JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 6 No. 1 (2025) 278-288
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Enhancing YOLO with Adversarial and Transfer Learning for UAV-Based Urban Wildlife Conservation

Thavavel Vaiyapuri[1]*, Elham Kariri[2], Ahmed Abdelrahman[1], Raya Aldawood[1]

[1] *Department of Computer Sciences, College of Computer Engineering and Science, Prince Sattam bin Abdulaziz, University, Al Kharj, SAUDI ARABIA*
[2] *Department of Information Systems, College of Computer Engineering and Science, Prince Sattam bin Abdulaziz, University, Al Kharj, SAUDI ARABIA*

*Corresponding Author: abbas.ahmed@spu.edu.iq
DOI: https://doi.org/10.30880/jscdm.2025.06.01.018

Article Info

Received: 28 October 2025
Accepted: 9 December 2025
Available online: 30 June 2025

Keywords

UN SDG, non-terrestrial networks, UAV imagery, YOLO, pretrained convolutional neural network, object detection, wildlife conservation

Abstract

Urban wildlife conservation has gained more importance worldwide, particularly in fast urbanizing regions like Kingdom of Saudi Arabia (KSA), where habitat fragmentation and biodiversity loss are major concerns. Unmanned aerial vehicle (UAV) imagery has emerged as a promising solution for wildlife monitoring, providing high-resolution, real-time data collection even in difficult-to-access urban areas. Nonetheless, the effective utilization of UAV images is seriously hampered by the metropolitan settings and interference from human activity. While deep learning (DL) models provide effective solutions for wildlife monitoring in UAV imagery, the scarcity of high-quality training datasets often limit their effectiveness. To the best of our knowledge, this is the first research to resolve these issues harnessing the combined power of adversarial and transfer learning. The research enhances the state-of-the-art YOLOv8 (You Only Look Once, version 8) model by integrating Generative Adversarial Networks (GANs) for data augmentation (DA) and EfficientNet-B3 for advanced feature extraction. Two research datasets namely, the Wildlife Animals Image Dataset (WAID) and the Animal Images Detection (AID) are used to evaluate the proposed model. Three comprehensive experimental analyses—DA, ablation, and cross-dataset validation (CDV) are carried out to prove its efficacy in urban wildlife monitoring. The results highlight the potential of the proposed model to considerably enhance detection accuracy and contribute to sustainable urban wildlife conservation efforts, which are consistent with the United Nations Sustainable Development Goals (UN SDGs).

1. Introduction

Urban wildlife conservation has become more important as urbanization transforms natural ecosystems into regions designed largely for human use [1]. Urban areas are typically connected with habitat degradation, but with the correct conservation methods in place, they may serve as critical biodiversity sanctuaries. Wildlife in urban areas contributes to biodiversity, human health, and environmental stability [2-3]. This adheres to the UN SDGs, notably Goal 11: Sustainable Cities and Communities, Goal 13: Climate Change, and Goal 15: Life on Land [4]. These Goals emphasize the significance of incorporating ecological preservation into urban development in order to create resilient, inclusive, and sustainable environments.

KSA has distinctive environmental challenges that make urban wildlife preservation more important in comparison to other countries [5]. Driven by Vision 2030's economic diversification initiatives, the Kingdom's fast urbanization and growth has led to notable habitat loss and fragmentation, endangering its great biodiversity including famous species such as the Arabian oryx and Arabian leopard [6]. Additionally, KSA's arid desert climate and scarcity of natural green spaces necessitate innovative strategies to create urban habitats that support wildlife and enhance ecosystem resilience. Vision 2030 emphasizes sustainable development and environmental protection, positioning urban wildlife conservation as a vital strategy for improving quality of life, mitigating climate change impacts such as rising temperatures and water scarcity, and promoting sustainable urban growth [4-6]. Similar to KSA, nations around the globe are tackling comparable challenges, highlighting the significance of incorporating urban wildlife conservation into broader sustainability initiatives.

Nonetheless, urban wildlife preservation poses distinctive challenges that vary from those faced in rural environments. Urban ecosystems are characterized by fragmented habitats, pollution, human-animal conflicts, and intricate infrastructure, which hampers the observation and conservation of wildlife [7]. Moreover, urban environments have logistical and legal limitations, like no-fly zones and restricted access, which further impede conservation initiatives. Unlike the expansive and stable habitats typical of rural areas, urban ecosystems are dynamic and heavily influenced by human activity, necessitating innovative approaches for effective conservation.

The aforementioned problems have lately prompted a plethora of solutions, including public awareness campaigns, green corridors, and habitat restoration. While conventional approaches set a standard, advancements in technology have significantly changed urban wildlife monitoring. Non-terrestrial networks (NTNs)—such as satellites, high-altitude platforms, and UAVs—are among the most effective tools for this purpose [8]. UAVs have become increasingly important as they are capable of generating high-resolution images, conducting low-altitude surveys, and providing on-demand data collection in urban areas that are spatially fragmented. UAVs are significant in observing urban biodiversity because of their flexibility and accuracy [9].

Integrating UAVs with DL algorithms improves their use in wildlife monitoring. YOLO is a popular deep learning framework for object recognition, valued for its real-time processing capabilities and impressive accuracy. YOLO's rapidity and efficacy facilitate the examination of extensive datasets, making it optimal for animal monitoring. However, issues like fluctuating illumination, occlusion from infrastructure, and a paucity of annotated urban wildlife datasets restrict YOLO's effectiveness in urban environments [7, 10].

