**ORIGINAL RESEARCH** **Open access**

# Enhanced YOLOv7 for UAV-Based Small Object Detection Using Dynamic Snake Convolution and EIoU Loss

Luca Ferraro[1*], Matteo Ricci[1], Giulia Moretti[2], Paolo Greco[3], Antonio Bianchi[4]

## Abstract

The swift progress of remote sensing technology has introduced multiple challenges to target detection, one key issue being the persistent lack of effective solutions for small targets in complex backgrounds with simple features. To tackle this, we have integrated dynamic snake convolution (DSC) into the YOLOv7 architecture. Additionally, the original spatial pyramid pooling structure has been replaced with SPPFCSPC, and the original loss function has been replaced with the EIoU loss function. This work was assessed on UAV image data (VisDrone2019) and benchmarked against mainstream algorithms. The experiments revealed that the proposed algorithm achieves favorable average precision. Specifically, compared with the original algorithm, the mAP0.5 of the present algorithm shows a 4.3% gain. Trials confirmed that this algorithm delivers superior results compared to alternative methods.

**Keywords** Object detection, YOLOv7, UAV, Dynamic serpentine convolution, SPPF

*Correspondence:
Luca Ferraro
luca.ferraro@unibo.it
[1] Department of Artificial Intelligence for Olive Oil Milling and Quality Traceability, Faculty of Agriculture, University of Bologna, Bologna, Italy
[2] Department of Precision Vineyard Terroir Mapping, Faculty of Agriculture, University of Turin, Turin, Italy
[3] Department of Smart Storage and Post-Harvest Loss Prediction, Faculty of Agriculture, Sapienza University of Rome, Rome, Italy
[4] Department of Machine Learning for Pest Trapping, Faculty of Agriculture, University of Naples Federico II, Naples, Italy

## Introduction

The application of UAV remote sensing has grown increasingly pervasive across numerous sectors in recent years. It demonstrates remarkable effectiveness in contexts such as battlefield reconnaissance, disaster relief, environmental assessment, power line maintenance, and surveillance inspection. Deploying drones has vastly elevated task completion efficiency. Although remote sensing imagery provides notable gains in resolution and precision over conventional satellite-based methods, it still struggles with significant standoff distances, small targets, heavy occlusion, and weakly distinguishable features. Moreover, the limited payload capacity of UAVs makes it difficult for airborne edge computing platforms to meet the computational requirements of typical deep learning models, thereby presenting an additional practical challenge.

Target detection from the UAV perspective is inherently difficult because UAV imagery features scale variability, irregular spatial distributions, and a larger share of small targets, which creates a tension between the high processing needs of high-resolution UAV imagery and the constrained computing capability of current low-power chips. When contrasted with ground-level natural imagery, the expansive viewpoint of a UAV provides richer contextual data but also entails more complex scenes and a greater diversity of objects, introducing additional, unproductive noise that interferes with detection. Moreover, from an overhead perspective, objects are often harder to spot due to factors such as long-range capture, background clutter, and lighting conditions, making high-resolution images indispensable. This practice

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

significantly increases both the computational load and the memory footprint of detection algorithms. Directly applying generic, non-specialized detection algorithms results in prohibitive computational and memory demands, further compounding the detection challenge. Practical application scenarios often require fine-grained distinctions, such as classifying vehicle models, and visually similar objects pose a formidable challenge for the model's recognition capabilities.

Classical target detection pipelines include region selection, feature extraction, and classification. A candidate region is first located within the input image, and its features are then extracted and classified. As the object's location, aspect ratio, and dimensions are unknown in advance, sliding windows of varying scales must traverse the entire image. While this technique can localize potential targets, it suffers from high computational complexity, window redundancy, and weak region alignment, all of which degrade the speed and quality of subsequent feature extraction. Indeed, due to time-complexity concerns and for targets with widely varying aspect ratios, properly matched regions remain elusive even after exhaustive traversal. In the feature extraction phase, descriptors such as local binary patterns, the scale-invariant feature transform, and histograms of oriented gradients are commonly used. The indeterminate morphology of targets, variability in illumination, and the intricacy of backgrounds make it extremely difficult to design robust features. Consequently, classical methods produce unstable results, are highly susceptible to environmental conditions, and are hard to operationalize effectively.

With ongoing technological advances, visual target detection has become a critical component of practical systems. The industry has recently seen the rise of several related unicorn companies, including Shangtang Technology and Kuangwei Technology. Concurrently, computer vision has become a cornerstone of autonomous driving, with firms such as Tesla exemplifying how visual perception drives progress in this domain. Despite considerable advances in UAV-based visual inspection, it still encounters many hurdles. Chiefly, aerial imagery differs substantially from natural scene imagery, making precise recognition difficult. Secondly, the human–machine target-detection task imposes stringent demands on both speed and accuracy.

To remedy the aforementioned issues, this work refines YOLOv7. Experiments on the public VisDrone-2019 dataset verified that the modified algorithm attains high detection precision. First, dynamic snake convolution (DSC) is incorporated into the Backbone, yielding a substantial boost in model accuracy. Second, an enhanced SPPF module replaces the standard spatial pyramid pooling structure. Finally, the original loss function is superseded by EIOU.

# Related Work

## Targeted detection

As a fundamental task in computer vision, object detection directly dictates the efficacy of downstream applications and has therefore drawn intense scrutiny from diverse industries and research communities. In academia, object detection is a focal topic in major computer vision publications, with thousands of new papers emerging each year. According to Google Scholar, more than 15,000 papers concerning object detection have been published in the last decade. In the industrial realm, technology giants—including Google, Shangtang, Kuangyi, Facebook, Huawei, and Baidu—have invested substantial resources and engineering talent in this area. From a governmental standpoint, object detection is categorized as a strategic artificial intelligence technology, prompting active competition and development across nations.

Initially, detection algorithms predominantly relied on handcrafted features and shallow classifiers, such as AdaBoost [1]. This era gave rise to landmark feature descriptors like Haar-like features and histograms of oriented gradients. Since 2012, the rapid evolution of deep learning, coupled with advances in computational hardware and the public release of large-scale datasets and benchmarks, has spurred a wave of seminal works. These include region-based convolutional neural networks [2], SSD [3], YOLO [4], and DETR [5]. In contrast to the manual feature engineering paradigm, deep learning obviates the tedious design cycle, enabling automatic acquisition of discriminative features. Furthermore, such

© 2024 The Author(s). | Published by Agro Intelligence Research Press

methods unify feature learning and classifier training within a single architecture. These breakthroughs are driving swift and transformative progress.

Single-stage detectors subdivide the image into a grid of cells, with each cell predicting the presence, category, and position of an object, as exemplified by the YOLO family and SSD [3]. Two-stage methodologies separate the task: an initial step generates a sparse set of candidate regions likely to contain objects, and a subsequent step classifies and precisely localizes these proposals before rendering a final verdict. Faster R-CNN [6] is a representative example. Two-stage detectors trade off real-time performance for higher accuracy and routinely excel on numerous benchmark datasets.

## Unmanned aerial vehicle target detection

Adopting a UAV viewpoint creates a series of obstacles for multi-target detection, among them a proliferation of small objects, a scarcity of feature richness within single-dimensional data, reduced detection throughput due to sparse and non-uniform class distributions, corruption of detection by disturbances, missed and erroneous detections triggered by scale fluctuations, and slow inference latency. This portion surveys the refinements that scholars have advanced to counter these challenges from two angles.

