon the success of Squeeze-and-Excitation YOLO (SE-YOLO), this research explores the effectiveness of integrating the Convolutional Block Attention Module (CBAM) within the YOLOv7 framework. Due to limitations in replicating the original SE-YOLO architecture details, a direct comparison is not possible.

1 INTRODUCTION

Wildlife detection using drones equipped with high-resolution cameras plays a critical role in conservation efforts. However, accurate detection, particularly for small targets, remains a challenge. Deep learning approaches like SE-YOLO have demonstrated success by leveraging the Squeeze & Excitation channel attention mechanism.Mou et al. (2023); Hu et al. (2018)

This research investigates the potential of CBAM-YOLO, proposing and evaluating its effectiveness for wildlife detection in drone imagery. CBAM incorporates both channel and spatial attention mechanisms, potentially leading to superior performance compared to SE-YOLO, especially with the limitations in replicating the original architecture.

In Summary, this work aims to answer the following questions:

1. Can a novel architecture, CBAM-YOLO, achieve comparable or even surpass the state-of-the-art (SOTA) performance in wildlife detection using drone (WAID dataset), set by the previously reported SE-YOLO model?
2. Since the exact details of the SE-YOLO architecture are unavailable, can the CBAM-YOLO architecture, which integrates the Convolutional Block Attention Module (CBAM) into YOLOv7, serve as a viable alternative for wildlife detection?
3. How important are attention mechanisms, specifically the CBAM module, in enhancing wildlife detection accuracy, particularly for small targets in drone imagery?

2 METHODOLOGY

This new methodology builds upon the foundation laid out in the previous approach by exploring modifications within the model architecture itself.

3 DATASET

WAID Dataset: A High-Quality Resource for Wildlife Detection in Drone Imagery containing 14,375 UAV imagesMou et al. (2023). The WAID dataset serves as the foundation for training and evaluating the wildlife detection models in this research. This dataset is a valuable resource due to its high quality and diverse composition. It incorporates images from various sources like web scraping, publicly available drone footage, and pre-existing datasets. A meticulous standardization process ensures consistent image resolution and removes irrelevant content. Professionals

Table 1: The table details the number of wildlife instances in each class within the training, validation, and testing sets.

Class	Training	Validation	Testing	Total
Sheep	3602	349	173	4124
Cattle	3301	943	471	4715
Seal	2709	330	329	3368
Camelus	512	149	82	743
Zebra	443	126	65	643
Kiang	551	157	74	782