This work combats these limitations by taking advantage of GANs for the first time to create realistic synthetic datasets. GANs enrich training datasets with multiple urban situations, allowing the YOLO model to be fine-tuned over a wide range of urban settings. Further, the proposed approach not only adopts transfer learning but also develops a variant of YOLO with the EfficientNet-B3, a well-known Google's Convolutional neural network (CNN) model, as its backbone. In this paper, efficient transfer learning on EfficientNet-B3 is used to improve the feature extraction performance of YOLO and promote detection accuracy and reliability in complicated urban driving scenarios. Transfer learning is applied in further optimizing the model. The two GAN-augmented study datasets on urban wildlife are used to fine-tuned the model, which leads to improved model generalization. Using cutting-edge features from the GAN and the EfficientNet-B3 model to fine-tune the YOLO, the proposed approach consistently makes correct detections.

To the best of the authors' knowledge, this work is the first to address the difficulties raised by urban wildlife monitoring by leveraging the benefits of transfer learning and adversarial learning, contributing to the domain with the following significant contributions:

- The study explores the impact of utilizing EfficientNet-B3 as the backbone for feature extraction within the YOLOv8 network, highlighting improvements in speed, accuracy, and overall performance for UAV-based wildlife monitoring.
- This study takes the benefits of transfer learning to train the developed YOLOv8 variant on the COCO (Common Objects in Context) dataset to build a robust pretrained model. The pretrained YOLOv8 model is then fine-tuned to improve its performance for wildlife monitoring using UAV imagery.
- This study is the first to use GANs for wildlife conservation, producing a diversified training dataset to overcome insufficient data. This enables the improved YOLOv8 model to generalize more well across various animals and urban areas.
- These innovations address critical issues with data availability, model efficiency, and detection accuracy, representing a significant step forward for urban wildlife monitoring.

1. Research Methodology

In this part of the article, the proposed model is introduced. As seen in Fig.1, it starts by creating varied training datasets adopting adversarial learning using GAN. Then, to improve the speed and accuracy of real-time urban wildlife monitoring, the proposed model uses EfficientNet-B3 as the YOLOv8 backbone network and applies transfer learning. The dual strategy used here efficiently overcomes the difficulties of data scarcity and

unpredictability, allowing for more accurate and efficient monitoring of wildlife in urban contexts. The next part describes the essential components of the proposed model.

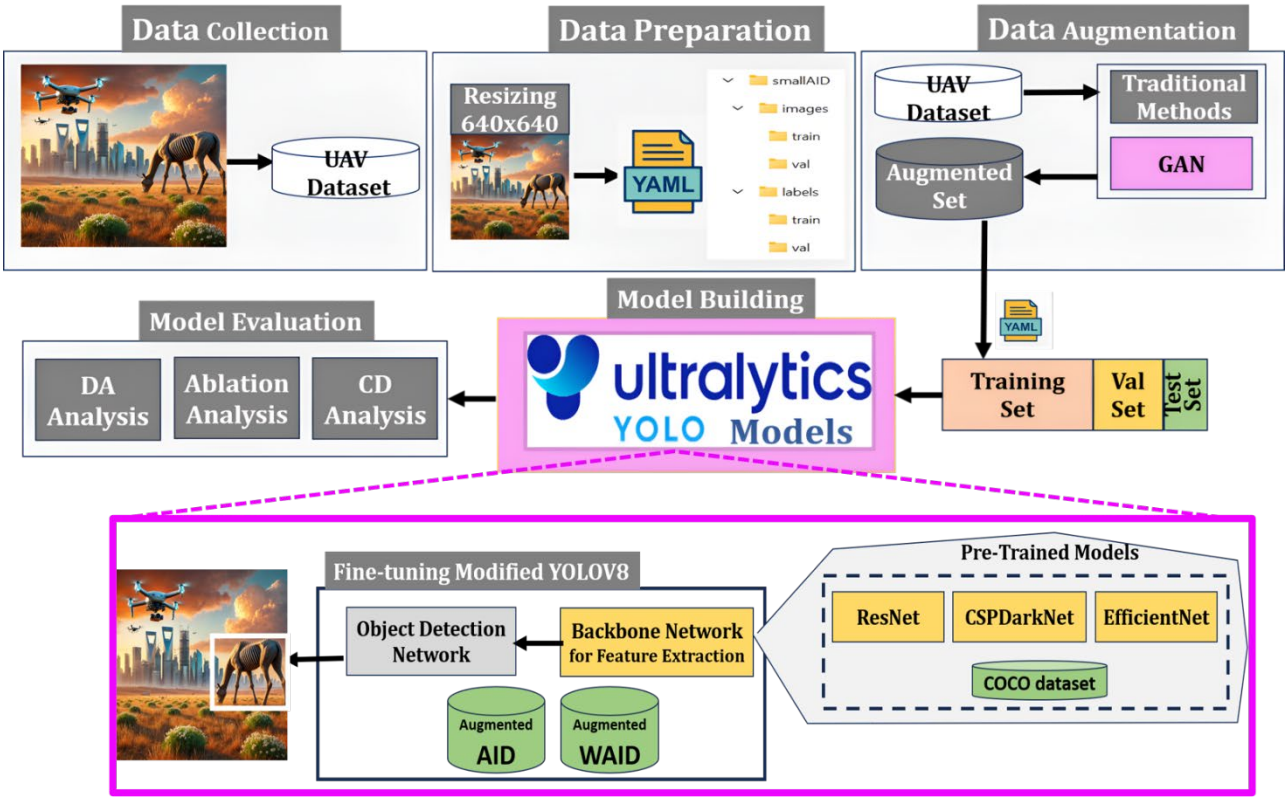

Fig. 1 *Proposed YOLOv8 variant with GAN and EfficientNet-B3 backbone for UAV-based wildlife detection*