For multi-target detection from the UAV viewpoint, single-stage frameworks from the YOLO lineage and SSD enjoy broad adoption due to their distinct strengths. A variety of researchers have confronted viewpoint-specific algorithm hurdles. (1) To manage the prevalence of minute objects in overhead imagery, Liu *et al.* [7] integrated Res Unit_2 into the Backbone and ResNet unit inside YOLO, fusing a pair of ResNet units within Darknet's Resblock while simultaneously lowering the chance of misses tied to a cramped receptive field, thereby remedying small-target dropouts that stem from a restricted sensing area. (2) Saetchnikov *et al.* [8] advanced the YOLOv4 eff architecture, which harnesses four cross-stage partial linkage sets to bridge the Backbone and neck, activates with the Swish function, and fixes the letter-box at 1 to uphold operational efficiency. (3) Striving to overcome erroneous detections in UAV aerial frames brought on by scale shifts, Li *et al.* [9] engineered an SSD variant that couples an attention module with dilated convolution, swapping the original convolution for the dilated version and merging shallow feature maps of small targets with deeper ones.

Two-stage detection pipelines stand apart from single-stage ones. Porting an algorithm from a standard viewpoint to UAV aerial footage yields poor performance and requires tailoring to the characteristics of objects captured in UAV video. The chief improvements are condensed here: (1) Avola *et al.* [10] designed a multi-stream topology for multi-scale image trials to cope with aerial scenes teeming with small items. Adopting this topology as the Fast R-CNN backbone yielded an MS-Faster R-CNN detector that uniformly and robustly pinpoints targets across UAV image sequences. Stadler deployed a Cascade R-CNN as the detector, reduced the default anchor dimensions by half to accommodate tinier targets, and doubled the number of predicted objects [10]. (2) Seeking to remedy the shortfall of single-dimensional feature content, Azimi *et al.* [11] enlisted a Siamese network to pull out visual cues and operated alongside graph convolutional networks and LSTM to fuse the shape, relational, and temporal signatures of targets. (3) To offset the sluggishness caused by the scattered placement of objects in aerial settings, Yang *et al.* [12] wove a clustering philosophy into detection work, unveiling the ClusDet framework, which first spawns target cluster zones via the CPNet clustering network, gauges the scale of targets within those zones through the ScaleNet network, and subsequently feeds the clustered regions into the DetecNet network to carry out detection, thus slashing the computational burden.

# Principles and Improvements

## YOLOv7

YOLOv7 falls under the single-stage detection paradigm and ranks among the most cutting-edge algorithms, outstripping previous YOLO versions. The YOLOv7 model comprises four core components: Input, Backbone, Neck, and Head [13].

© 2024 The Author(s). | Published by Agro Intelligence Research Press

The Backbone fuses Conv2D_BN_SiLu (CBS), a high-efficiency layer aggregation network (ELAN), a maximum-pooling layer (MP), and SPPCSPC. The convolutional portion bundles Conv2D, batch normalization (BN), and Selu—an amalgam of Sigmoid and linear rectification—to harvest image features at varied scales. The maximum-pooling layer (MP) is divided into twin branches. The left branch is downsampled via max pooling and then traverses a convolutional section to trim the channel count; the right branch runs through a 1 × 1 convolutional section and is instantly followed by a 3 × 3 convolutional section for downsampling, with both branches concatenated to sharpen object-focused extraction. The high-efficiency layer aggregation network (ELAN) relies on a four-branch stack, each branch hosting a distinct tally of Conv2D_BN_SiLu components, yielding a richer residual layout that eases optimization and helps dampen the gradient explosion threat inherent to deepening networks.

The Neck segment operates a path-aggregation network (PA-Net) blueprint to extract three feature tiers—middle, middle-lower, and bottom—at distinct network depths from the Backbone, and subjects them to a sequence of convolutions, max up-sampling, max down-sampling, and ELAN modules to amplify feature information.

The Head segment applies a RepConv operation to each of the three enriched feature maps from the Neck. This introduces a specialised residual arrangement that maintains predictive fidelity as the model's complexity decreases. The outcome then passes to YoloHead, which finalises class predictions and anchors object bounding boxes. **Figure 1** lays out the entire YOLOv7 network structure.

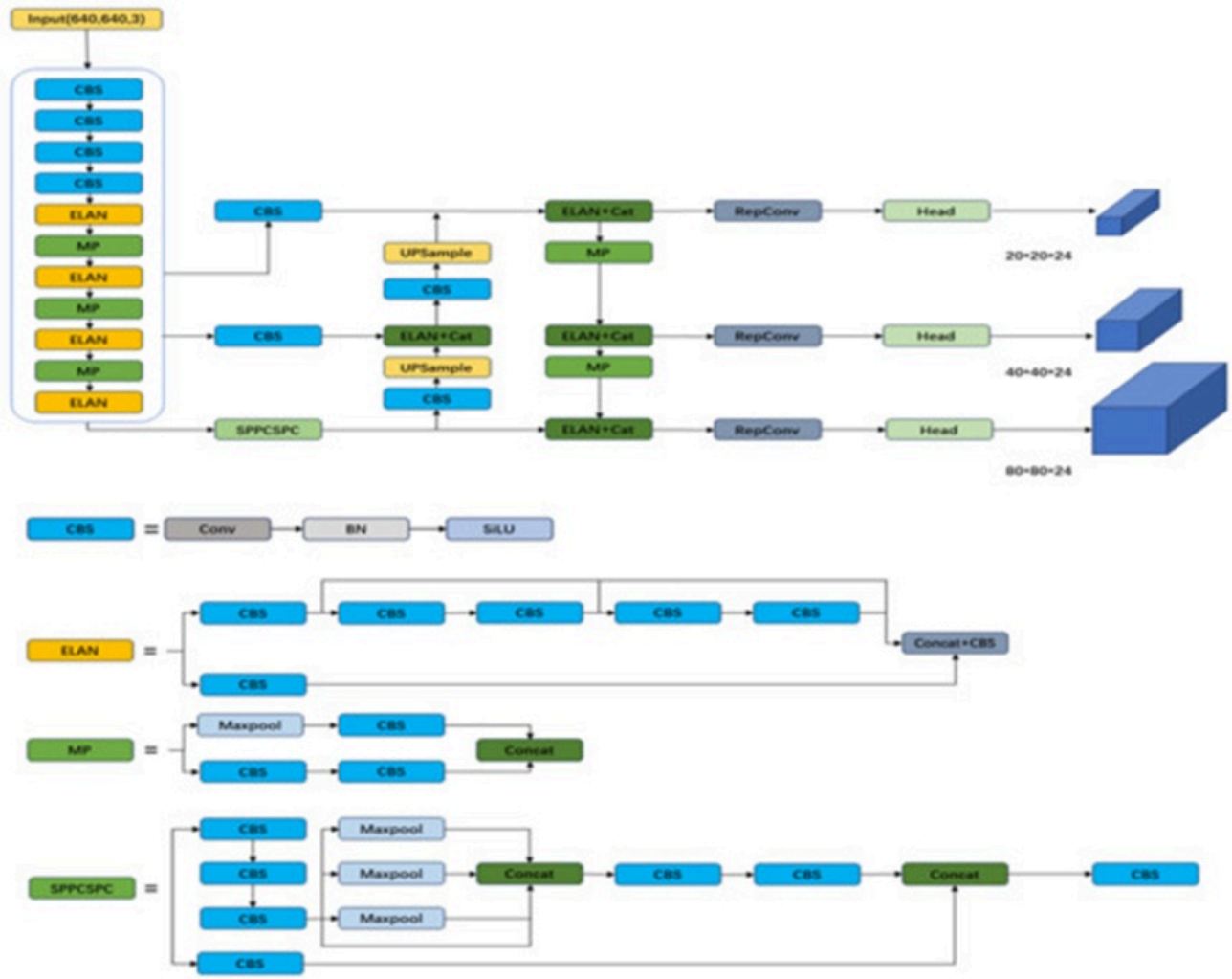



**Figure 1.** YOLOv7 network structure diagram.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

# Improvement

## Dynamic snake convolution

UAV remote sensing settings teem with extended, sinuous tubular dominant formations, gossamer and fragile local textures, and erratic, labyrinthine global arrangements. The standard convolution kernel aims to capture a subset of traits, such as local patterns. Building upon that, deformable convolution extends its usefulness by flexing to the geometric warping of different entities. Still, honing in on slim, meandering tubular shapes remains problematic due to the hurdles outlined earlier. Hence, we adopt the fresh DSCNet framework, which packs a tubular-sensitive dynamic snake convolution kernel, a multi-angle feature fusion tactic, and a topology-continuity penalized loss function. We unpack the derivation of the dynamic snake convolution's formulation below [14].