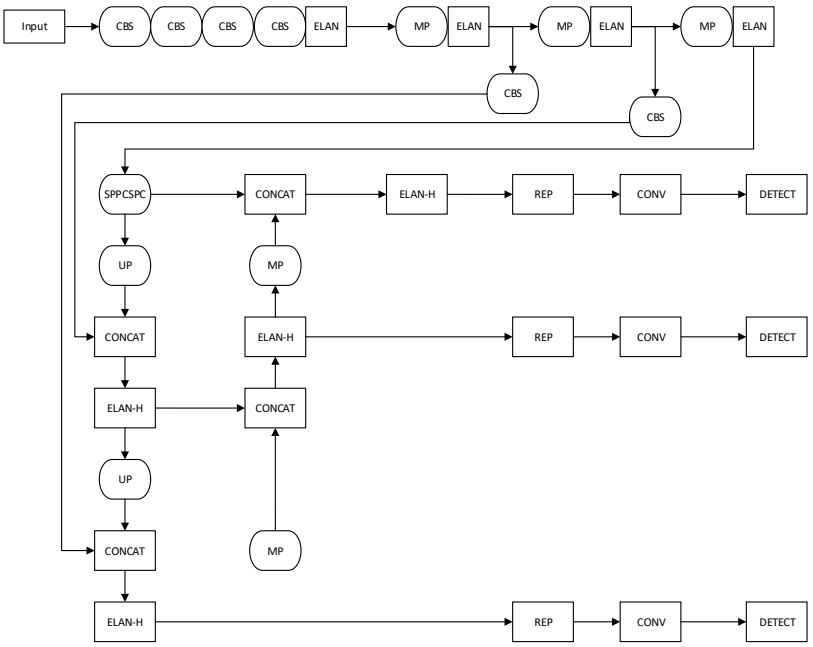

Figure 1: YOLOv7 Architecture

use Labeling to annotate the images with details like object position, size, and category. Rigorous quality assurance measures guarantee the accuracy and completeness of these annotations. Finally, the dataset is split into training, validation, and testing sets using a 70:20:10 ratio while maintaining category balance across all sets. This comprehensive dataset with its diverse wildlife categories (sheep, cattle, seal, camel, zebra, kiang) proves to be instrumental in advancing wildlife detection models for drone-based conservation efforts. Table 1 details the dataset composition of wildlife instances in each class.

3.1 Model Architecture

3.1.1 Convolutional Block Attention Module (CBAM)

CBAM, which stands for Convolutional Block Attention Module, is an attention mechanism designed to enhance feature representation in convolutional neural networks (CNNs).Woo et al. (2018)

Key Components of CBAM:

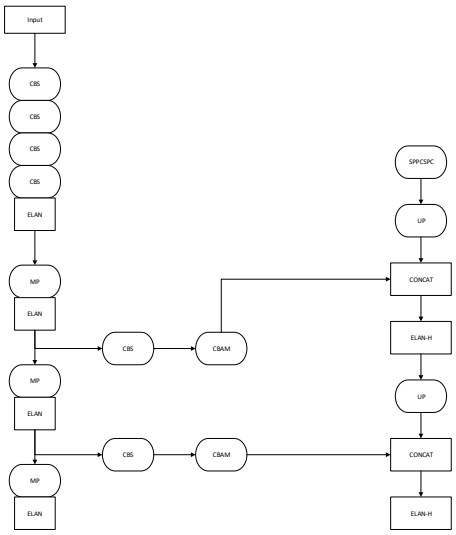

Figure 2: YOLOv7 + CBAM Architecture

- Channel Attention Module (CAM): The CAM focuses on inter-channel dependencies by computing the importance of each feature channel. It generates channel-wise attention maps by aggregating information across spatial dimensions.
- Spatial Attention Module (SAM): The SAM, on the other hand, attends to spatial relationships within feature maps. It captures dependencies between spatial locations by considering features across different channels.

CBAM integrates both the CAM and SAM components to enable the network to adaptively recalibrate feature maps based on channel-wise and spatial dependencies. By incorporating CBAM into CNN architectures, models can learn to selectively attend to informative features while suppressing irrelevant or noisy ones.

3.1.2 CBAM-YOLO

The CBAM-YOLO architecture builds upon the YOLOv7 object detection framework to improve wildlife detection in drone imagery.Wang et al. (2023); Lv & Su (2024); Jiang et al. (2022); Thakuria & Erkinbaev (2023) While the core feature extraction layers (backbone) remain unmodified to preserve their effectiveness in capturing essential details, the neck layer undergoes crucial changes. Here, CBAM blocks are strategically inserted as seen in Figure2. By incorporating CBAM blocks, the CBAM-YOLO architecture aims to achieve two key benefits: improved feature prioritization and reduced background noise. Feature prioritization allows the model to focus on informative details, especially for small targets with limited pixel representation. Additionally, background noise suppression by the CBAM blocks leads to cleaner feature representations for the detection head, potentially enhancing overall wildlife detection accuracy.

4 RESULTS

Table 2 displays a comparison of results between the Original architecture reproduced (YOLOv7) and the original architecture from the paper. The reproduced architecture achieved a higher mean Average Precision (mAP) of 0.976 compared to the original paper's results. However, there was a slight decrease in Precision, while Recall remained the same for both architectures.

Table 2: Results of YOLOv7

Architecture	*Precision*	*Recall*	*mAP*@0.5
YOLOv7 *(paper)*	**0.964**	0.953	0.972
YOLOv7 *(mine)*	0.963	0.953	**0.976**

Eight different configurations of CBAM-YOLO were evaluated, and the architecture presented in the figures and tables achieved the most promising results.

