A Deep Learning and UAV Technology-Based Approach to Wildlife Monitoring

Guoqing Zhang[1], Wei Luo[1,2,3,*], Guohong Li[1,2,3], Ruiyin Tang[1,2,3], Xuqing Li[1,2,3] and Bin Wen[1,2,3]

[1]North China Institute of Aerospace Engineering, Langfang 065000, China

[2]Aerospace Remote Sensing Information Processing and Application Collaborative Innovation Center of Hebei Province, Langfang 065000, China

[3]National Joint Engineering Research Center of Space Remote Sensing Information Application Technology, Langfang, China

* luowei@radi.ac.cn (*Corresponding author's email only)

ABSTRACT: *The Qinghai Tibet Plateau's wildlife faces threats to their survival, like poaching. It is essential to provide an intelligent monitoring system to safeguard their lives. On the basis of this, this paper suggests a deep learning and UAV-based multi-target tracking system. First, based on the YOLOv7 detection method, the SimAM attention mechanism is implemented to boost the accuracy of object detection. Next, a unique Deep SORT-based appearance feature extraction network was suggested. In order to minimize the frequency with which target IDs are switched during tracking, this network incorporates the idea of convolutional structure reparameterization and constructs a full-size feature extraction module to extract the target's full-scale appearance features. In addition, a visual servo controller designed especially for UAVs to lessen the impact of rapid movement on tracking. Lastly, using a self-built dataset, experiments were carried out and compared with the most advanced multi-objective tracking algorithms available. The experimental findings demonstrate the great tracking accuracy of the system. For a total of 67.8% and 79.6%, MOTA and MOTP increased by 6.7% and 3.7%, respectively, in contrast to the baseline method Deep SORT. This approach offers great technological support for preserving the diversity of wildlife by demonstrating reliability in complicated settings through thorough evaluation and analysis.*

Keywords: Wildlife monitoring; UAV; Deep learning; Deep SORT

1 Introduction

Procapra przewalskii, an endangered ungulate, thrives in elevated terrains and is one of the rare mammals adapted to high altitudes. It received classification as endangered in the 1999 Red Book that details China's threatened fauna and veterinary species. Despite the existence of legal protections, Procapra przewalskii remains vulnerable to lucrative illegal hunting, driven by the high market demand for its fur. This poaching significantly contributes to the rapid decline in its numbers. The encroachment of human activities continues to degrade its natural habitats, further threatening its survival. In response, the Chinese authorities have set up a specific nature reserve for the species, where ongoing surveys are conducted to assess both the population and the condition of its habitat. The harsh conditions of the high-altitude plateau, characterized by severe cold, reduced oxygen levels, and rugged terrain, pose significant challenges to these monitoring efforts (Luo et al., 2023).

The advancement of UAV technology has initiated a transformative period for wildlife monitoring. UAVs that operate autonomously and are equipped with embedded computing systems have the capability to observe and identify distinct wildlife populations dynamically without human intervention (Luo et al., 2024). The application of sophisticated UAV platforms for extensive surveillance of diverse targets has emerged as an increasingly effective strategy. Owing to the progressive enhancements in deep learning (DL), its amalgamation with UAV technology has become prevalent in the domain of wildlife observation (Zhang et al., 2024). UAVs augmented with DL technologies are capable of processing extensive datasets instantaneously, thereby addressing the challenges associated with the low accuracy, complex models, and sluggish response times inherent in conventional machine learning approaches. This synergy significantly alleviates the labor-intensive and inefficacious aspects of monitoring wildlife, particularly in the demanding conditions of elevated terrains.

Recent advances in DL, characterized by multi-layer neural networks (NNs) and fueled by extensive data, have revolutionized pattern recognition across several domains, including image classification (Jamil et al., 2022; Khan et al., 2022; Rawat et al., 2017). In particular, DL techniques have achieved high levels of precision in detecting targets (Han et al., 2018). Such techniques are commonly categorized into two distinct groups: the first includes two-stage detectors like Regional Convolutional Neural Networks (RCNN) (Girshick et al., 2015), Faster-RCNN (Ren et al., 2015), and Mask-RCNN (He et al., 2017), which have markedly enhanced the precision of pinpointing targets. The second category comprises one-stage detectors, notably the YOLO series (Redmon et al., 2017; Redmon et al., 2020; Bochkovskiy et al., 2020), designed to amalgamate the processes of target classification and localization, thereby substantially boosting the speed of detection. As a result, the field of wildlife detection is evolving into one enriched with data, propelled by advancements in computer vision technology. This progress has facilitated the automated identification of various wildlife species (Duporge et al., 2021; Eikelboom et al., 2019; Gonçalves et al., 2020).

In the domain of computer vision, multi-target tracking (MOT) is crucial, involving the identification of multiple targets within a video sequence, maintaining consistent IDs for each throughout, and documenting their movement paths. This technology has seen extensive application in tracking animals recently. (Sun et al., 2020) proposed an innovative approach by integrating multi-channel color features with pig silhouette data, thus advancing the precision of tracking mechanisms while preserving the efficiency required for real-time applications. (Xiao et al., 2019) enhanced pig identification techniques by leveraging color-based data, employing a novel set of association rules termed DA-ACR to optimize tracking processes. Designing the underlying features for these methods involves substantial effort and demands precise environmental conditions, which are challenging to fulfill in real-world applications. (Zhang et al., 2021) enhanced the YOLO v3 network and integrated it with the Deep SORT algorithm for effective MOT of cows. (Tu et al., 2022) combined YOLO v5 with the improved Deep SORT algorithm to realize the tracking of pigs. These approaches adhere to the tracking-by-detection (TBD) paradigm where results from detectors feed into backend optimization algorithms like SORT (Bewley et al., 2016) and Deep SORT (Wojke et al., 2017), which use Kalman filtering and the Hungarian algorithm for tracking. In this paradigm, frames with low detection scores are discarded to reduce processing demands, which facilitates matching in subsequent frames. This practice may lead to the omission of actual targets and fragmented trajectories, particularly in complex wildlife monitoring scenarios where obstacles and variable lighting conditions are common. This research is dedicated to tracking wildlife populations in mountainous terrains, a task compounded by the harshness of the environment.