First, taking the conventional convolution coordinates $K$, the central coordinates are fixed. It is then written as:

$$K = \{(x-1, y-1), (x-1, y), \cdots, (x+1, y+1)\} \quad (1)$$

To let it keep fixating on geometrically intricate object features, a deformation offset Δ is injected. Were the above model left wholly unshackled to learn offsets, those offsets would tend to be unfettered and especially liable to stray from the target when contending with stretched-out structures. Accordingly, an iterative scheme is pursued, serially projecting the next target location to be treated, thereby safeguarding continuity in handling the target and forestalling sizeable offsets.

Second, the convolution kernel is linearised along both the x-axis and y-axis. Picking each grid point inside the kernel $K$ carries a compounding influence. Originating from the start spot $Ki$, the coordinates of locations distant from the centre commonly lean on the previous grid point $Ki$+1, incorporating an offset Δ={$\delta$|$\delta$∈[−1,1]} with respect to $K$. The offsets thereby accumulate Σ and are deployed to prevent the kernel from rupturing the linear morphological shape.

As **Figure 2** shows, the shift in the x-axis orientation is:

$$K_{i\pm c} = \begin{cases} (x_{i+c}, y_{i+c}) = (x_i + c, y_i + \sum_{i}^{i+c} \Delta y), \\ (x_{i-c}, y_{i-c}) = (x_i - c, y_i + \sum_{i-c}^{i} \Delta y), \end{cases} \quad (2)$$

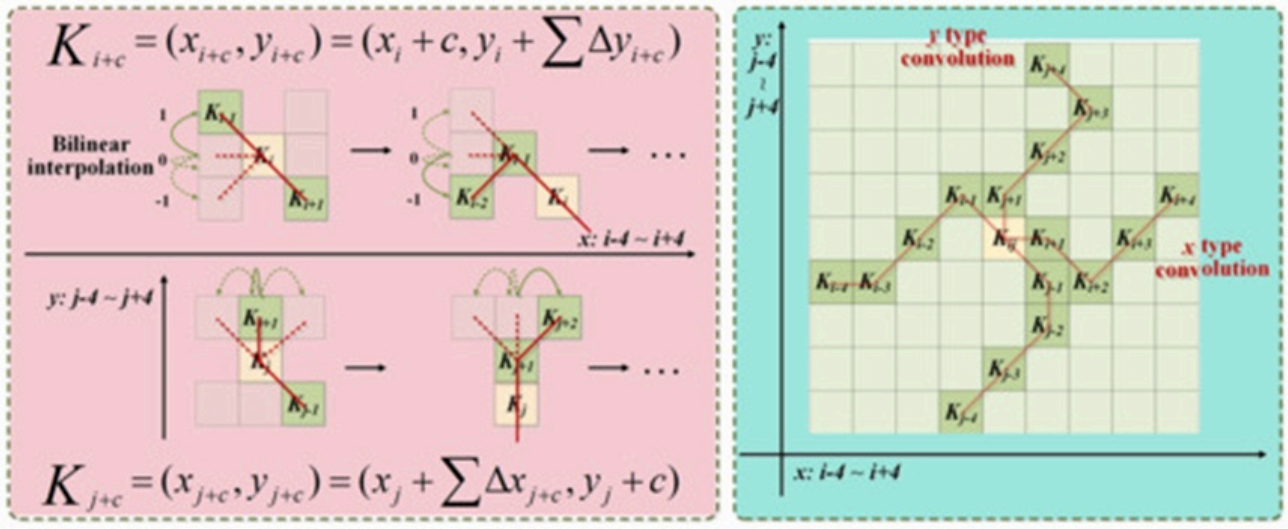



**Figure 2.** Schematic diagram of the dynamic serpentine convolution kernel coordinates computation and optional receptive fields.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

The shift in the y-axis orientation is:

$$K_{J\pm c} = \begin{cases} (x_{j+c}, y_{j+c}) = (x_j + \sum_{j}^{j+c} \Delta x, y_j + c), \\ (x_{j-c}, y_{j-c}) = (x_j + \sum_{j-c}^{j} \Delta x, y_j - c), \end{cases} \quad (3)$$

Given that offsets are at core fractional quantities, whereas coordinates gravitate toward integers, bilinear interpolation is applied as shown below:

$$K = \sum_{K'} B(K', K) \cdot K' \quad (4)$$

In the above expression, $K$ stands for the fractional seat of the convolution kernel on the coordinate plane, all points inhabit integer space, and $B$ functions as a bilinear interpolation kernel, thus factorable into a product of one-dimensional kernels:

In contrast to deformable convolution in **Figure 3**, deformable convolution grants the network unbounded freedom to absorb geometric distortions, leading to receptive field roaming, especially on fine tubular structures [15].

$$B(K, K') = b(K_x, K'_x) \cdot b(K_y, K'_y) \quad (5)$$

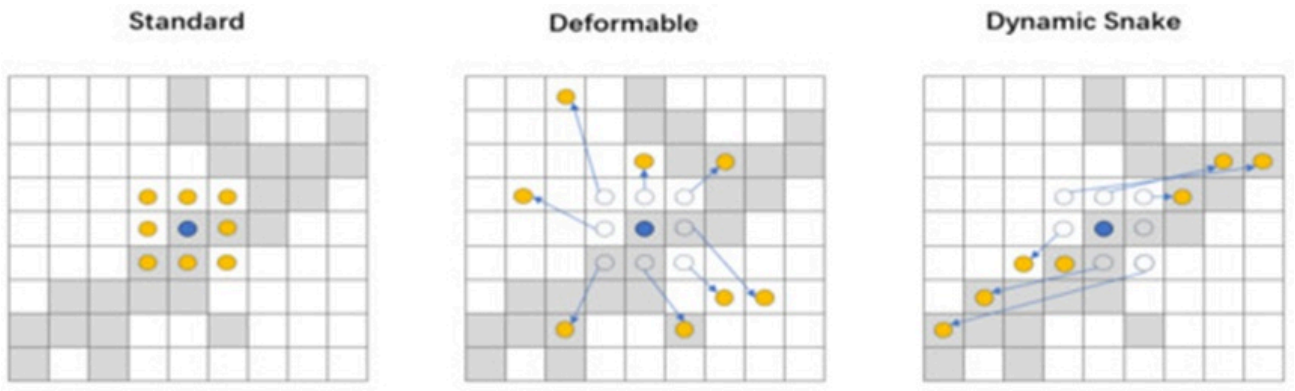



**Figure 3.** Receptive fields for standard convolution, deformable convolution, and dynamic serpentine convolution.

DSConv zeroes in on the curvilinear nature of pipeline geometries and explicitly reinforces the perception of pipeline morphology through constraint-guided adaptive learning. Cross-viewpoint feature amalgamation procedures are triggered when confronting the difficulty of an unstable overall shape. The approach crafts a diverse morphological kernel model via DSConv, examines the object's structural properties across multiple dimensions, and consolidates key elements to achieve robust feature fusion. The uptick in feature integration can strain network computation and lead to data redundancy, so a hierarchical random dropout scheme is enforced during the feature-blending training stage to reduce processing burden and curb overfitting. A topological coherence loss function (TCLoss) grounded in persistent homology (PH) is formulated to address the tendency of pipeline structures to lose connectivity. PH traces the birth-to-dissolution trajectory of topological signatures and polishes salient topological details buried within noisy data.

