Improving Real-Time UAV Target Recognition via Super-Resolution, InceptionNeXt, and Attention-Enhanced YOLOv8

Gangeshwar Mishra [1], Rohit Tanwar [2,*], and Prinima Gupta [1]

[1] Department of Computer Science and Technology, Manav Rachna University, Faridabad, Haryana, India
[2] School of Computer Science and Engineering, Shri Mata Vaishno Devi University, Katra, Jammu & Kashmir, India
Email: gangeshwarmishra045@gmail.com (G.M.); rohit.tanwar.cse@gmail.com (R.T.); prinima@mru.edu.in (P.G.)
*Corresponding author

***Abstract*—Accurate detection and recognition of clustered targets in aerial imagery remains difficult in Unmanned Aerial Vehicle (UAV) applications because targets are often small, blurred by low spatial resolution, affected by scale variation, partially occluded, and captured under uneven lighting. To address these issues, we present You Only Look Once (YOLO)v8-SR++, an improved UAV target recognition framework built on YOLOv8 that improves detection quality while preserving real-time speed. Rather than introducing entirely new component designs, the framework integrates and adapts several proven techniques within a single UAV-oriented pipeline. First, a Wide Activation Super-Resolution (WDSR) module enhances low-resolution inputs and restores fine details, helping the detector learn sharper features. Next, InceptionNeXt blocks are introduced in the backbone to strengthen multi-scale representation without heavy computation. In addition, a Separate and Enhanced Attention Module (SEAM) is placed in the neck to highlight important target regions and suppress background noise. Finally, to improve localization in dense and overlapping scenes, we use a hybrid loss that combines Complete Intersection-over-Union (CIoU) with Normalized Wasserstein Distance (NWD). Experiments on the VisDrone dataset show that YOLOv8-SR++ achieves 49.6% mAP@0.5 and 32.1% mAP@0.5:0.95, with 76.2% precision, 69.4% recall, and an F1-Score of 72.6%, while maintaining real-time inference at 43 FPS. These results indicate that YOLOv8-SR++ improves detection reliability for UAV aerial surveillance without sacrificing efficiency.**

***Keywords*—target recognition, Unmanned Aerial Vehicle (UAV), You Only Look Once (YOLO)v8, super-resolution, Wide Activation Super-Resolution (WDSR), Separate and Enhanced Attention Module (SEAM), Complete Intersection-over-Union (CIoU), drone**

I. Introduction

Recently, aerial imaging technologies have evolved rapidly, leading to a surge in applications involving Unmanned Aerial Vehicles (UAVs), high-resolution satellites, and airborne observation systems. These technologies provide critical infrastructure for operations such as disaster monitoring, urban planning, traffic management, agricultural inspection, and border surveillance. At the core of these applications lies the ability to reliably detect and recognize targets in complex and cluttered environments. However, accurate target recognition in aerial imagery remains a significant challenge in computer vision. Targets frequently reside in constrained pixel areas and are vulnerable to occlusion, blurriness, noise, and scale variation, leading to suboptimal predictions from traditional detectors.

Aerial imagery presents several inherent challenges for target recognition. First, the spatial resolution of UAV or satellite images is often low due to high-altitude capture, resulting in coarse texture details that hinder discriminative feature extraction by Convolutional Neural Network (CNN)-based models. Second, aerial environments are typically crowded with overlapping targets, where background clutter and inter-class similarity further degrade detection reliability. Third, variations in lighting, weather, and orientation introduce inconsistencies that can significantly impact recognition performance, particularly in real-time systems constrained by onboard processing power.

Traditional object detection frameworks such as Faster R-CNN, SSD, and early You Only Look Once (YOLO) models [1, 2] were not originally designed for aerial target recognition. Their reliance on coarse feature maps, large receptive fields, and fixed anchors often leads to missed detections and misclassifications. To overcome these limitations, recent research has explored advanced techniques such as multi-scale feature fusion, anchor refinement, and context-aware learning. Among them, Super-Resolution (SR), attention mechanisms, and loss optimization have emerged as highly effective methods for improving target recognition.

Super-resolution increases the resolution of input imagery before detection, restoring lost spatial information resulting from downsampling. Among super-resolution models, Wide Activation Super-Resolution (WDSR) achieves high-quality reconstruction based on wide

Manuscript received January 29, 2026; revised February 24, 2026; accepted April 1, 2026; published September 11, 2026.

activation layers and residual learning, where strong visual detail is typically critical for UAV-based imagery [3].

Complementing SR, attention mechanisms make the network focus on informative regions. The Separate and Enhanced Attention Module (SEAM) applies spatial and channel-wise attention to highlight the important visual cues and suppress the irrelevant background information, which helps enhance overall recognition accuracy.