Swift changes in UAV trajectories during wildlife surveillance can significantly modify the visual scope of the onboard camera, frequently resulting in interruptions in tracking or total loss of visual contact with the subject. With advancements in autonomous UAV technologies, extensive research has focused on image-based visual servoing. This approach allows UAVs to navigate and avoid obstacles using a single camera. The capability of UAVs to fly slowly, hover, move laterally, and maneuver in confined spaces enhances the effectiveness of visual servo control systems. Such systems are well-suited for tasks including inspection, surveillance, and environmental monitoring. Currently, various control methods are in use, such as adaptive control (Zhang et al., 2017), PID control (Subramanian et al., 2017), sliding model control (Ma et al., 2018), and neural network control (Guo et al., 2017). Among these, the PID controller remains a widely used and effective choice despite its traditional nature.

In conclusion, employing UAVs for wildlife surveillance in intricate and high-altitude conditions poses significant research challenges. This manuscript presents an UAV-based MOT system designed specifically for complex and crowded settings. The major contributions of this study are detailed as follows:

1. The integration of the SimAM attention mechanism into the YOLOv7 detection framework significantly improves the precision of object detections by refining the algorithm ability to focus on relevant features within an image.

2. The development of a novel appearance feature extraction network within the framework of Deep SORT, which incorporates convolutional structure reparameterization. This network establishes a comprehensive feature extraction module designed to decrease the frequency of target ID switches and to mitigate the impacts of occlusions during the tracking process.

3. The design of a visual servo controller ensures automatic tracking by maintaining the target within the visual field of the UAV camera.

This document is structured as follows: Section 2 delineates the scope, subjects, and methodologies employed in the study. Section 3 elaborates on and analyzes the experimental results. Section 4 provides a summary of the conclusions derived from the research.

2 Materials and methods

2.1 Materials

2.1.1 Video acquisition

The subject of the video analysis was Procapra przewalskii, captured around Qinghai Lake in Gonghe County, Qinghai Province. In June 2023, aerial footage was obtained using a P600 UAV operating at elevations between 30 and 50 meters. This aerial survey spanned an extensive region of 2100 km2 and effectively amassed substantial video data that depicted Procapra przewalskii within the designated study zone (Figure 1).

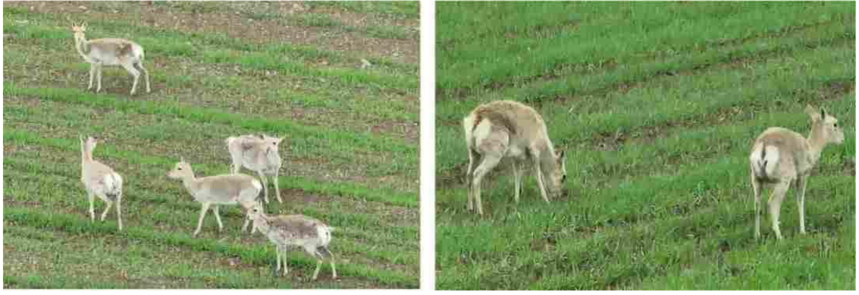

Figure 1: Procapra przewalskii captured in the research area.

2.1.2 Dataset construction

In the designated research area, select a dataset consisting of 20 video sequences featuring Procapra przewalskii to be utilized for detection and tracking purposes. Each of these datasets should encompass a total of 10 distinct video clips. These sequences have been established as benchmarks for determining the location and movements of Procapra przewalskii, and are subsequently analyzed using the YOLOv7 model.

To improve the identification capabilities for Procapra przewalskii and enhance the resolution of fur texture details, a dataset consisting of 2400 images was compiled from 10 distinct video sequences designated for object detection. This dataset was divided into sets for training, validation, and testing, following a distribution ratio of 7:2:1. At the same time, the YOLOv7 model was employed to refine model parameters and boost stability.

In order to better track and monitor objects using the Deep SORT algorithm, a large dataset of common antelopes is needed to extract their appearance features. From 10 video sequences, frame images are captured every 10 frames. These images are then resized to uniform dimensions of 460 × 460 pixels in JPEG format, resulting in a total of 3000 images. The LabelImage software is utilized for annotating these images, storing them in XML format to form the core of the target tracking dataset. The dataset is partitioned into subsets for training, testing, and validation using ratios of 70%, 20%, and 10% respectively.

2.2 Methods

2.2.1 General overview

Figure 2 illustrates the architecture of this article. Video sequences are input into the detection system to obtain the bounding box for the object. The tracking module employs parameterization techniques to forecast trajectories, which aids in constructing a comprehensive feature extraction bottleneck for effective MOT.

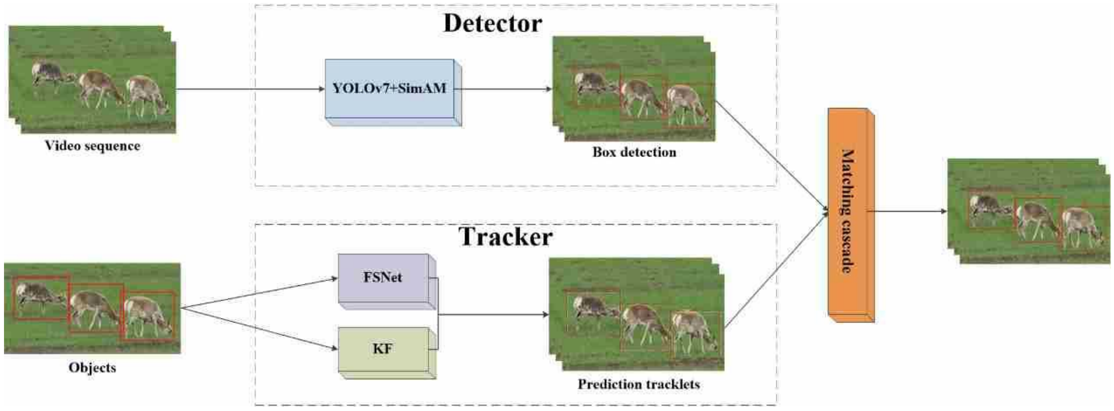

Figure 2: Overall system architecture.