## Refinements to SPPCSPC

Within the bottommost layer of the Backbone, PANET uses SPPCSPC, a component that combines spatial pyramid pooling (SPP) with cross-stage partial (CSP) pyramid junctions. This structure slices convolutional kernels into four distinct dimensions—1 × 1, 5 × 5, 9 × 9, and 13 × 13—to perform Maxpool operations that separate dissimilar objects,

© 2024 The Author(s). | Published by Agro Intelligence Research Press

expand the receptive field, and extract critical feature cues. Although its capabilities exceed those of the SPPF module in YOLOv5, it places greater strain on inference latency [16]. Motivated by this observation, the present study borrows the streamlined design philosophy of SPPF, which substantially accelerates inference with negligible degradation to detection fidelity. Accordingly, as illustrated in **Figure 4**, the current work overhauls SPPCSPC by consulting the SPPF network template, implementing the following modification: three kernels of 5 × 5, 9 × 9, and 13 × 13 are arranged to execute Maxpool operations one after another in series, thereby reducing structural complexity while simultaneously aggregating target information across all scales [17].

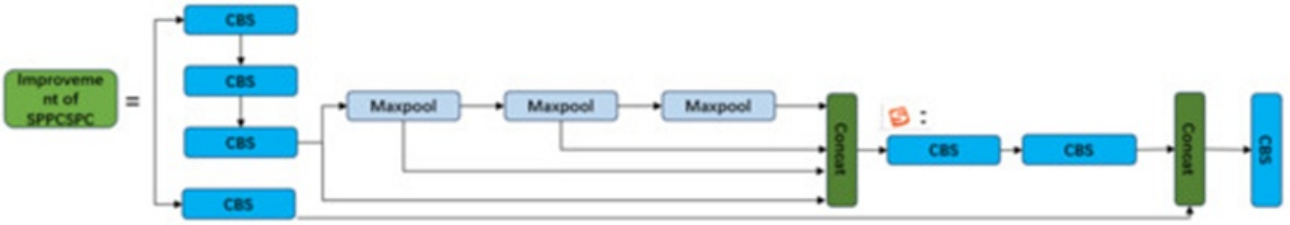



**Figure 4.** Improved SPP network.

## Refinements to IoU loss

The YOLOv7 stock employed CIoU as the coordinate regression loss [18]. As shown in **Figure 5**, the original formulation combines three geometric components: the overlapping region, the gap between centroids, and the aspect ratio. The loss is specified between the forecasted and ground-truth rectangles as set out below.

$$L_{CIOU} = 1 - IOU + \frac{p^2(b, b^{gt})}{c^2} + \alpha v \quad (6)$$

In these expressions, $b$ and $b^{gt}$ correspond to the centre points of B and $B^{gt}$, respectively, $(\cdot)=||b-bst||2$ stands for the Euclidean distance. At the same time, $c$ is the diagonal span of the minimal bounding rectangle that encloses both boxes.

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \quad (7)$$

Meanwhile,

$$\alpha = \frac{v}{(1-IOU)+v} \quad (8)$$

Captures the mismatch in aspect ratios.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

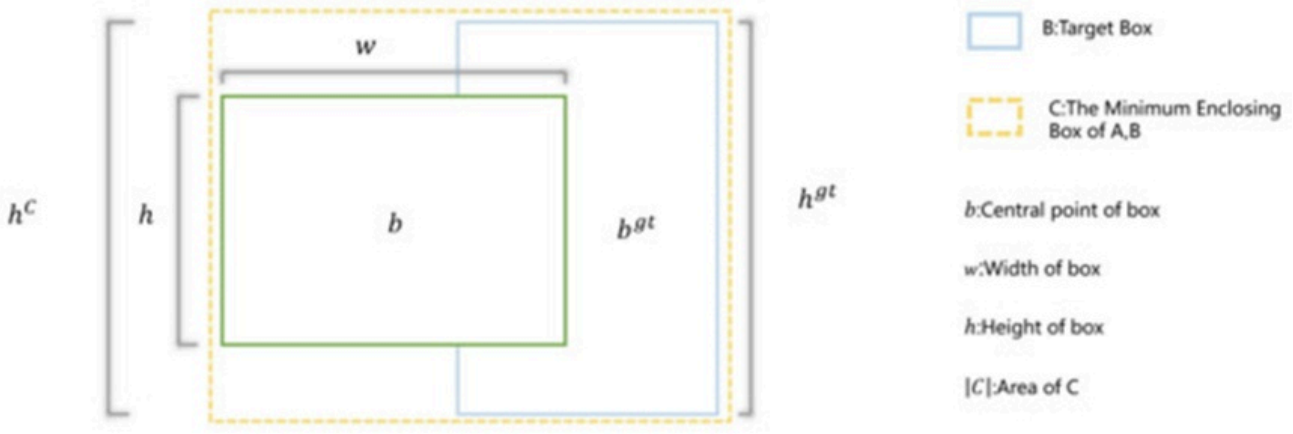


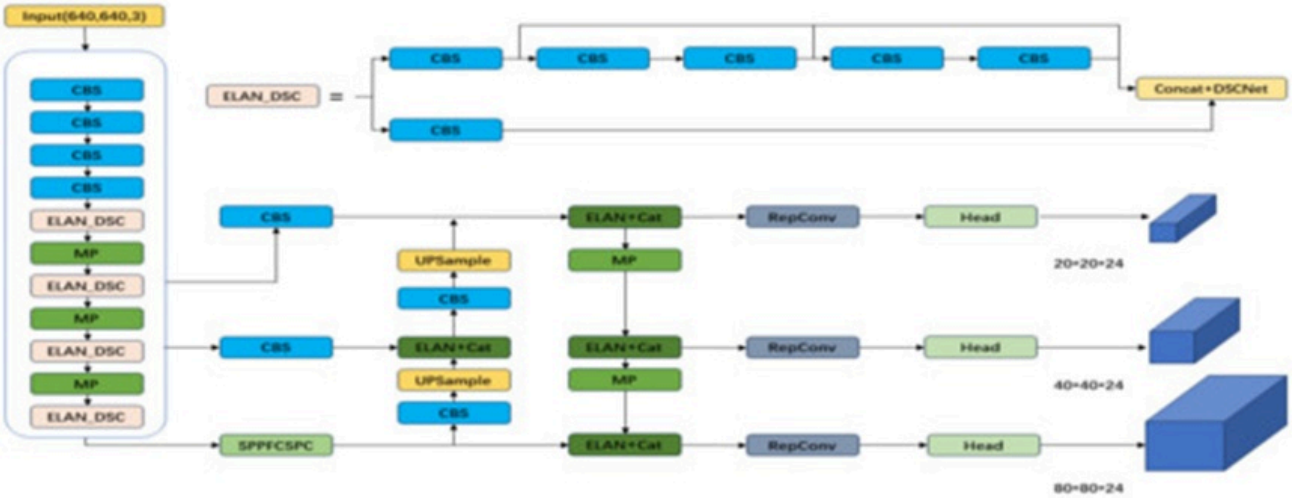

**Figure 5.** Loss function.

When placed alongside earlier loss variants, CIoU delivers a notable leap in both convergence rate and classification precision. That said, the formulation and refinement of its trailing term remain inadequately clarified, a shortcoming that drags down CIoU's convergence velocity on multiple fronts. (1) The quantity $v$ reflects aspect ratio divergence exclusively rather than the genuine association between $w$ and $wgt$ or $h$ and $hgt$; stated differently, for any collection of bounding boxes satisfying $\{(w=kw^{gt},h=kh^{gt})|kR^{+}\}$, the value of $v$ stays at zero, rendering the approach unworkable. (2) Expressed via the relationship $\frac{\partial v}{\partial w}=-\frac{h}{w}\frac{\partial v}{\partial h}$, the term $\partial v/\partial v$ moves in the reverse direction of $\partial v/\partial h$. If either dimension $w$ or $h$ grows, the remaining one is forced to contract, a behaviour that defies logical expectation. (3) Since $v$ signals disparity in aspect ratio, any resulting penalty or benefit steers similarity optimisation along an undesirable path.