1.1 Generative Adversarial Network (GAN)

GANs are a powerful method for DA, particularly in scenarios, where acquiring large, diverse datasets is challenging, such as urban wildlife conservation. GANs are made up of two neural networks: the generator and the discriminator, which are trained concurrently in a competitive manner [11]. The generator generates synthetic data (such as images), while the discriminator tries to discriminate between actual and produced data. Over time, the generator improves at producing realistic, high-quality data that closely mimics the original dataset. The overall objective function for GANs may be written as follows [12]:

$$GAN\ Loss = \min_{\theta} \max_{\emptyset} E_{x\sim p(x)} [\log D_{\emptyset}(x)] + E_{z\sim p(z)} [\log 1 - D_{\emptyset}(G_{\theta}(z))] \quad (1)$$

Within DA, GANs creates varied training samples by creating variants of current data, such images of wildlife from many angles, lighting conditions, or environments [13,14]. This method enhances the amount of training data without requiring expensive and labor-intensive manual data collection, leading to better accuracy and reliability in machine learning.

1.2 You Only Look Once version 8 (YOLOv8)

YOLOv8 is an advanced architecture for object detection and tracking that enhances the achievements of earlier versions by providing notable advancements in speed and accuracy. This model uses a single neural network to analyze the whole image at once, allowing for real-time detection and tracking of several objects at the same time. YOLOv8 comprises two modules viz, Feature extraction and Object detection as shown in Fig. 1. The feature extraction module also called backbone network employs a series of convolutional layers to extract features from the input image. Subsequently, the object detection module uses anchor boxes to predict bounding boxes and class probabilities for each detected object [15, 16]. This unified architecture simplifies the detection process and decreases the latency linked to multi-stage models, making it well-suited for applications that need quick response times, like monitoring wildlife in changing environments. YOLO models aim to reduce the loss function, consisting of three main components: classification loss ($LOSS_{cls}$), object loss ($LOSS_{obj}$), and localization loss ($LOSS_{LOC}$) [15].

$$LOSS_{YOLO} = LOSS_{LOC} + LOSS_{obj} + LOSS_{cls} \quad (2)$$

In our research methodology for monitoring urban wildlife, we utilize YOLOv8, leveraging its excellent capability to detect small and closely grouped objects. In urban settings where infrastructure and different challenges may hide wildlife, this capability is especially important. In addition, YOLOv8's efficiency allows for real-time processing, which making it possible for the continuous, lag-free monitoring of animal behaviors and movements. As a pretrained model, YOLOv8 may use the weights gained from the COCO dataset and get fine-tuned with the research wildlife datasets via transfer learning, improving accuracy even when training data is low [17, 18]. This flexibility is especially useful in urban wildlife conservation, where gathering large labeled datasets presents considerable difficulties. This study's overarching goal is to strengthen YOLOv8 by integrating EfficientNet-B3 as its backbone. With this integration, we may build a robust and effective wildlife monitoring system that advances conservation efforts in urban settings by increasing detection rates.

1.3 Pretrained Convolutional Neural Networks (CNN)

The efficacy of pretrained CNNs in feature extraction has transformed several computer vision applications, such as image classification, object identification, and segmentation. Pretrained CNNs use transfer learning, enabling models trained on huge datasets like ImageNet to provide a robust basis for particular applications [19]. By using the features learned from these large datasets, pretrained models can perform impressively well even when working with smaller labeled datasets [20]. This method speeds up the training process and enhances model accuracy by offering strong feature representations that can be applied across various domains and applications. EfficientNet-B3 has emerged as a premier architecture for feature extraction among several pretrained CNNs, owing to its novel design and effective scalability [21]. EfficientNet-B3 utilizes a compound scaling approach that effectively balances network depth, breadth, and resolution, guaranteeing that each dimension enhances performance. This design has depthwise separable convolutions, enabling it to attain great accuracy with a very minimal parameter count. Consequently, EfficientNet-B3 excels in capturing fine features and complicated patterns in pictures, rendering it very efficient for tasks necessitating exact feature representation, such as item detection and classification in diverse contexts.

The state-of-the-art object identification model YOLOv8 is built on top of EfficientNet-B3, which is ideal for this role because to its outstanding performance. In this study, we improve YOLOv8's detection speed and accuracy by using EfficientNet-B3 as its backbone and taking use of its strong feature extraction capabilities. This combination enables effective animal tracking and identification, particularly in urban environments where real-time processing is essential. Due to its efficiency, which helps maintain a low computational load, EfficientNet-B3 is well-suited for use on devices with limited resources, such as UAVs. Animal conservation efforts that rely on precise detection and monitoring gain significant advantages from using EfficientNet-B3 alongside YOLOv8.