2.2.2 Detector

Contemporary MOT methodologies predominantly adhere to the detection-tracking framework. Accordingly, this research employs a similar detection-centric MOT approach, emphasizing detection as the fundamental initial phase. The objective of this study is the surveillance of densely populated wildlife areas. YOLOv7 has been chosen as the foundational algorithm due to its high accuracy and swift processing capabilities. The implementation of YOLOv7, conducted using PyTorch, ensures compatibility with embedded systems and mobile platforms.

The YOLOv7 model architecture is segmented into three distinct sections: the input layer, the backbone, and the prediction head. At the initial stage, raw images undergo several preprocessing techniques including data augmentation, dynamic resizing, and computation of adaptive anchor boxes, ultimately standardizing the image to a 460×460 RGB RGB format suitable for further processing. Within the backbone, multiple layers such as BConv, EELAN, and MPConv are utilized to decrease spatial dimensions, augment channel capacities, and perform feature extraction, resulting in a three-tiered feature map. The head of the model integrates an SPPCPC layer, additional BConv layers, EELAN-H layers, MPCnv layers, Catconv layers, and a RepVGG block, all of which collaborate to output three distinct feature map layers using the backbone processing capabilities. The three principal tasks in image detection—namely classification, differentiation between foreground and background, and bounding box estimation—are efficiently addressed by these layers. Computations for the final outcomes are executed by leveraging the synergy between the RepVGG block and the convolutional layers.

Given the objective of this research to monitor a densely populated wildlife area where many subjects are occluded, it becomes essential to enhance the feature details in regions of occlusion. This focus ensures that even partially obscured subjects are identifiable, thereby increasing the accuracy of detection and setting a groundwork for future tracking of such targets. To this end, this manuscript introduces an efficient attention module for convolutional NNs, SimAM. Unlike conventional channel and spatial attention modules that necessitate additional parameters in the network, SimAM computes the 3D attention weights directly from the feature maps within a specific layer, also accounting for targets that are partially hidden. This module is seamlessly integrated into the head section of the YOLOv7 architecture without altering the main structure of the network. The adapted YOLOv7 architecture incorporating the SimAM module is depicted in Figure 3.

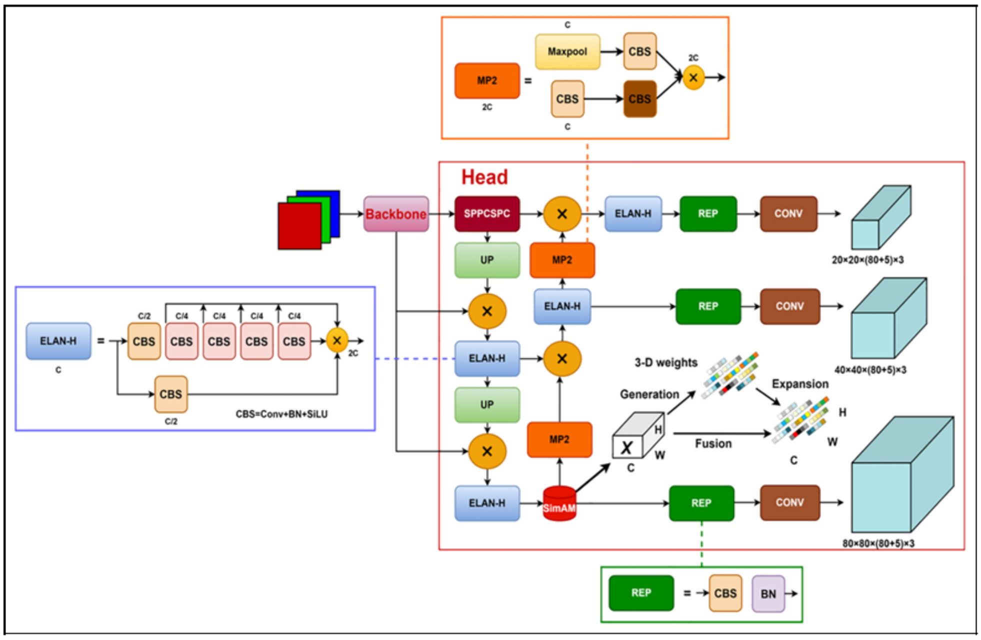

Figure 3: Improved YOLOv7 network structure.

Current methods in attention mechanisms typically compute weights in one or two dimensions from a feature matrix, designated as X. These weights are then uniformly applied across either the channel or spatial dimensions. In the case of channel attention, which is one-dimensional, it differentiates among channels while treating all spatial locations identically. Spatial attention, which is two-dimensional, distinguishes among various locations but applies uniform treatment across all channels. This approach constrains their capacity to identify more nuanced discriminative features. The assignment of three-dimensional weights to the network channels significantly reduces limitations, improving performance compared to conventional one-dimensional and two-dimensional attention mechanisms. The SimAM attention module computes these 3D weights autonomously. It ensures uniform color assignment across all channels, spatial positions, and elemental points within subgraphs. By incorporating the SimAM module, the network experiences a substantial improvement in 3D feature extraction, thus increasing the accuracy of detection.