For these reasons, we switch to the EIoU loss as proposed by Zhang *et al.* [19].

$$\frac{\partial v}{\partial w}=-\frac{h}{w}\frac{\partial v}{\partial h} \quad (9)$$

In the notation above, $h_w$ and $h_c$ indicate the breadth and height of the smallest enclosing rectangle spanning both boxes. Put another way, the sub-penalty function splits into three segments: $IOU$ together with $LIOU$, the separation penalty $Ldis$, and the orientation penalty $Lasp$. As a consequence, the desirable properties of the original penalty are preserved. At the same time, the function directly reduces the height–width gap between the predicted and anchor boxes, enabling faster convergence and sharper localisation. The fully enhanced network blueprint is shown in **Figure 6**.




© 2024 The Author(s). | Published by Agro Intelligence Research Press

**Figure 6.** Improved YOLOv7 overall network.

# Analysis of Experimental Results

## Datasets

To measure the algorithm's capability, the research team conducted evaluations on the VisDrone2019 benchmark. Assembled by the AISKYEYE group at Tianjin University, this resource comprises 288 videos, 261,908 frames, and 10,209 static pictures. The data were gathered from numerous drone photography sources, spanning a wide variety of environments, including 14 Chinese cities over considerable distances, a mix of rural and urban backdrops, and assorted object categories (such as pedestrians, vehicles, and so forth) appearing at varying densities.

VisDrone2019 contains 6471 training pictures, 548 validation pictures, and 1610 test pictures, spanning an extensive assortment of traffic configurations—motorways, intersections, and T-junctions among them—as well as a rich diversity of climatic and illumination settings, from clear day and pitch night to haze-filled and rainy atmospheres. The dataset is well-suited to validating detection performance on small ground-level targets captured from UAVs. Every technique in the experiment is learned on the training partition and scored on the validation partition.

## Experimental steps

**Table 1** lists the hardware configuration used for the experiments: an Intel(R) Core(TM) i9-113500F CPU clocked at 3.50 GHz, with model training on a GeForce GTX 3090 with 24 GB of video RAM and 40 GB of system RAM. The trials were run under Windows, programmed in Python 3.8.6, and built on the PyTorch 1.11.0 ecosystem, harnessing CUDA 11.6 to accelerate computation.

**Table 1.** Experimental parameter configuration.

| Configuration | Name | Type |
|---|---|---|
| | CPU | Intel(R) Core(TM) i9-113500F |
| Hardware | GPU | NVIDIA GeForce GTX3090 |
| | Memory | 40GB |
| | CUDA | 11.6 |
| Software | Python | 3.8.6 |
| | Pytorch | 1.11.0 |
| | Learning rate | 0.01 |
| | Image size | 640 × 640 |
| Hyperparameters | Workers | 8 |
| | Batch size | 16 |
| | Maximum training epochs | 300 |

The study used the Adam optimiser during model training to fine-tune the network's weights. The optimiser was configured with specific hyperparameter choices: a batch size of 16, a learning rate of 0.01, momentum of 0.937, weight decay of 0.0005, and a training duration of 300 epochs to ensure exhaustive network learning.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

This investigation employs the Adam optimiser to tune our model, combining the advantages of both Momentum and RMSprop for effective weight adaptation. The heart of Adam, or adaptive moment estimation, lies in its power to calculate both the average (first-order moment) and the uncentered variance (second-order moment) of the gradients. It subsequently customises the learning rate individually for each parameter, drawing upon these quantities. This design permits Adam to adjust its step size based on the parameter update history, thereby achieving faster convergence while improving both training throughput and stability compared to classical stochastic gradient descent (SGD) schemes.

## Evaluation indicators

The model leans on precision ($P$), recall ($R$), average precision ($AP$), and mean average precision ($mAP$) as yardsticks to judge its performance. AP measures per-class detection fidelity, while mAP is computed by totalling AP values across all classes and then dividing by the overall class count. In this study, mAP0.5 denotes the mAP at a threshold of 0.5, where IoU quantifies the overlap between the predicted and ground-truth bounding boxes.

$$P = TP/(TP + FP) \quad (10)$$

$$R = TP/(TP + FN) \quad (11)$$

$$AP = \int_0^1 P(R)dR \quad (12)$$

$$mAP = \frac{1}{N}\int_0^1 P(R)dR \quad (13)$$

In the sphere of model performance appraisal, $TP$ stands for the tally of positive instances correctly flagged as positive by the model. $FP$ denotes the count of negative instances mistakenly labelled as positive. Meanwhile, $FN$ represents the number of positive instances erroneously classed as negative. These quantities are indispensable for deriving precision, recall, and associated performance measures.

## Ablation experiments

To verify the merit of the newly inserted components, this work executed ablation trials on the VisDrone2019 dataset using YOLOv7 as the starting-point algorithm. These tests focused on recording mean average precision (mAP), parameter volume, and frames per second (FPS) as indicators of performance gains. The findings from these runs are presented in **Table 2**.

**Table 2.** Ablation experiment of the improved point.

| Method | mAP% | Parameters/M | R% | P% |
|---|---|---|---|---|
| YOLOv7 | 50.47 | 35.51 | 50.6 | 59.7 |
| YOLOv7_dscnet | 53.07 | 51.83 | 53.1 | 62.3 |
| YOLOv7_SPPF | 51.37 | 30.01 | 51.8 | 60.8 |

© 2024 The Author(s). | Published by Agro Intelligence Research Press

| | | | | |
|---|---|---|---|---|
| YOLOv7_EIOU | 50.90 | 35.51 | 51.2 | 60.3 |
| YOLOv7_dscnet_SPPF | 53.45 | 46.82 | 53.5 | 62.8 |
| YOLOv7_dscnet_EIOU | 53.28 | 51.83 | 53.4 | 62.5 |
| YOLOv7_SPPF_EIOU | 52.01 | 30.01 | 52.3 | 61.3 |
| Ours | 54.7 | 46.82 | 54.9 | 64.1 |

Seven ablation runs were performed under matched settings, elaborated here as per **Table 2**:

· The opening experimental run, representing the baseline model—namely the YOLOv7 algorithmic model—acts as the control, scoring an mAP of 50.47% on the VisDrone2019 dataset.

· The second run swaps out the baseline ELAN for the enhanced ELAN_DSC; the parameter volume swells by 16.32M, yet mAP climbs by 3.4%, precision edges up by 2.6%, and recall rises by 3.5% against the baseline. The core reasons for the model's uplift are as follows: the target, being a delicate structure, occupies a minuscule slice of the image, has a limited pixel budget, and is easily thrown off by cluttered surroundings. Such conditions make it difficult for the model to register fine-grained target shifts faithfully. Yet, integrating dynamic snake convolution into ELAN enables the network to focus tightly on long, curved objects, thereby boosting detection capacity. Since dynamic snake convolution offers finer segmentation and greater complexity compared to standard convolution, the parameter footprint of the revised module increases relative to the original model.

· The third run installs the improved SPPF module; a 1.3% gain over the baseline model is recorded, with precision advancing by 1.1% and recall by 1.2%, while the parameter count drops. The sources of mAP, P, and R improvements, along with the parameter reduction, are interpreted as follows: the revamped SPPF module subjects kernels of non-identical sizes to Maxpool operations, enabling it to tell apart different objects, widen the receptive field, and harvest extra salient feature signals; hence, mAP, P, and R all increase. The redesigned module processes kernel sizes in a chain rather than in parallel, thereby reducing model complexity and, consequently, trimming the parameter tally.

· The fourth run switches the loss function to EIOU; mAP increases by 0.43%, precision by 0.5%, and recall by 0.6%, while the parameter volume stays flat, owing to the network architecture remaining untouched and thus unaltered compared to the baseline. Interpreting the uplift in detection capability: the loss term directly shrinks the height and width differential between the predicted and anchor rectangles, yielding faster convergence and more exact localisation.