Table 1 *Architectural design structure of GAN*

Layer	Output Size	Filters/ Neurons	Act. func
Generator			
Input	(100,)		
Fully connected	(512,)	512	ReLU
Reshape	75x75x128		
ConvTranspose2D	150x150x128	256	ReLU
ConvTranspose2D	300x300x64	128	ReLU
ConvTranspose2D	300x300x3	3	Tanh

Table 2 *Class distribution details in the research dataset*

Dataset	WAID			AID	
Classes	Train	Val	Test	Train	Test
Camelus	512	149	82	67	27
Sheep	3602	349	173	99	74
Zebra	443	126	65	181	31
Fox	-	-	-	148	69
Leopard	-	-	-	123	57
Ostrich	-	-	-	136	76

2. Experimental Setup

2.1 Model Development

The generator in the proposed GAN is structured to transform a random noise vector of dimension 100 into a RGB image of size 300×300×3. The noise vector is processed through a series of dense layers with 512 neurons and reshaping layer to converts it into a three-dimensional tensor of size 75×75×128. Following this, transposed convolutional layers progressively upscale the tensor, first to 150×150×128 and then to 300×300×64. Each stage

is supplemented with batch normalization to stabilize the training process and dropout layers with 0.2 to mitigate overfitting. While the discriminator used the same architecture but in reverse order. Both networks used the leaky ReLU activation function to enhance gradient flow, as shown in Table I.

We used a 24 GB GPU to improve performance and built YOLOv8 on top of EfficientNet-B3 using the Ultralytics and Keras Python libraries on the Google Colab platform. The Ultralytics library offers an easy-to-use interface for working with YOLOv8, making it simple to develop and test the model for object detection tasks [22, 23]. Similarly, the Keras library, recognized for its user-friendliness and adaptability, was utilized to construct and modify the EfficientNet-B3 backbone, facilitating improved feature extraction suited to the particular needs of wildlife monitoring.

2.2 Research Dataset

This study uses two datasets: the AID Dataset from Kaggle [24] and the WAID dataset from GitHub [25]. Both provide useful resources for confirming DL models in UAV-focused urban wildlife conservation. The Kaggle dataset features almost 29,000 high-quality images along with bounding box annotations, serving as a reliable resource for object detection tasks and showcasing a variety of animals in both natural and urban environments. In contrast, the WAID dataset has a larger volume of UAV images depicting diverse species, providing a broad environmental context that ranges from urban areas to wild habitats, which is crucial for constructing robust models. The use of both datasets enhances the experiments' capacity to evaluate the proposed model's efficacy in animal monitoring, guaranteeing robust generalization across diverse landscapes critical for urban wildlife conservation. From the images presented, only images of six animals were selected for the training and evaluation of the proposed model, as seen in Table 2.

2.3 Dataset Preparation

The dataset preparation encompasses critical measures to augment model correctness, optimize training efficiency, and guarantee superior generalization across varied datasets and situations, which is vital for resilient real-world performance. The datasets used for this investigation are prepared as follows:

2.3.1 Data Preprocessing

Following the acquisition of data from open sources, it is essential to preprocess the images to maintain uniformity in their dimensions. The animal images in the chosen datasets exhibit variable dimensions, which might provide considerable hurdles during the training process [26]. To overcome this, all images are reduced to 300 x 300 x 3 to match the input layer size of the EfficientNet-B3, which is used as the backbone of the produced YOLOV8. This scaling phase is critical because it provides consistency across both datasets, enabling the model to handle data more efficiently. Furthermore, downsizing reduces computing complexity while maintaining crucial information required for successful animal recognition and tracking, improving both accuracy and efficiency in the proposed model.

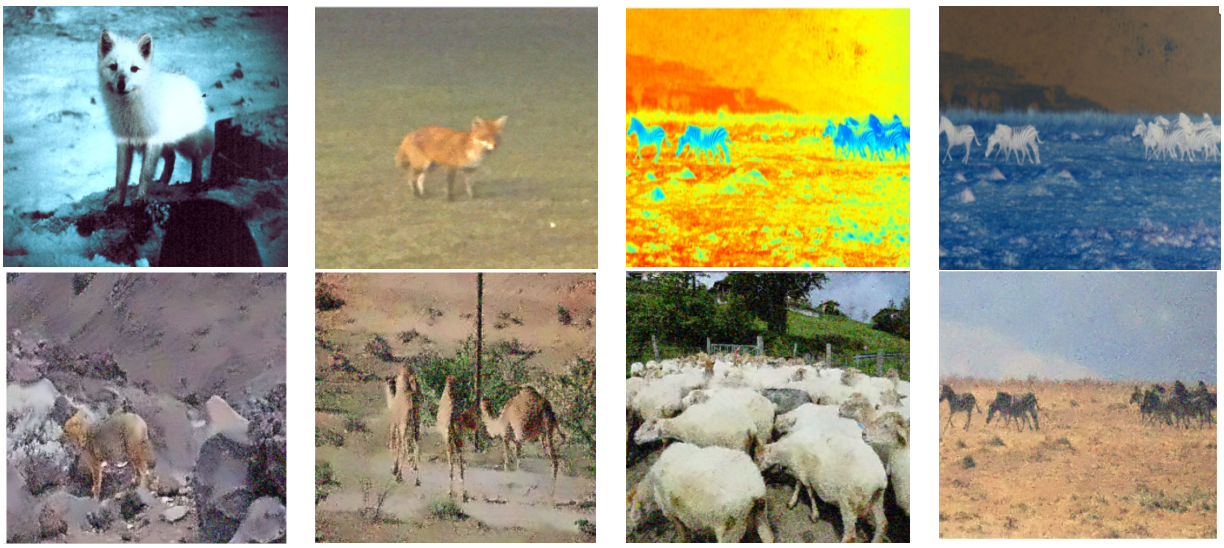

Fig. 2 *Synthetic images generated by traditional methods (HE, CS, TI, NI) in (Row-1) and GAN (Row-2)*

2.3.2 DA Process

In the proposed study, after the preprocessing of both datasets to standardize image dimensions, DA using fundamental classical techniques such as histogram equalization (HE), contrast stretching (CS), and GAN-based augmentation is used to increase the quality and variety of the training dataset. HE produces images by reallocating intensity levels, hence enhancing feature distinguishability. CS produces images by augmenting the contrast between objects and their backdrop, which is essential for differentiating animals from their surroundings in wildlife monitoring [27]. Thermal (TI) and negative (NI) images were made to provide additional variation to the collection and simulate different environmental scenarios. The developed GAN model generates synthetic images that mimic the attributes of the original images. By supplementing the training dataset with rich and better set of generated images, the created model may learn more effectively and improve accuracy in detecting and tracking animals in a variety of environmental settings. Fig. 2 shows samples of synthetic images made using the developed GAN and traditional techniques.