2.2.3 Tracker

The Deep SORT algorithm augments the SORT algorithm by incorporating a similarity metric for evaluating appearance features of targets, coupled with a cascade matching mechanism. These enhancements reduce ID switches when targets are occluded and increase the overall robustness of the model. Originally, the SORT algorithm uses a Kalman filter to forecast the motion states of objects and applies the Intersection over Union (IOU) metric to assess data association (IOU Match), determined by comparing the predicted target bounding boxes against those produced by the network. The associations are finalized through the Hungarian algorithm, which is pivotal for maintaining continuous tracking of targets. The efficiency of the tracking process is markedly influenced by the capabilities of the feature extraction network, which is pivotal in capturing precise and comprehensive visual details of the subjects. Depicted in Figure 4, the architecture of the Deep SORT algorithm consists of two primary elements: the deep appearance descriptor branch and the motion prediction branch. The latter predicts the trajectory states through Kalman filtering, leveraging data from preceding frames to forecast the positions of targets in the subsequent frame. Discrepancies in space and time between the predicted trajectory and actual detections are quantified using the Mahalanobis distance. Operating as a straightforward convolutional network, the deep appearance descriptor branch undertakes image classification and distills appearance features from the detected frames into a vector of appearance features. The cosine distance metric is employed to evaluate the similarity of these feature vectors. Track segments are associated through a matching cascade algorithm that incorporates both cosine and martens distances. Finally, during the track management phase, the tracks are updated, initialized, and deleted.

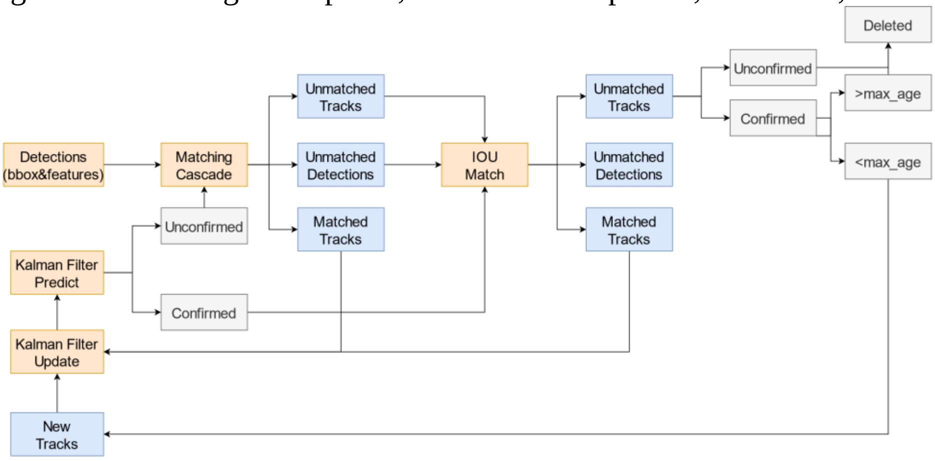

Figure 4: Deep SORT flowchart.

This section introduces FSNet, which facilitates the extraction of target appearance features for MOT applications. Initially, the method of convolutional structure parameterization is described. This is followed by an explanation of the comprehensive feature extraction block and the subsequent feature aggregation module.

Utilizing the principles outlined in the RepVGG framework (Ding et al., 2021), convolutional structure reparameterization separates the training network from the inference network. The methodology utilizes a configuration where a multi-branch convolutional bottleneck is employed during training, with a transition to a single-branch convolutional bottleneck for the inference phase. A detailed representation of this architectural setup is provided in Figure 5.

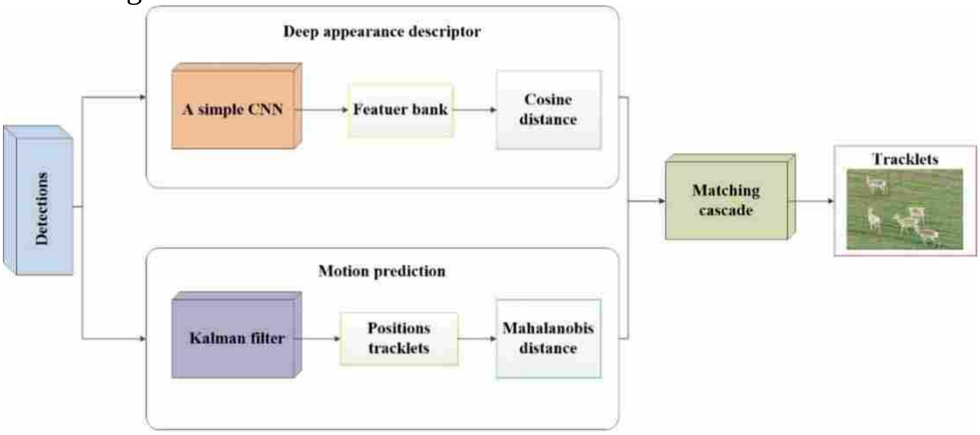

Figure 5: The MOT process in this study.

REPLConv exhibits distinct configurations during its training and deployment phases. In the training phase, the architecture includes a branch composed of Lwtconv3x3+BN layer, alongside a separate BN branch. The outputs of these branches are combined, as depicted in Figure 6a. For deployment, the architecture is simplified by reparameterizing the branch parameters into the primary branch, utilizing only the Lwtconv3x3+BN as the output, illustrated in Figure 6b. The Lwtconv3x3 module integrates a Conv1x1, a DWConv3x3, a BN layer, and a ReLU activation. The DWConv3x3, as a depthwise separable convolution (Howard et al., 2017; Chollet et al., 2017), reduces parameter counts significantly compared to traditional convolution methods. For a given input tensor of dimensions $x \in R^{h \times w \times c}$, conventional convolutions compute parameters as $k^2 \times c \times c'$, where k represents the kernel size, and c the channel count. In contrast, depthwise separable convolutions calculate parameters as $k^2 \times c + c \times c'$, where $k^2 \times c$ accounts for channel-wise convolution and $c \times c'$ for point-wise convolution. Figure 7 illustrates the Lwtconv3x3 structure.