· Spanning the second through to the seventh ablation runs, the injection of DSCNet delivered the decisive lift, supplying a 3.4% mAP boost, as can be seen in **Figure 7**.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

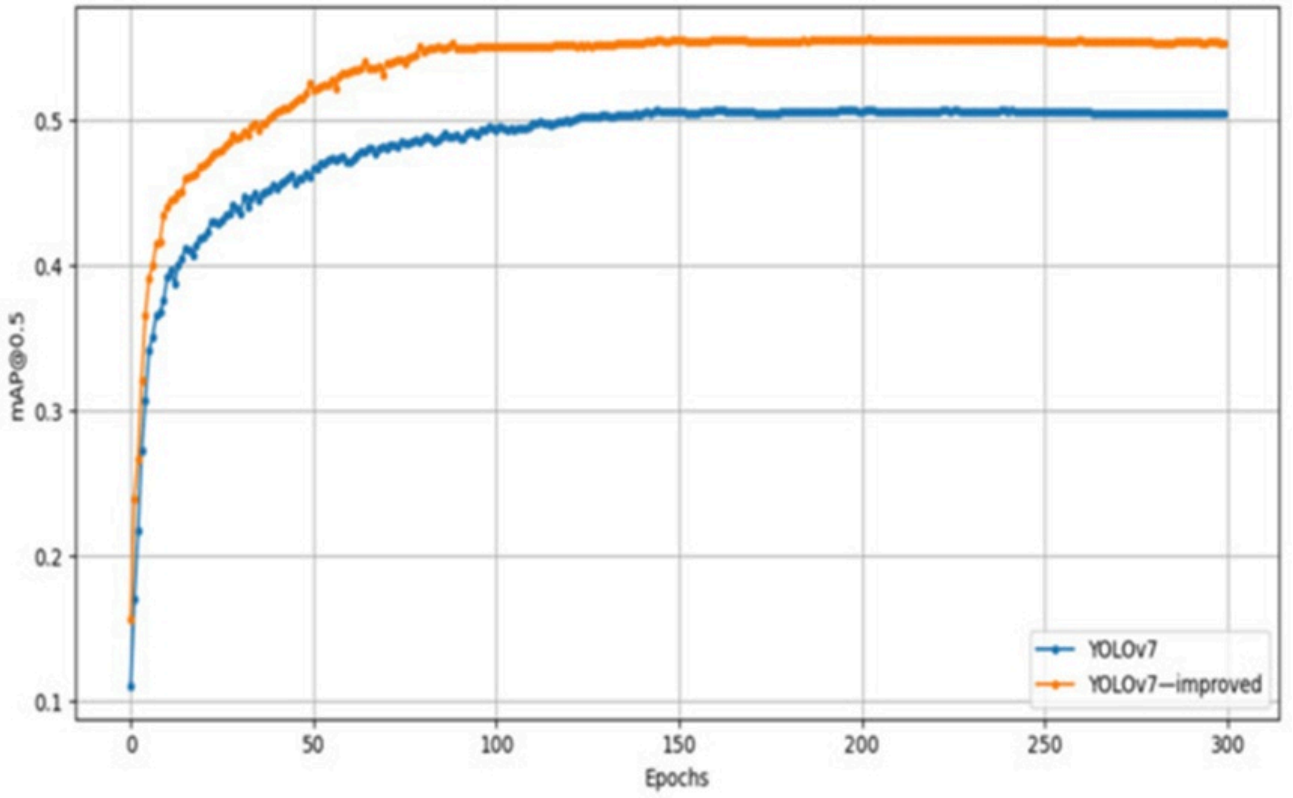



**Figure 7.** Comparison curve of mAP between the benchmark algorithm and the improved algorithm during the training process.

The figure that follows plots the progression and gain in mAP over the training span for both the final enhanced algorithm and the baseline. It is readily apparent that the original algorithm's mAP shoots up steeply across the first 50 epochs, then rises at a more measured pace from epoch 50 onward, plateauing at its ultimate training mAP near epoch 150, after which it settles. By contrast, the enhanced algorithm exhibits a steep mAP increase within the opening 30 epochs, a gentler ascent from epoch 30 to roughly epoch 90, and convergence thereafter. It is evident that the enhanced algorithm converges more quickly and achieves a 4.33% improvement in mAP over the baseline.

## Comparative experiments

A selection of UAV aerial image detection methods—among them YOLOv4, YOLOv3-LITE, YOLOv5s, Faster R-CNN, DMNet, and others—are pitted against the enhanced algorithm proposed in this study, with testing conducted on the VisDrone2019 test partition. Looking at **Table 3**, one can see that, compared with competing techniques, this algorithm posts a 33.0% mAP increase over Faster R-CNN, a 11.6% edge over YOLOv4, a 24.4% advantage over DMNet, and a 23.6% lift over YOLOv5s. Beyond these hefty mAP margins over mainstream detectors, the algorithm also outpaces others markedly in per-class AP; for example, detection accuracy on cars reaches 82.4%, on vans 58.6%, and on trucks 51.7%. Attributable to the properties of competing detection pipelines, these experimental outcomes demonstrate the utility and practicality of the algorithm for pinpointing faint, small objects in UAV aerial photographs. The experiments underscore the algorithm's effectiveness and real-world value for detecting subtle targets in UAV-captured aerial imagery.

**Table 3.** Comparative experiments with different detection algorithms.

| Method | mAP | | | | | | x | | | | AP% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Motor | A-T | Tricycle | Truck | Bus | Van | Car | Bicycle | Person | Pedestrian |

© 2024 The Author(s). | Published by Agro Intelligence Research Press

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Faster RCNN | 21.7 | 20.7 | 7.7 | 13.1 | 19.0 | 31.4 | 29.5 | 51.7 | 6.7 | 15.6 | 21.4 |
| YOLOv4 | 43.1 | 22.1 | 8.1 | 11.4 | 22.7 | 44.3 | 22.4 | 64.3 | 8.6 | 13.1 | 25.0 |
| CDNet | 34.2 | 29.3 | 25.4 | 33.0 | 38.2 | 49.5 | 42.1 | 55.8 | 13.8 | 19.2 | 35.6 |
| DMNet | 30.3 | 29.2 | 14.0 | 22.6 | 30.1 | 47.1 | 37.9 | 56.8 | 15.9 | 20.4 | 28.5 |
| RetinaNet | 13.9 | 11.8 | 4.2 | 6.3 | 11.5 | 17.8 | 19.9 | 45.5 | 1.4 | 7.9 | 13.0 |
| Cascade R-CNN | 23.2 | 21.4 | 8.6 | 14.8 | 21.6 | 34.9 | 31.5 | 54.6 | 7.6 | 14.8 | 22.2 |
| CenterNet | 26.2 | 23.7 | 17.4 | 20.1 | 21.3 | 37.9 | 24.0 | 59.7 | 14.6 | 20.6 | 22.6 |
| YOLOv3-LITE | 28.5 | 32.7 | 6.2 | 15.3 | 21.9 | 40.9 | 31.3 | 70.8 | 7.9 | 23.4 | 34.5 |
| MSC-CenterNet | 31.1 | 27.5 | 21.6 | 29.2 | 34.1 | 42.2 | 40.5 | 55.2 | 12.1 | 15.2 | 33.7 |
| DBAI-Det | 28.0 | 16.6 | 16.9 | 23.4 | 41.4 | 31.9 | 38.0 | 47.4 | 14.7 | 12.8 | 36.7 |
| YOLOv5s | 31.1 | 35.4 | 11.1 | 20.6 | 29.5 | 41.0 | 31.5 | 65 | 10.1 | 30.5 | 35.8 |
| YOLOv7 | 50.47 | 55.8 | 19.9 | 39.8 | 46.5 | 59.7 | 53.4 | 77.3 | 26.1 | 46.5 | 53.3 |
| Ours | 54.7 | 60.4 | 24.6 | 44.4 | 51.7 | 64.7 | 58.6 | 82.4 | 29.3 | 51.1 | 57.6 |

In addition, to highlight the forward-looking nature of the proposed approach and to situate it alongside state-of-the-art YOLOv7-derived algorithms in the UAV domain, YOLOv7-UAV [20], PDWT-YOLO [21], and a previously improved YOLOv7 variant are selected for direct comparison [22]. Inspection of **Table 4** shows that while YOLOv7-UAV surpasses the proposed algorithm in parameter economy, the proposed method outperforms both PDWT-YOLO and the prior improved YOLOv7 variant on this criterion. Beyond that, the proposed algorithm outperforms all the aforementioned methods on the mAP metric, achieving 54.7% on the VisDrone2019 dataset.