2.3.3 Training Procedure

Transfer learning is employed to tackle the problem of limited sample size and speed up network convergence by using EfficientNet-B3 model which is ImageNet pretrained as the backbone network's initialization weights and then training the developed YOLOV8 model on the study dataset. Selecting the optimal training hyperparameters to maximize model performance on both datasets required some deliberation. In order to promote continuous convergence, an Adam optimizer—renowned for its adjustable learning rates—was used with a learning rate of 0.001. Training with a batch size of 16 made efficient use of Google Colab's 24GB GPU, enabling quick image processing. We discovered an optimal training duration of 50 epochs that would allow us to train the model thoroughly while keeping the overfitting risk to a minimum.

3. Result and Discussion

This section describes the tests carried out to assess the usefulness of the suggested model in monitoring urban animals using UAV images. The results of these studies are described, demonstrating the model's capabilities and potential for accurate identification and tracking of animals in urban situations.

3.1 DA Analysis

The experiment sought to assess the impact of various DA approaches on the efficacy of the proposed model, specifically using the two research datasets. DA is essential for improving model generalization, particularly when training data is few or insufficiently diverse. In this case, three DA approaches were examined: traditional methods, GAN-based methods, and a combination of both. Each approach was assessed based on precision (P), recall (R), and accuracy (Acc) to determine its influence on the model's effectiveness. The results of this analysis are summarized in Table 3.

Table 3 *DA analysis results*

DA Methods	P	R	Acc
Traditional	0.937	0.927	0.945
GAN	0.914	0.902	0.922
Both (Proposed)	1	0.959	0.987

The first method, classical DA, yielded commendable performance, with a precision of 0.937, a recall of 0.927, and an accuracy of 0.945. This indicates that conventional procedures, such as flipping, cropping, and modifying picture brightness, may significantly enhance model performance by including basic variations in the data. Nonetheless, while these strategies provide enhancements, they are relatively constrained in the variety they include. Thus, the model's generalization capability is modest, indicating possibility for improvement, particularly in the context of more intricate data variances.

The second strategy emphasized GAN-based DA, which produced synthetic data to enhance the size and variety of the training set. This method led to a minor decrease in performance relative to conventional augmentation, achieving a precision of 0.914, recall of 0.902, and accuracy of 0.922. GANs are capable of generating a wider range of intricate images, mimicking data from different perspectives and situations that might not exist in the initial dataset. However, GANs may contribute artificial noise or unrealistic components, which may explain the minor performance drop. Although GANs can alleviate data shortages, relying only on synthetic data might reduce the model's accuracy and recall.

The most promising outcomes came from combining classical and GAN-based augmentation. This strategy considerably improved the model's performance, resulting in perfect precision (1.000), recall (0.959), and

accuracy (0.987). By integrating the advantages of both techniques, the model gained from the simplicity of classical augmentation and the variety supplied by GAN-generated images. This combination resulted in a more robust training dataset, enhancing the model's capacity to generalize to various settings. This is particularly important for practical applications like animal monitoring in complex urban environments.

Lastly, the results of this experiment highlight the significant role of employing a varied DA technique to enhance the model's dependability for wildlife identification tasks in changing urban environments. The PR curve is a useful tool for determining how well DL models perform, particularly when the datasets are not balanced [28, 29]. This curve sheds light on the trade-off between accuracy and recall, providing for a complete understanding of the suggested model's promises. We can assess the proposed model's efficacy across several threshold settings by looking at the area under the curve (AUC) in the PR curve in Fig. 3. This research helps to make intelligent decisions on the appropriate balance of false positives and false negatives for certain applications.

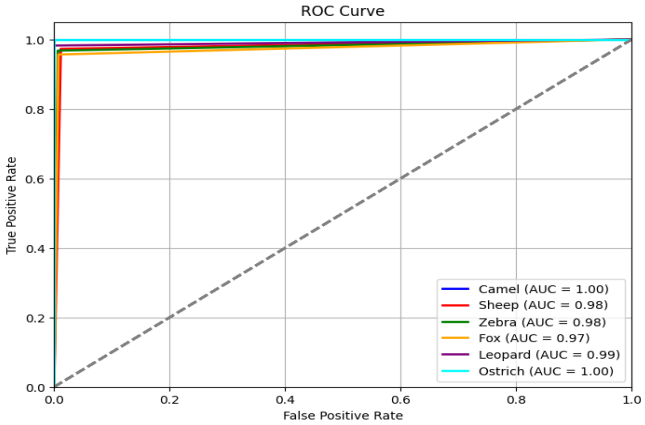

Fig. 3 *PR curve on AID testing dataset*

3.2 Ablation Analysis

Next, an ablation analysis was performed on the two chosen study datasets by creating variants of YOLOv8 using different backbone networks, such as ResNet50 and EfficientNet-B3. This analysis is essential for showing how various elements of the proposed model impact its overall performance, especially regarding wildlife detection and monitoring. Model optimization relies heavily on ablation analysis, which verifies the proposed model's design choices. We can evaluate the impact of each backbone network on critical performance measures like Acc, P, and R by systematically modifying the networks. The results of this experiment are summarized in Table 4. Due to space constraints, sample outputs from the suggested model on the AID test set are displayed in Fig. 4, providing visual evidence of the model's capabilities.