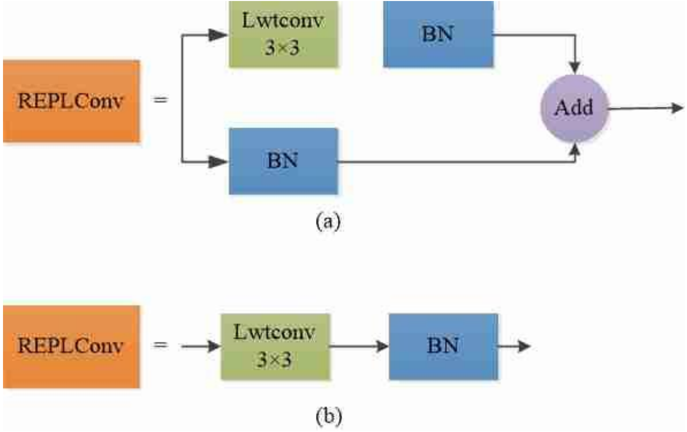

Figure 6: (a) The REPLConv structure during the training. (b) The REPLConv structure during the inference.

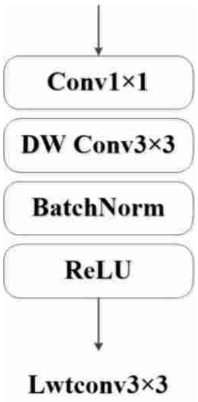

Figure 7: The structure of Lwtconv3×3.

Full-scale feature extraction block. The architecture for extensive feature extraction is detailed by incorporating layers known as REPLConv, as depicted in Figure 7. This structure embraces a multi-branch design to facilitate the learning of features across various scales. Each branch is equipped with distinct REPLConv layers, enabling the differentiation in receptive fields, thus permitting the extraction of scale-specific features. Within this framework, the variable 't' denotes the number of layers stacked within the REPLConv configuration, influencing the receptive field size to be calculated as (2t+1) by (2t-1), provided t exceeds 1. Through a series of experiments, it has been established that setting t to 4 optimizes the trade-off between network depth and the efficacy of feature extraction. This optimal setting allows for the assimilation of fine-scale features as well as the comprehension of spatial features over larger extents. Details of this configuration are illustrated in Figure 8.

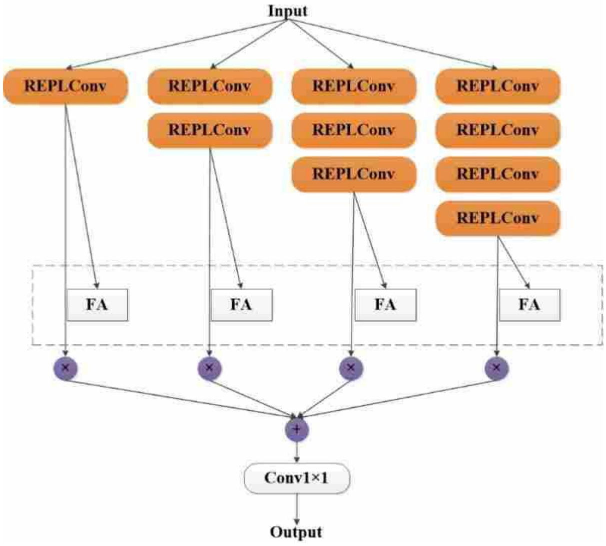

Figure 8: The structure of the full-scale feature extraction block.

Function aggregation. The feature extraction module in each branch captures specific dimensions and scales of features but lacks comprehensiveness. To achieve a holistic feature integration, a Feature Dynamic Aggregation (FA) network architecture is introduced. This compact convolutional neural structure enhances feature synthesis across varying channel dimensions through several components: a global average pooling layer, a densely connected layer, and a layer that includes an activation function.

2.2.4 Visual servo control

This research employs a servo control mechanism within a MOT, integrated with a UAV and a camera. The configuration encompasses four distinct control variables: lateral control, two vertical controls, and control over transverse angular velocity.

The objective of lateral control is to keep the central frame of the camera aligned with the horizontal midpoint of the target. This is achieved through a PID controller, which processes three types of inputs: the cumulative horizontal displacement of each target as a proportional component, the difference in current and preceding central positions as a derivative component, and the total accumulated discrepancy as an integral component.

The longitudinal control of the helicopter modifies the speeds for advancement and retreat by analyzing the bounding box heights of objects, which represent distances from the camera. This adjustment employs a PID controller that computes inputs for velocity adjustments by determining the total deviations between current heights and both maximum and minimum thresholds. Additionally, the controller calculates inputs based on the collective rate of changes in the heights across observed objects, serving as a derivative input for managing speed.

The vertical control of the UAV adjusts the altitude within a predefined range. Unlike the response to lateral velocities, the system exhibits a slower reaction to low vertical velocities, facilitating more precise altitude regulation. Consequently, the altitude of the UAV typically remains unchanged immediately after the autopilot receives a command for vertical velocity.

Yaw rate control allows a helicopter to pivot about its vertical axis, ensuring alignment perpendicular to a line that spans the extremities of objects within the camera field of view. The calculation of the yaw angle is based on the ratio of the horizontal distance to the image width, coupled with the elevation difference among object types relative to a standardized height. The computed angle undergoes modification by the processing time duration and is scaled by a predefined factor to establish the angular velocity necessary for traversal.

3 Experimental results and discussion

3.1 Experimental platform

Table 1 presents the setup details of the experimental environment utilized in our study.

Table 1: Experimental environment.