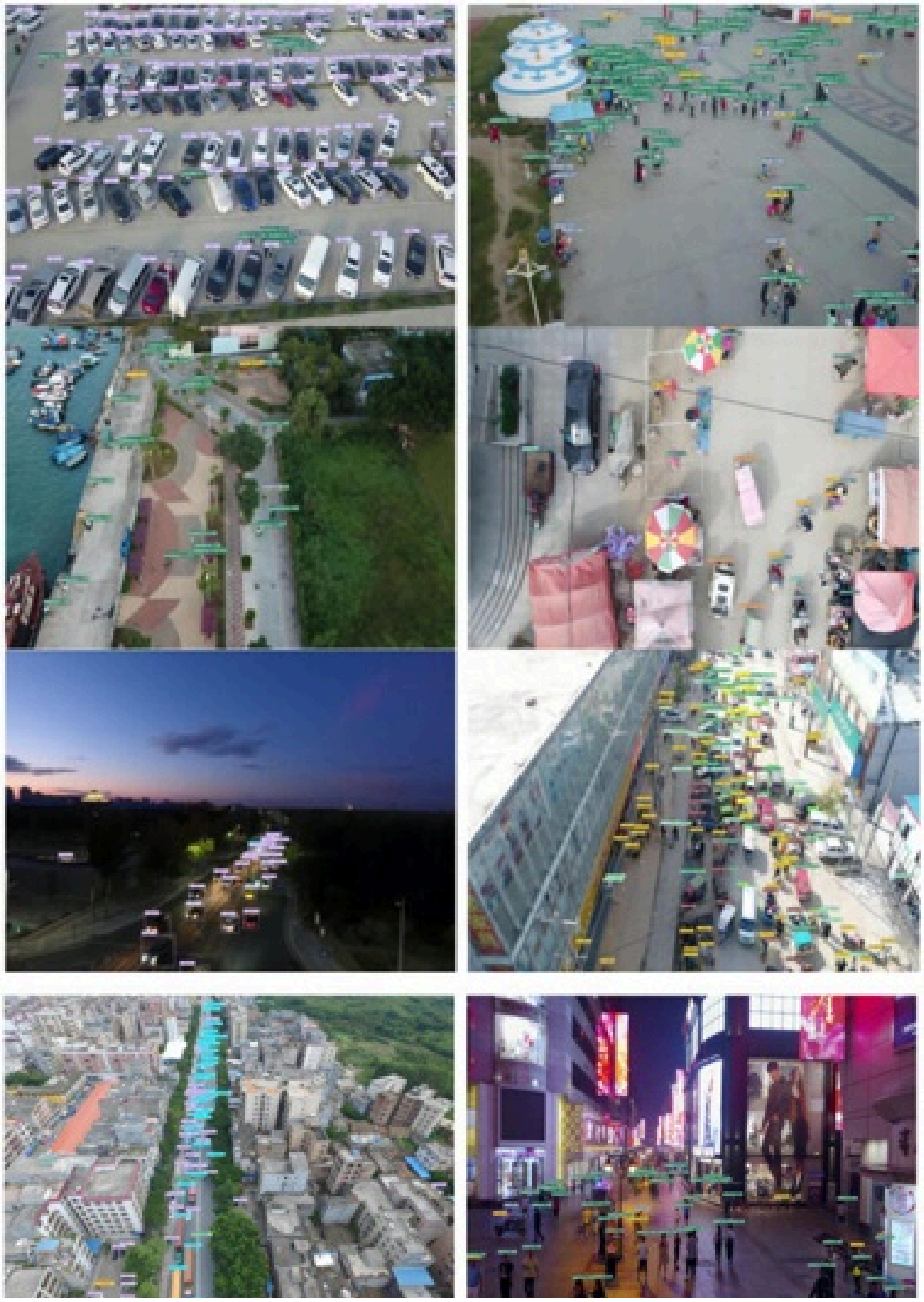
**Table 4.** Comparison of this algorithm with the latest improved algorithm based on YOLOv7.

| Method | mAP% | Parameters/M |
|---|---|---|
| YOLOv7-UAV | 52.21 | 3.07 |
| PDWT-YOLO | 41.2 | 24.2 |
| Improved YOLOv7 | 45.30 | 26.77 |
| Ours | 54.7 | 16.82 |

## Analysis of detection effects

Aerial photos taken by UAVs across a variety of demanding real-world scenes from the VisDrone2019 test partition are selected for evaluation, and the resulting detection outputs are displayed in **Figure 8**. One can observe that the algorithm advanced in this study suppresses interference from vegetation and structures within the busy image backdrop and properly segments and pinpoints the target, even when confronted with identical small objects set amid complex environmental contexts. This indicates that the proposed approach delivers more robust detection behaviour under authentic conditions, spanning varying illumination, dissimilar surroundings, and differing object distributions. Further, the confidence threshold is set to 0.25; image confidences below this cut-off are suppressed from view.