Table 4 *Ablation analysis results*

Dataset	AID			WAID		
YOLOv8 Backbones	P	R	Acc	P	R	Acc
CSPDarkNet53	0.969	0.942	0.953	0.966	0.941	0.958
ResNET50	0.881	0.874	0.861	0.779	0.775	0.761
EfficientNetB3 (Proposed)	1.000	0.979	0.987	0.972	0.953	0.964

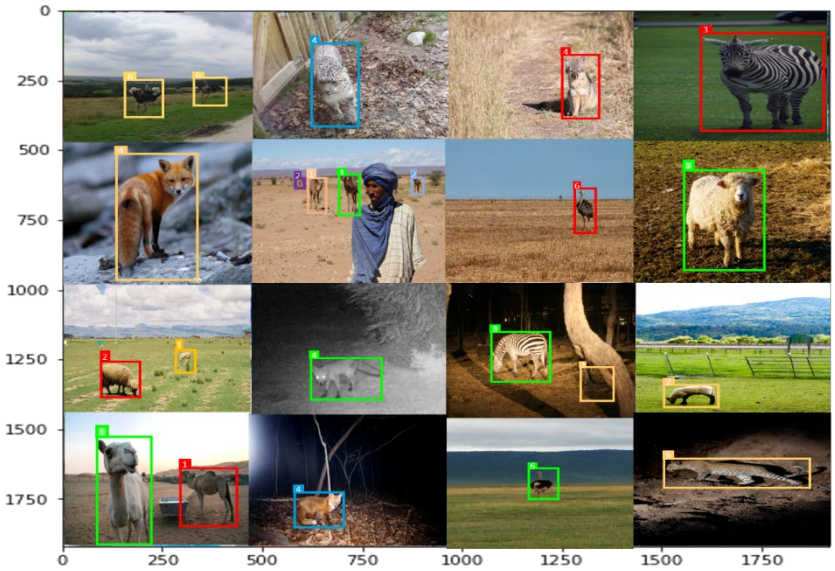

Fig. 4. *Sample animal detection output images from the proposed model on AID testing set*

The analysis results reveal clear distinctions in how well each backbone - CSPDarkNet53, ResNet50, and EfficientNet-B3 supports the model's ability to detect wildlife in two different datasets, AID and WAID. This experiment enables a thorough understanding of how different pretrained CNN backbones contribute to overall detection accuracy, precision, and recall, all of which are crucial in real-world applications. With the CSPDarkNet53 backbone, the model performs well on both datasets. It delivers great precision, recall, and accuracy with little differences among datasets. On the AID dataset, CSPDarkNet53 produces P: 0.969, R: 0.942, and Acc: 0.953, but on the WAID dataset, it produces P: 0.966, R: 0.941, and Acc: 0.958. These findings show that CSPDarkNet53 is capable of handling both datasets, however it is not the best-performing backbone in our investigation. The effectiveness of this backbone's performance is rooted in its design and depth, making it well-suited for feature extraction in complex urban settings for wildlife monitoring.

The performance of the ResNet50 backbone is, however, noticeably lower than that of the other backbones. On the AID dataset, ResNet50 records P: 0.881, R: 0.874, and Acc: 0.861. However, its performance decreases on the WAID dataset, showing P: 0.779, R: 0.775, and Acc: 0.761. The results indicate that ResNet50 has difficulty managing the variability and complexity of the datasets utilized in this study. In spite of its extensive use in several image identification tasks, this backbone completely performs low in this specific context.

EfficientNet-B3 backbone produces outstanding results when compared with CSPDarkNet53 and ResNet50. On the AID dataset, EfficientNet-B3 achieves detection performance with perfect precision (P: 1.000) along with R: 0.979, and Acc: 0.987. The results on the WAID dataset demonstrate resilience across both datasets, with P: 0.972, R: 0.953, and Acc: 0.964. All of these outcomes prove that EfficientNet-B3 is the best backbone when compared to its competitors in terms of depth, breadth, and resolution balance. Its robust accuracy and recall across several images demonstrate its applicability to intricate applications such as urban animal identification, where images may be very noisy and subject to variation.

The visual outputs presented in Fig. 4 further corroborate these findings, showcasing the model's ability to detect animals accurately in test images from the AID dataset. The bounding boxes around the observed animals give a clear visual representation of the EfficientNet-B3 performance, demonstrating its great precision and accuracy in real-world detection. This not only confirms the quantitative findings of the ablation study but also demonstrates the model's practical relevance in field deployments for urban wildlife monitoring. Overall, the ablation study shows that EfficientNet-B3 is the best backbone for YOLOv8 when it comes to detecting animals in urban areas. It outperforms CSPDarkNet53 and ResNet50 in terms of precision, recall, and accuracy. Its high feature extraction capacity makes it ideal for dealing with the complexities of urban wildlife identification jobs across a variety of datasets.