Name	Configuration information
Operating system	Windows 10
Graphics card	NVIDIA GeForce RTK 4070
CPU	AMD Ryzen 9 5900X
Software	Python 3.8 , Pycharm 2020.1

3.2 Evaluating indicator

In the domain of target detection, models typically identify multiple categories of targets, with each category capable of being represented through a PR curve. This curve facilitates the computation of the Average Precision (AP) value, a numerical measure of the area beneath the PR curve. The mAP is determined by averaging the AP values across all target categories, as outlined in Eq. (1):

$$mAP = \frac{1}{class\_number}\sum_{1}^{class\_number} AP \tag{1}$$

MOT accuracy (MOTA) serves as a comprehensive indicator of tracking efficacy, focusing on both object detection and trajectory continuity without considering the precision of object spatial location. As MOTA values escalate, tracking performance improves. The calculation of MOTA involves:

$$MOTA = 1 - \frac{\sum_t FN + FP + IDSW}{\sum_t GT_t} \tag{2}$$

MOT precision (MOTP) measures the accuracy of positioning. An increased value of MOTP reflects improved accuracy, as determined by Eq. (3):

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t C_t} \tag{3}$$

The IDF1 metric, defined by Eq. (4), is the Identification F1 score which evaluates the accuracy of identity matches. This score comprises IDTP, the count of accurate identity matches, while IDFP and IDFN denote the counts of erroneous matches and failures to match identities, respectively.

$$IDF1 = \frac{2 \times IDTP}{2 \times IDTP + IDFP + IDFN} \tag{4}$$

FP denotes the number of alerts for errors related to the mis-predicted trajectories of objects.

FN accounts for objects that were neither detected nor tracked.

IDs reflect changes in the frequency of identity assignments.

3.3 Benchmarking evaluation

This study introduces a self-made dataset named Procapra przewalskii dataset for object detection tasks. The dataset was partitioned into designated training, validation, and test sets for model evaluation. Enhancements to the YOLOv7 model incorporated the SimAM attention mechanism, leading to a version termed the enhanced YOLOv7. Performance comparisons between the enhanced YOLOv7, the standard YOLOv7, and YOLOv5s models were meticulously conducted. The outcomes, detailed in Table 2, affirm the superiority of the enhanced YOLOv7 model developed in this research.

Table 2: Accurate evaluation of YOLO series algorithms.

Model	**Test size**	**Precision↑(%)**	**Recall↑(%)**	**mAP↑(%)**
YOLOv5s	460	70.5	62.6	68.4
YOLOv7(Wang et al., 2023)	460	82.8	68.7	72.8
YOLOv7+SimAM	460	89.3	80.2	83.6

As demonstrated in Table 2, the enhancement of our YOLOv7 model through the integration of the SimAM module has markedly elevated the detection precision, surpassing not only the original YOLOv7 but also achieving significant advancements over the YOLOv5s model. This substantiates the effectiveness of incorporating the SimAM module into the YOLOv7 framework.

Econometric evaluation demonstrates that the tracking algorithm introduced in this study exhibits superior performance when benchmarked against several leading counterparts. As evidenced in Table 3, this novel algorithm achieves the highest scores in terms of MOTA, MOTP, and IDF1 indicators.

Table 3: Comparison analysis of outcomes using different methodologies.

Model	MOTA↑(%)	MOTP↑(%)	IDF1↑(%)	FP↓	FN↓	IDs↓
SORT	46.8	65.1	46.1	13445	7846	334
Deep SORT	61.1	75.9	58.3	10829	4979	115
JDE (Wang et al., 2021)	61.8	76.7	60.9	10686	4863	235
FairMOT (Zhang et al., 2021)	63.0	78.4	64.4	12683	2275	128
Ours	67.8	79.6	65.8	9688	3359	95

Meticulously evaluated the efficacy of several advanced models, underscoring their potential in wildlife tracking and surveillance systems. As evidenced in Table 3, the MOTA score for the SORT model significantly trails that of the Deep SORT model, which incorporates feature extraction networks. This discrepancy suggests that the integration of a feature extraction network enhances the extraction of more comprehensive target information during the tracking process, thereby elevating the MOTA during feature matching phases and facilitating precise target tracking and monitoring. In comparison to other models, ours excels in multiple metrics, including MOTA, MOTP, and IDF1. Our model shows an enhancement of 6.7% in MOTA, 3.7% in MOTP, and 7.5% in IDF1 over the baseline Deep SORT model. These increments confirm the superiority of our method.

In this investigation, the utilization of a visual servo controller facilitated the adjustment of parameters concerning transverse, longitudinal, vertical, and yaw rate dimensions. This enhancement was specifically designed to improve the P600 intelligent UAV capabilities in detecting and tracking multiple maritime targets. Ordinarily, targets follow a linear trajectory, allowing for independent assessment of the controller. However, to fully evaluate its efficacy in all directions, an experiment with more intricate target movements was organized. For this purpose, two types of antelope were employed to test the UAV tracking precision. Procapra przewalskii were maneuvered along concentric arcs of varying radii to prevent their paths from intersecting with the UAV flight trajectory, as depicted in Figure 9. This arrangement facilitated a detailed analysis of the UAV proficiency in following targets on curved trajectories.

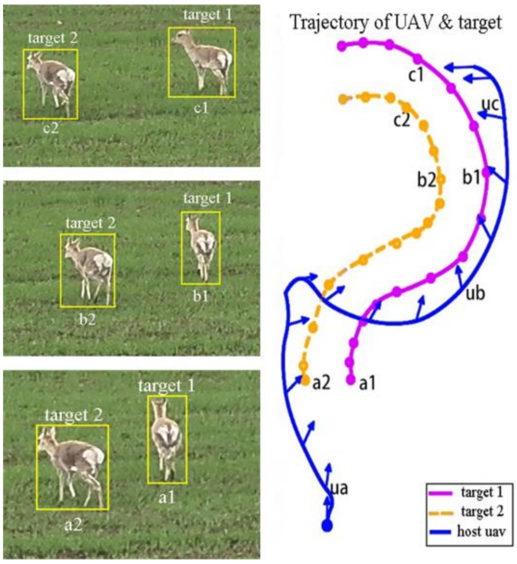

Figure 9: The flight paths of UAV (blue line) and the targets (purple and orange lines). Displayed on the left, three images capture the perspectives from the UAV at specific points labeled as ua, ub and uc.