InceptionNeXt introduces a lightweight, flexible, and multi-scale architecture inside the detection backbone, balancing computational efficiency with feature diversity [4]. Its parallel convolutional paths and grouped operations expand the receptive field, enabling the detector to capture rich contextual cues at different scales—an important ability in aerial environments with diverse object sizes and densities.

Bounding-box regression remains a vital part of object detection. For general objects, traditional Intersection-over-Union (IoU)-based loss functions, such as Generalized Intersection-over-Union (GIoU), Distance Intersection-over-Union (DIoU), and Complete Intersection-over-Union (CIoU), work effectively. but they can become unstable when targets are numerous or overlap. To address this, the Normalized Wasserstein Distance (NWD) metric was introduced. It measures the distance between bounding-box distributions instead of overlaps. Combining CIoU and NWD in a hybrid loss formulation results in more stable training and superior localization accuracy.

In this work, the proposed model is evaluated on the VisDrone dataset, which comprises thousands of real-world UAV images containing dense and complex scenes (Fig. 1). The experimental results show that YOLOv8-SR++ outperforms both the baseline YOLOv8 and all SR-based variants in terms of mAP, precision, and accuracy, without losing real-time inference performance. The remainder of this paper is structured as follows: Section II reviews recent advancements in target recognition, super-resolution integration, attention mechanisms, and loss optimization. Section III details the proposed YOLOv8-SR++ architecture. Section IV describes the experimental setup, dataset specifications, evaluation metrics, and ablation studies. Section V discusses results, findings, and limitations. Section VI concludes with final remarks and future research directions [5].

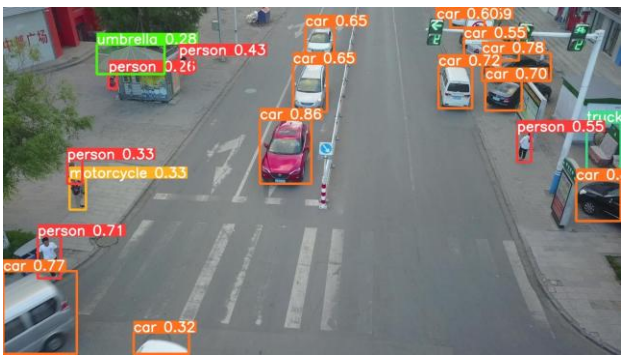


Fig. 1. An instance from the VisDrone collection and its annotations.

II. Literature Review

Target recognition has become an important problem in computer vision because it is used in many real-world applications such as surveillance, search and rescue, traffic monitoring, precision agriculture, and military reconnaissance. Unlike general object detection, where objects are large and occupy a significant part of the image, targets in aerial images are usually tiny and cover only a few pixels. Because of this, it becomes harder to extract useful features, bounding boxes are localized less accurately, and the system becomes more sensitive to occlusion and noise. As a result, traditional detectors designed for larger objects do not perform well under these conditions.

Over the past decade, object detection methods have advanced quickly. The YOLO family of models has played a key role in real-time detection. YOLOv1 introduced a unified framework that predicts bounding boxes and class probabilities directly from the image [6]. YOLOv2 and YOLOv3 improved this further by using anchor boxes, multi-scale detection, and residual connections. YOLOv4 boosted both speed and accuracy by refining the CSPDarknet53 backbone and applying stronger data augmentation. YOLOv5 from Ultralytics focused on modular design and easy deployment [7 8]. YOLOv8 then moved to an anchor-free design, decoupled classification and regression heads, and upgraded backbone components to increase flexibility and accuracy [9].

However, robust target recognition in aerial images is still challenging. Standard architectures often produce coarse feature maps, which cause small targets to disappear in deeper layers [10–12]. Anchor boxes may not match the actual object sizes, and NMS can mistakenly remove valid overlapping detections. These problems become more serious in UAV datasets, where targets like vehicles, pedestrians, and infrastructure elements are very small, appear at different angles, and are captured under changing lighting conditions.

To tackle these challenges, several benchmark datasets have been created for aerial imagery. The most commonly used ones include UAVDT, DOTA, VisDrone, and xView. These datasets present crowded scenes with complex backgrounds, challenging detection algorithms to their limits. For example, UAVDT targets drone-based traffic monitoring under different weather conditions, DOTA covers multi-oriented objects in satellite images, and VisDrone provides thousands of UAV images from varied environments, making it a strong benchmark for evaluating detection models [13, 14].