3.3 Cross – Dataset Validation (CDV) Analysis

CDV is essential for assessing the proposed model, especially because the YOLOv8 backbone is built on a transfer learning technique. This validation assures that the model generalizes successfully beyond the dataset on which it was fine-tuned, demonstrating its resilience under a variety of scenarios. Since YOLOv8 relies on transfer learning to employ pre-trained weights, CDV is necessary to assess its flexibility in dealing with different types of data and real-world scenarios. In this experiment, we train and test the proposed model with different

combinations of the two datasets to see how it does in different scenarios. The experimental findings are shown in Table 5, which demonstrates the model's robustness and generalizability.

Table 5 *CDV analysis results*

Training Set	Test Set	P	R	Acc
WAID	WAID	0.972	0.953	0.964
WAID	AID	0.583	0.629	0.635
AID	WAID	0.776	0.719	0.743
AID	AID	1.000	0.959	0.987
{AID, WAID}	WAID	0.981	0.974	0.988
{AID, WAID}	AID	1.000	0.991	1.000

The results of the CDV analysis provide crucial details on the suggested model's performance across different dataset combinations during training and testing. When the model is trained and tested on the same dataset, either WAID or AID, it demonstrates excellent performance. For example, the model trained and evaluated on the WAID dataset has 0.972 precision, 0.953 recall, and 0.964 accuracy. In the same way, when the model is trained and tested on the AID dataset, it achieves outstanding scores, with precision and accuracy both at 1.000 and recall at 0.991. This shows that the model successfully reflects the traits of each dataset when there is consistency between training and testing data, indicating a strong ability for precise detection within the same data environment. Nonetheless, the model's performance decreases noticeably when evaluated on a dataset different from the one it was trained on, highlighting the difficulties of adapting to new domains. For example, the precision falls to 0.583, recall to 0.629, and accuracy to 0.635 when trained on WAID and evaluated on AID. In the same way, when the model is trained on AID and then tested on WAID, it shows a slight improvement with a precision of 0.776, recall of 0.719, and accuracy of 0.743, yet these figures remain significantly lower than the results obtained from the same dataset. This performance decline emphasizes the difficulty of the model in generalizing across datasets with different features, so stressing the need of careful CDV to evaluate robustness in practical applications where data distributions vary.

Finally, when the model is trained on a combination of both datasets (AID + WAID) and tested on each individually, the results show substantial improvement. For WAID, the model shows a precision of 0.981, a recall of 0.974, and an accuracy of 0.988, while for AID, it achieves perfect precision and accuracy scores of 1.000. This shows that using a wider variety of datasets improves the model's ability to apply its knowledge to various animal species, making it more reliable and flexible in handling new data situations. These findings highlight the significance of CDV in facilitating model deployment for urban wildlife conservation efforts.

4. Conclusion

This study presents an innovative DL model for urban wildlife monitoring using UAV imagery, integrating adversarial and transfer learning techniques within the YOLOv8 framework. The model makes detection much more accurate, reliable, and fast by using GANs for DA and EfficientNet-B3 as the backbone for feature extraction. Three experimental analyses—DA, ablation, and CDV—proved the success of the proposed approach. Specifically, DA analysis highlighted the advantages of combining traditional and GAN-based augmentation for achieving superior model generalization. EfficientNet-B3 was found to be the best backbone for improving detection accuracy through ablation analysis. CDV demonstrated the model's adaptability to a wide range of animal species and its application to real-world urban settings.

As previously stated, the proposed approach in this study may be applied to meet the UN SDGs, notably Goals 11,13, 15. The improved wildlife detection capabilities of the proposed model may also serve as a practical tool for increasing the ecosystem resilience and biodiversity of urban landscapes, which can benefit conservation and sustainability efforts worldwide.

Future research could investigate expanding the model to incorporate other species or areas, so enhancing adaptation to different ecological environments. Moreover, incorporating more advanced DA strategies or by integrating hybrid DL architectures may further refine its performance. Looking forward, the proposed model could open the path for more extensive uses of conservation technology, so guaranteeing that it will have a long-term effect on conservation initiatives in urban areas and beyond.

Acknowledgement

The author extends her appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2024/01/29695).

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm their individual contributions as follows: Thavavel Vaiyapuri contributed to ***conceptualization, methodology design, software implementation, secured funding, and manuscript drafting****. Elham Kariri was involved in* ***data collection, data exploratory analysis, and manuscript review****. Ahmed Abdelrahman contributed to* ***data preparation, experimental setup, and manuscript editing****. Raya Aldawood conducted the* ***literature review, and conducting experiments****. All authors reviewed and approved the final manuscript for publication.*

References

[1] Merri K Collins, Seth B Magle, and Travis Gallo (2021). Global trends in urban wildlife ecology and conservation, *Biological Conservation*, 261:109236.

[2] Wael A Aboneama (2021). Applying sustainable principles to create new urban areas and developing existing cities in 2030 vision of Saudi Arabia. *In Resilient and Responsible Smart Cities*, 219–231.

[3] Seth B Magle and Dave Aftandilian (2024). Urban Wildlife: Threats, Opportunities, and Religious Responses, *Routledge*, 257–276.

[4] Chen Fu, Feifei Jiang, Jing Ma, Mohammed A. Alghamdi, Yanfeng Zhu, and Jean Wan Hong Yong. Intersecting planetary health: Exploring the impacts of environmental stressors on wildlife and human health. *Ecotoxicology and Environmental Safety* 283 (2024): 116848.

[5] Abdulaziz S Alatawi (2022). Conservation action in Saudi Arabia: Challenges and opportunities, *Saudi Journal of Biological Sciences*, 29(5), 3466–3472.