To further authenticate the yaw rate of the visual servo controller, this investigation proceeded along the trajectory depicted in Figure 9, where the starting point was marked with color-coded solid dots. Along the path of UAV P600, fifteen arrows were utilized to signify the orientation of the UAV axis at the moment. Each target path was adorned with fifteen hollow circles, each aligning with the arrows observed during the experimental phase. These specific instances were captured and illustrated in Figure 10, which also delineates the angular relationships between the connecting lines of the two objects and Xworld, as well as the angular trajectory between YUAV and Xworld. Here, YUAV refers to the longitudinal velocity of the UAV, whereas Xworld denotes the positive horizontal semi-axis of the established coordinate system for tracking the two targets.

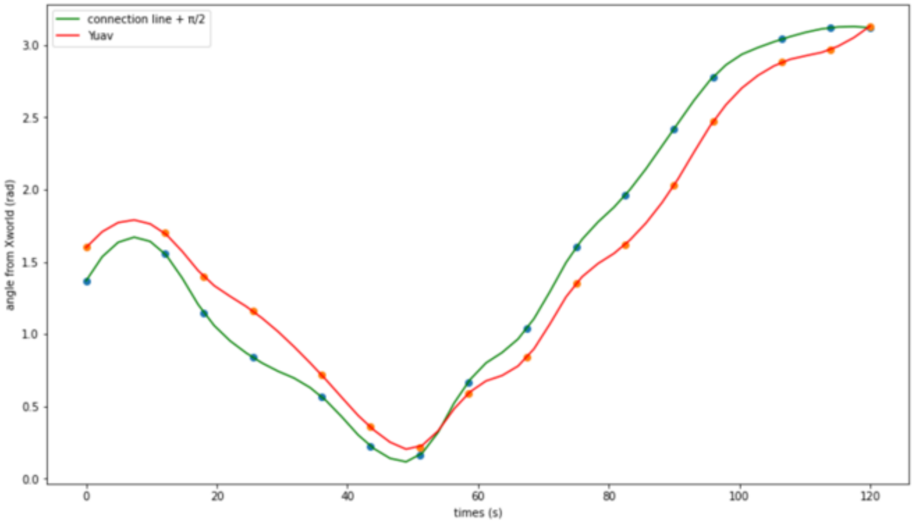

Figure 10: The sum of the angles formed by two connecting objects with an additional Π /2, alongside the angular relationships between Xworld and YUAV compared to Xworld.

In Figure 10, a maximum deviation of -0.362 radians between the UAV trajectory and that of the two targets was recorded at 26.352 seconds. This deviation can be attributed to the swift changes in the positions (Xworld,Yworld) of the targets, which the yaw rate of the visual servo controller could not promptly adjust to. In the final phases of the experiment, the controller maintained a consistent separation of approximately 0.25 radians. During the culmination of the process, it managed the YUAV in an orientation nearly orthogonal to the axis of the connection. These outcomes lend support to the efficacy of the vision servo controller.

Figure 11 presents the outcomes of the algorithm tracking performance using actual Procapra przewalskii population data, highlighting the method efficacy in navigating intricate and crowded environments.

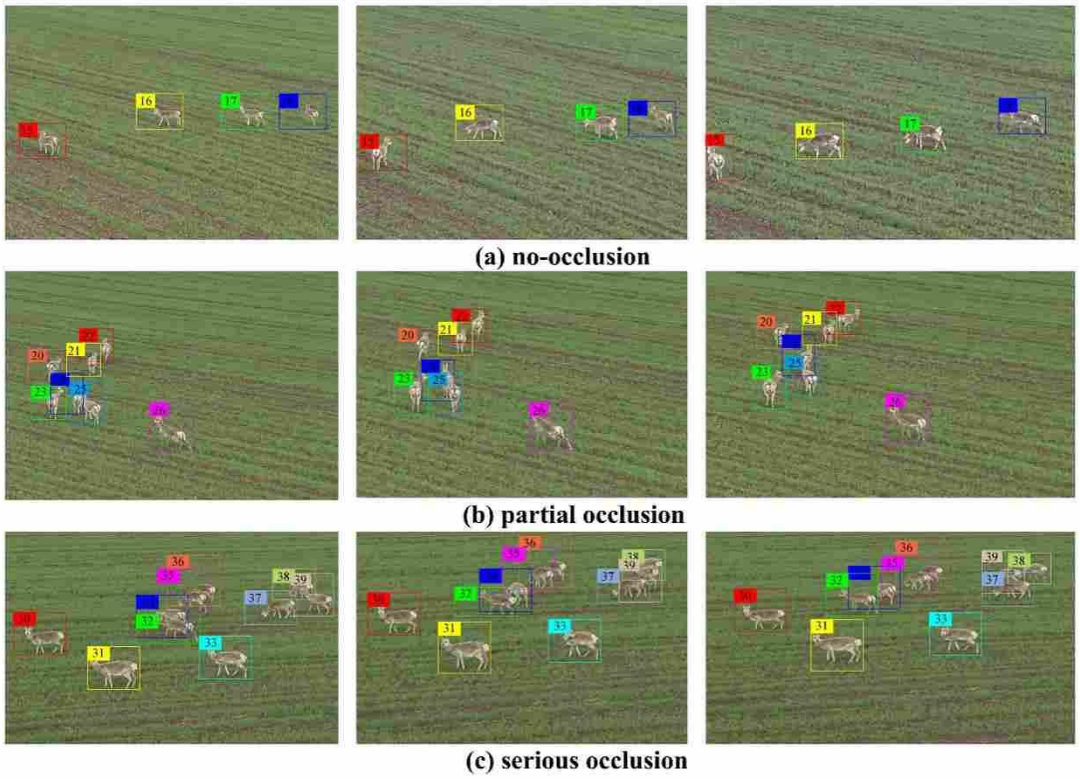

Figure 11: Actual tracking results in different scenarios.

Utilizing the tracking outcomes from actual population data of Procapra przewalskii, the system outlined in this study is capable of performing MOT effectively even within densely populated and intricate environments, while maintaining a high degree of robustness.