Super-Resolution (SR) has emerged as an important technique for improving detection performance. SR methods enrich low-resolution images by recovering high-frequency details such as textures and edges, which are crucial for accurate recognition. SRCNN was the first deep learning–based SR model and mapped low- to high-resolution images using a shallow convolutional network [15]. FSRCNN improved speed with deconvolution layers to support near real-time use. SRGAN introduced adversarial learning to generate visually realistic details [16]. EDSR removed unnecessary

normalization layers and achieved state-of-the-art PSNR performance. Later, RDN and SwinIR further improved feature propagation and representation using dense connections and transformer-based designs [17, 18].

In UAV scenarios, both computation and power are limited, so efficient models are essential. The WDSR uses wide activation layers to provide high accuracy with relatively low computational cost [9]. Lightweight GAN-based methods such as Fast-SRGAN have also been successfully used for aerial images [19]. Using SR as a preprocessing step for UAV imagery allows detectors to work with richer visual information, which improves both precision and recall [12, 20, 21].

Super-resolution can be used in two main ways: as a separate preprocessing step or as part of an end-to-end model. In preprocessing, the SR network first enhances the input image, and then a detector is applied. This is simple to implement but less flexible. In joint optimization, SR and detection are usually trained together in a single model [20]. This approach can produce sharper and more informative features, leading to higher detection accuracy [10]. Studies such as YOLO-SR and YOLO+SR show that even a lightweight SR module can significantly improve mAP and precision while adding only a small computational overhead. As a result, joint SR-detection frameworks are especially promising for UAV-based applications[12].

Backbone networks have also evolved. Early backbones such as VGG and ResNet were only designed for image classification, not for dense target recognition, which needs stronger multi-scale features. After that, Feature Pyramid Network (FPN) was introduced to improve multi-level feature fusion, and PANet added bottom-up paths to further enhance localization [22]. CBNetV2 went a step further by using multiple backbones together, increasing feature reuse and improving overall robustness.

Transformers brought global self-attention into vision tasks. ViT extended this idea directly to images [23], while Swin Transformers improved efficiency for dense prediction tasks. However, most of these transformer-based models are often computationally heavy and may not fit well on resource-limited UAV platforms. Hybrid architectures that combine convolutional layers with transformer modules offer a better balance between accuracy and efficiency. InceptionNeXt is one such design, enabling strong feature extraction for real-time aerial target recognition while keeping hardware demands manageable.

Attention mechanisms also play an important role by helping the network focus on the most informative regions and channels while suppressing noise. Squeeze-and-Excitation (SE) modules recalibrate channel-wise responses to emphasize important features. Convolutional Block Attention Module (CBAM) applies both channel and spatial attention in sequence to refine features further [24]. Transformer-based attention offers powerful global reasoning but typically increases computational cost [25, 26]. SEAM introduces parallel channel and spatial attention, improving localization and recognition in dense and cluttered UAV scenes [27].

Loss functions for bounding box regression have also been improved to handle localization difficulties. Early approaches used Smooth L1 and IoU-based losses, but these work poorly when predicted boxes do not overlap with the ground truth. GIoU resolved this by providing meaningful gradients even for non-overlapping boxes [28]. DIoU adds center-distance information to speed up convergence [29]. Focal Loss tackled class imbalance in dense detection tasks [30]. More recently, Normalized Wasserstein Distance (NWD) was proposed, modeling each bounding box as a probability distribution to provide smoother gradients and better training stability [31]. Combining CIoU with NWD gives a good balance between geometric accuracy and distribution-based stability, which is particularly useful for localizing small aerial targets.

Looking across the literature, several key points are clear. First, super-resolution helps recover fine details that standard detectors often miss, improving both recall and localization. Second, flexible backbones with strong multi-scale features—such as those using FPN, PANet, or Inception-style modules—are essential for handling objects of different sizes. Third, attention mechanisms, including SE, CBAM, and SEAM, help the network focus on important fine details, which is especially helpful in crowded UAV scenes. Fourth, new loss functions like NWD are important for stabilizing training and improving the localization of very small objects.

Despite this progress, important challenges remain. Many SR-based detection models are still too slow for real-time UAV deployment because they require heavy computation. Transformer-based backbones also need further optimization to meet strict power and latency limits. Attention modules, while beneficial for accuracy, increase model complexity and make deployment on embedded hardware more difficult. Finally, although NWD improves localization consistency, the best way to balance it with IoU-based losses is still not fully understood. These open issues must be addressed to achieve truly accurate, efficient, and real-time target recognition in next-generation UAV systems.

III. Methodology

The proposed YOLOv8-SR++ framework is designed to enhance target recognition in UAV imagery through the integration of super-resolution, multi-scale feature extraction, attention refinement, and improved localization loss. Instead of claiming novel standalone modules, the framework adapts established components within a unified detection pipeline tailored to aerial imagery, where small targets, dense scenes, and background clutter are common.