[6] Ismaila Rimi Abubakar and Umar Lawal Dano (2020). Sustainable urban planning strategies for mitigating climate change in Saudi Arabia, *Environment, Development and Sustainability*, 22, 5129–5152.

[7] Kyeong-Tae Kim, Hyun-Jung Lee, Seung-Wook Jeon, Won-Kyong Song, and Whee-Moon Kim (2023). Development of urban wildlife detection and analysis methodology based on camera trapping technique and yolo-x algorithm, *Journal of the Korean Society of Environmental Restoration Technology*, 26, 17–34.

[8] L J Mangewa, P A Ndakidemi, and L K Munishi (2019). Integrating UAV technology in an ecological monitoring system for community wildlife management areas in Tanzania, *Sustainability*, 11 (21), 6116.

[9] Sowmya Sankaran (2024). Multi-species object detection in drone imagery for population monitoring of endangered animals, *arXiv preprint arXiv*:2407.00127.

[10] Johnwesily Chappidi and Divya Meena Sundaram (2024). Novel animal detection system: Cascaded yolov8 with adaptive preprocessing and feature extraction, *IEEE Access (early access)*.

[11] Yuzhen Lu, Dong Chen, Ebenezer Olaniyi, and Yanbo Huang (2022). Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review, *Computers and Electronics in Agriculture*, 200, 107208.

[12] Manas Ranjan Prusty, Vaibhav Tripathi, and Anmol Dubey (2021). A novel data augmentation approach for mask detection using deep transfer learning, *Intelligence-Based Medicine*, 5, 100037.

[13] Qiaoyi Zhang, Xiaoli Yi, Jiali Guo, Yadong Tang, Tao Feng, and Rui Liu (2023). A few-shot rare wildlife image classification method based on style migration data augmentation, *Ecological Informatics*, 77, 102237.

[14] Jaekwang Lee, Kangmin Lim, and Jeongho Cho (2022). Improved monitoring of wildlife invasion through data augmentation by extract–append of a segmented entity. *Sensors*, 22(19), 7383.

[15] Gang Wang, Yanfei Chen, Pei An, Hanyu Hong, Jinghu Hu, and Tiange Huang (2023). UAV-YOLOv8: A small-object-detection model based on improved yolov8 for UAV aerial photography scenarios, *Sensors*, 23(16), 7190.

[16] Yiting Li, Qingsong Fan, Haisong Huang, Zhenggong Han, and Qiang Gu (2023). A modified yolov8 detection network for UAV aerial image recognition, *Drones*, 7(5), 304.

[17] Zuxiang Situ, Shuai Teng, Wanen Feng, Qisheng Zhong, Gongfa Chen, Jiongheng Su, and Qianqian Zhou (2023). A transfer learning-based yolo network for sewer defect detection in comparison to classic object detection methods. *Developments in the Built Environment*, 15, 100191.

[18] Subek Sharma, Sisir Dhakal, and Mansi Bhavsar (2024). Evaluating transfer learning in deep learning models for classification on a custom wildlife dataset: Can YOLOv8 surpass other architectures?, *preprint arXiv:2408.00002*.

[19] Javaria Amin, Irum Shazadi, Muhammad Sharif, Mussarat Yasmin, Nouf Abdullah Almujally, and Yunyoung Nam (2024). Localization and grading of NPDR lesions using ResNet-18-YOLOv8 model and informative features selection for DR classification based on transfer learning, *Heliyon*, 10.

[20] Esraa Hassan (2024). Enhancing coffee bean classification: a comparative analysis of pre-trained deep learning models, *Neural Computing and Applications*, 36(16), 9023–9052.

[21] Tay Shiek Chi, Mohd Nadhir Ab Wahab, Ahmad Sufril Azlan Mohamed, Mohd Halim Mohd Noor, Khaw Beng Kang, Lim Lay Chuan, and Liau Wei Jie Brigitte (2024). Enhancing efficientnet-YOLOv4 for integrated circuit detection on printed circuit board (PCB)(December 2023), IEEE Access (early access).

[22] P Naik, G Naik, and M Patil (2022). Conceptualizing python in google colab. India: *Shashwat Publication*.

[23] Vishal Nagpal and Manoj Devare (2024). Computer vision in the sky: Ultralytics yolov8 and deep sort synergy for accurate vehicle speed monitoring in drone video, *Journal of Electrical Systems*, 20(10s), 116–122.

[24] [Dataset]. URL: https://www.kaggle.com/datasets/antoreepjana/animals-detection-imagesdataset

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

[26] Thavavel Vaiyapuri, Ashit Kumar Dutta, Mohamed Yacin Sikkandar, Deepak Gupta, Bader Alouffi, Abdullah Alharbi, Hafiz Tayyab Rauf, and Seifedine Kadry (2022). Design of meta-heuristic optimization-based vascular segmentation techniques for photoacoustic images. Contrast Media & Molecular Imaging, 2022(1):4736113.

[27] Ahmet Haydar Ornek and Murat Ceylan (2019). Comparison of traditional transformations for data augmentation in deep learning of medical thermography. 42nd International Conference on Telecommunications and Signal Processing (TSP), 191–194, IEEE.

[28] Thavavel Vaiyapuri and Adel Binbusayyis (2020). Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: a comparative evaluation, *PeerJ Computer Science*, 6, 1–26. doi: 10.7717/peerj-cs.327

[29] Dhas, M. M., & Singh, N. S. (2024). Breast Cancer Diagnosis Using Majority Voting Ensemble Classifier Approach. *Journal of Soft Computing and Data Mining*, *5*(1), 152-169.