4 Conclusion

This research introduces a sophisticated system for tracking multiple targets, employing unmanned aerial vehicles and advanced DL techniques. The system utilizes YOLOv7 combined with the SimAM attention mechanism, enhancing the accuracy of object recognition. The system integrates an enhanced version of the DeepSORT algorithm by incorporating FSNet, a novel network that extracts multi-depth appearance features, ensuring comprehensive feature representation of targets. The study presents the development of a visual servo controller, designed to keep the targets within the UAV camera view, thereby automating the tracking process effectively.

Through comparative analysis of different MOT algorithms, it was determined that the specified method outperforms others in terms of MOTA, MOTP, and IDF1 metrics. Comparison to the baseline Deep SORT, the newly implemented UAV-based MOT system demonstrated notable enhancements, achieving a 6.7% improvement in MOTA, a 3.7% increase in MOTP, and a 7.5% rise in IDF1. These results validate the efficacy of the proposed approach. In summary, UAVs employing this innovative method can effectively deter poaching by conducting real-time surveillance. This study is critically important and offers substantial promotional value for safeguarding the universal gazelle and its natural habitat.

5 Reference

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016, September). Simple online and realtime tracking. In 2016 IEEE international conference on image processing (ICIP) (pp. 3464-3468). IEEE.

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection. arxiv preprint arxiv:2004.10934.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making vgg-style convnets great again. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13733-13742).

Duporge, I., Isupova, O., Reece, S., Macdonald, D. W., & Wang, T. (2021). Using very-high-resolution satellite imagery and deep learning to detect and count African elephants in heterogeneous landscapes. Remote Sensing in Ecology and Conservation, 7(3), 369-381.

Eikelboom, J. A., Wind, J., van de Ven, E., Kenana, L. M., Schroder, B., de Knegt, H. J., ... & Prins, H. H. (2019). Improving the precision and accuracy of animal population estimates with aerial image object detection. Methods in Ecology and Evolution, 10(11), 1875-1887.

Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

Gonçalves, B. C., Spitzbart, B., & Lynch, H. J. (2020). SealNet: A fully-automated pack-ice seal detection pipeline for sub-meter satellite imagery. Remote Sensing of Environment, 239, 111617.

Guo, Z., Pan, Y., Sun, T., Zhang, Y., & Xiao, X. (2017). Adaptive neural network control of serial variable stiffness actuators. Complexity, 2017.

Han, J., Zhang, D., Cheng, G., Liu, N., & Xu, D. (2018). Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Processing Magazine, 35(1), 84-100.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arxiv preprint arxiv:1704.04861.

Jamil, S., Abbas, M. S., & Roy, A. M. (2022). Distinguishing malicious UAVs using vision transformer. AI, 3(2), 260-273.

Khan, W., Raj, K., Kumar, T., Roy, A. M., & Luo, B. (2022). Introducing urdu digits dataset with demonstration of an efficient and robust noisy decoder-based pseudo example generator. Symmetry, 14(10), 1976.

Luo, W., Zhao, Y., Shao, Q., Li, X., Wang, D., Zhang, T., ... & Yu, Z. (2023). Procapra Przewalskii Tracking Autonomous Unmanned Aerial Vehicle Based on Improved Long and Short-Term Memory Kalman Filters. Sensors, 23(8), 3948.

Luo, W., Zhang, G., Shao, Q., Zhao, Y., Wang, D., Zhang, X., ... & Yu, Z. (2024). An efficient visual servo tracker for herd monitoring by UAV. Scientific Reports, 14(1), 1-16.

Ma, Z., & Sun, G. (2018). Dual terminal sliding mode control design for rigid robotic manipulator. Journal of the Franklin Institute, 355(18), 9127-9149.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. Neural computation, 29(9), 2352-2449.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arxiv preprint arxiv:1804.02767.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Subramanian, R. G., Elumalai, V. K., Karuppusamy, S., & Canchi, V. K. (2017). Uniform ultimate bounded robust model reference adaptive PID control scheme for visual servoing. Journal of the Franklin Institute, 354(4), 1741-1758.

Sun, L., Chen, S., Liu, T., Liu, C., & Liu, Y. (2020). Pig target tracking algorithm based on multi-channel color feature fusion. International Journal of Agricultural and Biological Engineering, 13(3), 180-185.

Tu, S., Zeng, Q., Liang, Y., Liu, X., Huang, L., Weng, S., & Huang, Q. (2022). Automated behavior recognition and tracking of group-housed pigs with an improved DeepSORT method. Agriculture, 12(11), 1907.

Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). 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).

Wang, Y., Kitani, K., & Weng, X. (2021, May). Joint object detection and multi-object tracking with graph neural networks. In 2021 IEEE international conference on robotics and automation (ICRA) (pp. 13708-13715). IEEE.

Wojke, N., Bewley, A., & Paulus, D. (2017, September). Simple online and realtime tracking with a deep association metric. In 2017 IEEE international conference on image processing (ICIP) (pp. 3645-3649). IEEE.

Xiao, D., Feng, A., & Liu, J. (2019). Detection and tracking of pigs in natural environments based on video analysis. International Journal of Agricultural and Biological Engineering, 12(4), 116-126.

Zhang, D., & Wei, B. (2017). A review on model reference adaptive control of robotic manipulators. Annual Reviews in Control, 43, 188-198.

Zhang, G., Zhao, Y., Fu, P., Luo, W., Shao, Q., Zhang, T., & Yu, Z. (2024). A reliable unmanned aerial vehicle multi-target tracking system with global motion compensation for monitoring Procapra przewalskii. Ecological Informatics, 81, 102556.

Zhang, H., Wang, R., Dong, P., Sun, H., Li, S., & Wang, H. (2021). Beef cattle multi-target tracking based on DeepSORT algorithm. Trans Chin Soc Agric Mach, 52, 248-56.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. International journal of computer vision, 129, 3069-3087.