However, due to limitations in replicating the original SE-YOLO architecture, I devised my own variant, termed SE-YOLO(mine) in the table 3. Here, I replaced the CBAM blocks within the best-performing CBAM-YOLO architecture with SE (Squeeze-and-Excite) blocks. This approach enabled to compare between CBAM and SE Attention mechanism.

Table 3: Results of Modified Architecture

Architecture	*Precision*	*Recall*	*mAP*@0.5
SE-YOLO (paper)	**0.978**	**0.969**	**0.983**
SE-YOLO (mine)	0.94	0.93	0.956
CBAM-YOLO (mine)	0.948	0.92	0.954

5 DISCUSSION

While CBAM-YOLO performed well, it failed to surpass the performance of my SE-YOLO implementation (replaced CBAM blocks with SE blocks in the best-performing CBAM-YOLO architecture). This suggests that for this specific dataset, the SE attention mechanism might be more effective than the CBAM mechanism in enhancing wildlife detection. It is unfortunate that I was unable to replicate the original SE-YOLO architecture due to missing details. This makes a direct comparison between the reported SOTA (State-of-the-Art) performance and my implementation challenging.

Some of the reasons for CBAM-YOLO not performing as well could be:

1. The specific configurations or hyperparameters used in the original SE-YOLO may not be suited for CBAM-YOLO to achieve those top results.
2. CBAM blocks incorporate both channel-wise and spatial attention. While spatial attention can be helpful in some cases, it might introduce noise for small wildlife targets. By focusing solely on channels, in some cases, SE blocks might be more effective in directly highlighting relevant information for detection.
3. It is also possible that the WAID dataset's characteristics favor a channel-wise attention mechanism. The specific image quality, variations in target size and pose, or the types of wildlife present might make SE blocks a better fit for extracting the most relevant features.

These are just potential explanations, and further investigation is needed to definitively determine why SE outperformed CBAM in this case. While the CBAM mechanism might appear theoretically advantageous, the practical results from our experiments demonstrate that it doesn't necessarily translate to superior performance in real-world applications. This highlights the importance of evaluating approaches through experimentation, as theoretical promise may not always hold true in practice.

6 Conclusion

In conclusion, the research sought to enhance the performance of YOLOv7 for wildlife detection by integrating CBAM attention mechanisms, aiming to surpass the state-of-the-art performance established by SE-YOLO. Despite extensive experimentation, neither the CBAM-YOLO nor the modified SE-YOLO configurations managed to outperform the benchmark set by the original SE-YOLO. However, the replacement of CBAM with SE blocks in the best-performing CBAM-YOLO architecture did yield a variant that outperformed the other CBAM configurations, suggesting some superiority of SE blocks in this application.

The inability to replicate the original SE-YOLO architecture fully and precisely might have contributed significantly to the inability to achieve new state-of-the-art results. Future research could focus on further exploring the interplay of different attention mechanisms within YOLO architectures and other model frameworks, potentially looking at hybrid approaches that combine the strengths of both channel and spatial attention mechanisms.

References

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Kailin Jiang, Tianyu Xie, Rui Yan, Xi Wen, Danyang Li, Hongbo Jiang, Ning Jiang, Ling Feng, Xuliang Duan, and Jianjun Wang. An attention mechanism-improved yolov7 object detection algorithm for hemp duck count estimation. *Agriculture*, 12(10):1659, 2022.

Meng Lv and Wen-Hao Su. Yolov5-cbam-c3tr: an optimized model based on transformer module and attention mechanism for apple leaf disease detection. *Frontiers in Plant Science*, 14:1323301, 2024.

Chao Mou, Tengfei Liu, Chengcheng Zhu, and Xiaohui Cui. Waid: A large-scale dataset for wildlife detection with drones. *Applied Sciences*, 13(18):10397, 2023.

Angshuman Thakuria and Chyngyz Erkinbaev. Improving the network architecture of yolov7 to achieve real-time grading of canola based on kernel health. *Smart Agricultural Technology*, 5: 100300, 2023.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7464–7475, 2023.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.