© 2024 The Author(s). | Published by Agro Intelligence Research Press



© 2024 The Author(s). | Published by Agro Intelligence Research Press



**Figure 9.** Small targets (baseline algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press



**Figure 10.** Small targets (improved algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press



**Figure 11.** Complex background (baseline algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press

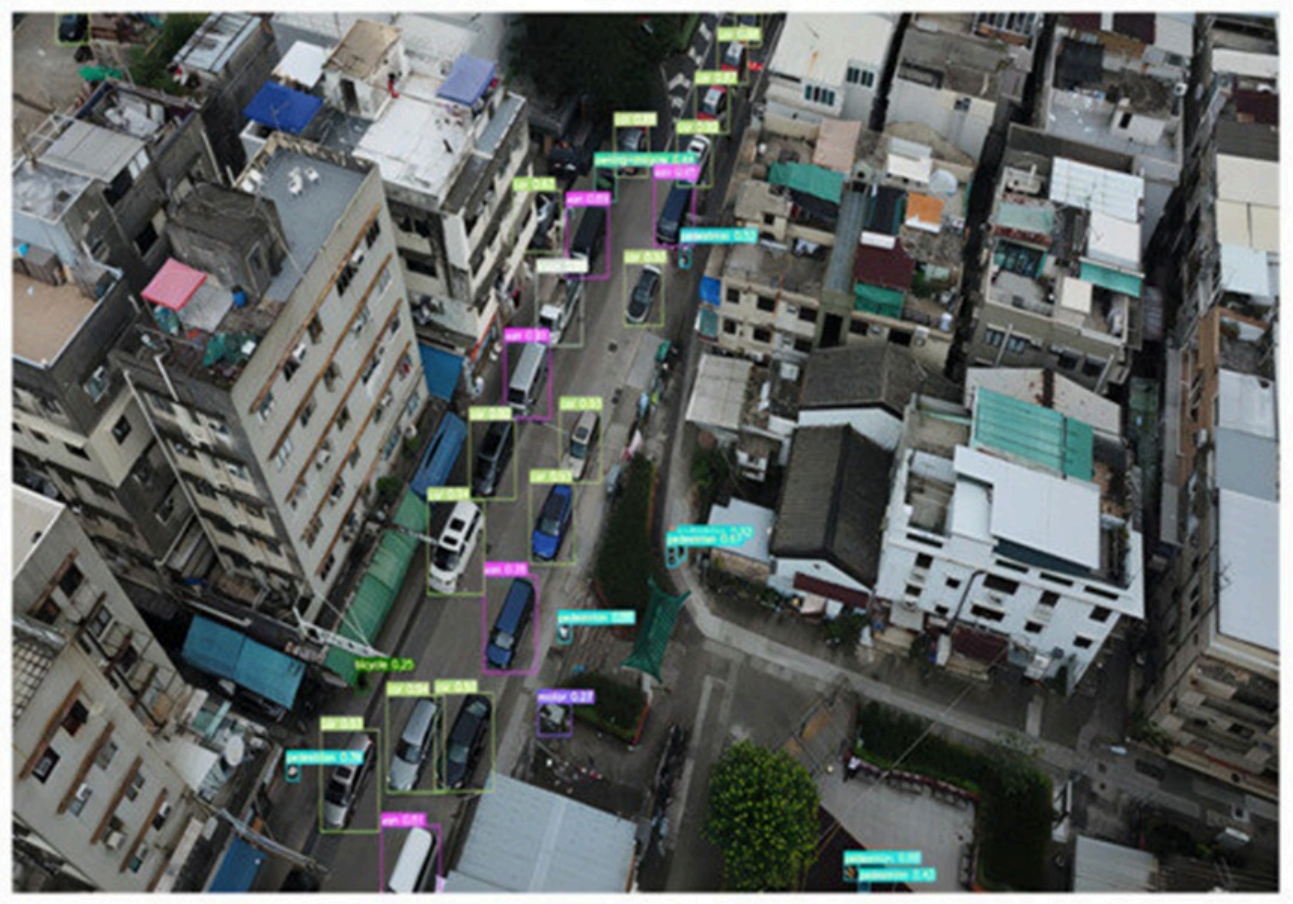


**Figure 12.** Complex background (improved algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press




**Figure 13.** Target occlusion (baseline algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press




**Figure 14.** Target occlusion (improved algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press



**Figure 15.** Dark background (baseline algorithm).

© 2024 The Author(s). | Published by Agro Intelligence Research Press



**Figure 16.** Dark background (improved algorithm).

# Conclusion

The present investigation successfully addresses the challenge of pinpointing small objects lodged within intricate surroundings—a persistent obstacle in UAV ground target detection—by integrating dynamic snake convolution (DSC), a reworked SPPCSPC built on the YOLOv7 framework, and introducing the EIoU loss formulation. Experimental validation confirms that the upgraded algorithm presented here delivers outstanding detection performance across a range of aerial imaging conditions and achieves top-tier results across all nine object classes, affirming its considerable practical value and operational effectiveness. Recognising the parallel processing requirements shared between satellite imagery interpretation and UAV aerial photography—particularly with respect to object detection, handling complex backgrounds, and detecting small targets—we deem the approach equally relevant to satellite image processing. Satellite images are widely used across geographic information systems (GIS), environmental surveillance, urban planning, and disaster response, all fields where the accurate detection and categorisation of small objects carries equal weight. While satellite images may differ from UAV captures in resolution and spatial scale, the refined algorithm detailed herein can nonetheless play a meaningful role in satellite image analysis, whether through proper parameter recalibration or minor structural adaptation.

# Future Work

Prospective lines of inquiry can gravitate toward the following focal domains:

© 2024 The Author(s). | Published by Agro Intelligence Research Press

(1) Application transfer and algorithmic generality: Probing the portability of the enhancement techniques crafted in this work—such as dynamic snake convolution (DSC), the re-engineered spatial pyramid pooling architecture (SPPFCSPC), and EIoU loss terms—onto alternative backbones, for instance, YOLOv8 or SSD. Study the manner in which these refinements can be moulded to fit the idiosyncrasies of different algorithmic blueprints, and how hyperparameters might be recalibrated throughout the transfer to uphold or even augment detection fidelity and throughput.

(2) Cross-domain applicability exploration: Moving past UAV image processing, investigate the promise of these improved algorithms when carried over to neighbouring fields, including satellite image examination and traffic flow surveillance. Place special emphasis on analysing algorithm behaviour when fed image data at disparate resolutions and spatial scales, and on tailoring algorithms to accommodate the bespoke demands of these emerging domains.

(3) Real-time throughput and edge deployment: Given that both UAV and satellite image analysis often impose real-time constraints, subsequent investigations could focus on lightweighting and computational acceleration to support edge computing environments. Delve into strategies for deploying deep learning architectures onto hardware-limited platforms while retaining efficient compute throughput and dependable detection precision.

(4) Multi-source data integration: In both UAV and satellite image analytics, heterogeneous data streams—such as optical, infrared, and radar modalities—are commonly encountered. Future research can examine methods for integrating these disparate data channels cohesively.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

Innovation Guidance Fund Project of Light Power Innovation Research Institute, Chinese Academy of Sciences (CXYJJ20-ZD-03); funded by National Key R&D Program (No. 2022YFF1302000).

## Ethics statement

None

Received: 19 Nov 2022 Revised: 24 Jan 2023 Accepted: 10 Feb 2023
Published online: 25 May 2024

### Rights and permissions

**Open Access** The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

# References

1. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. Stat Its Interface. 2009;2:349-60.

2. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA; 2014. p. 580-7.

3. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: Single shot multibox detector. In: Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands; 2016. p. 21-37.

4. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA; 2016. p. 779-88.

5. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision, Glasgow, UK; 2020. p. 213-29.

6. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile; 2015. p. 1440-8.

7. Liu M, Wang X, Zhou A, Fu X, Ma Y, Piao C. UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. Sensors. 2020;20:2238.

8. Saetchnikov I, Skakun V, Tcherniavskaia E. Efficient objects tracking from an unmanned aerial vehicle. In: Proceedings of the 2021 IEEE 8th International Workshop on Metrology for AeroSpace (MetroAeroSpace), Naples, Italy; 2021. p. 221-5.

9. Li Z, Liu X, Zhao Y, Liu B, Huang Z, Hong R. A lightweight multi-scale aggregated model for detecting aerial images captured by UAVs. J Vis Commun Image Represent. 2021;77:103058.

10. Avola D, Cinque L, Diko A, Fagioli A, Foresti GL, Mecca A, et al. MS-Faster R-CNN: Multi-stream backbone for improved Faster R-CNN object detection and aerial tracking from UAV images. Remote Sens. 2021;13:1670.

11. Azimi SM, Kraus M, Bahmanyar R, Reinartz P. Multiple pedestrians and vehicles tracking in aerial imagery: A comprehensive study. arXiv. 2020, arXiv:2010.09689.

12. Yang F, Fan H, Chu P, Blasch E, Ling H. Clustered object detection in aerial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea; 2019. p. 8311-20.

13. Wang CY, Bochkovskiy A, Liao HYM. 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, Vancouver, BC, Canada; 2023. p. 7464-75.

14. Qi Y, He Y, Qi X, Zhang Y, Yang G. Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France; 2023. p. 6070-9.

15. Dai J, Qi H, Xiong Y, Li Y, Zhang G, Hu H, et al. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy; 2017. p. 764-73.

16. Li C, Li L, Geng Y, Jiang H, Cheng M, Zhang B, et al. Yolov6 v3. 0: A full-scale reloading. arXiv. 2023, arXiv:2301.05586.

17. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37:1904-16.

18. Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans Cybern. 2021;52:8574-86.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

19. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. Neurocomputing. 2022;506:146-57.

20. Zeng Y, Zhang T, He W, Zhang Z. Yolov7-uav: An unmanned aerial vehicle image object detection algorithm based on improved yolov7. Electronics. 2023;12:3141.

21. Zhang L, Xiong N, Pan X, Yue X, Wu P, Guo C. Improved object detection method utilizing yolov7-tiny for unmanned aerial vehicle photographic imagery. Algorithms. 2023;16:520.

22. Li X, Wei Y, Li J, Duan W, Zhang X, Huang Y. Improved YOLOv7 algorithm for small object detection in unmanned aerial vehicle image scenarios. Appl Sci. 2024;14:1664.

© 2024 The Author(s). | Published by Agro Intelligence Research Press