A. Dataset: VisDrone (UAV)

This work tested the performance of the proposed YOLOv8-SR++ (Fig. 2) on the VisDrone-DET benchmark dataset. The AISKYEYE team collected this challenging large-scale dataset for object detection tasks over complex aerial scenes. The dataset contains 10,209 high-quality static images captured from drones over 14 Chinese cities, covering a wide range of terrains: from crowded urban

streets and suburban landscapes to rural highways. The dataset contains 2.6 million annotated bounding boxes from 10 target categories, namely pedestrian, person, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motorbike, captured in different environmental conditions, including day and night lighting, occlusions, motion blur, and different weather. This heterogeneity makes VisDrone a suitable benchmark for testing the generalization and robustness of aerial detection models.

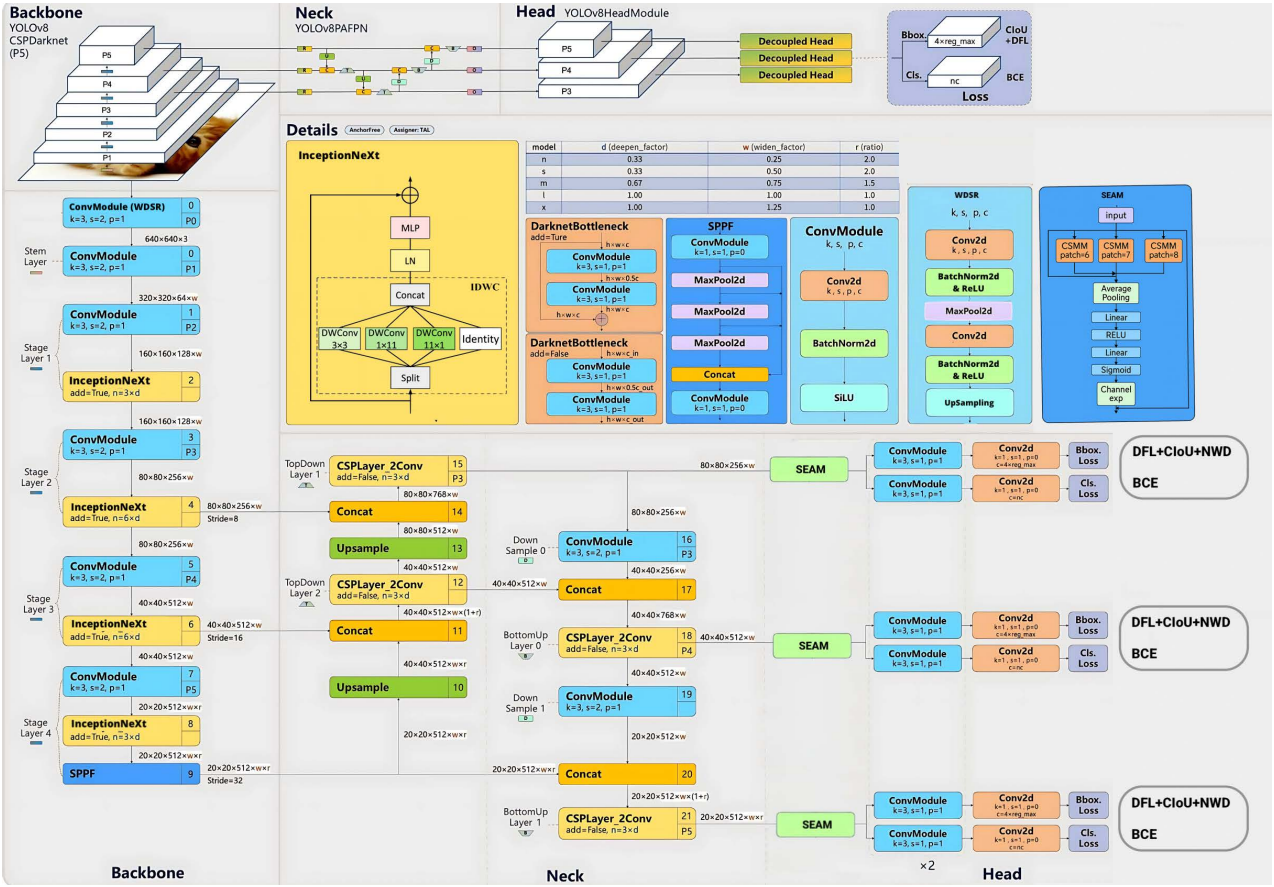


Fig. 2. Integration of inception-NeXt and SEAM with DLF in YOLOv8-SR.

The dataset follows the COCO evaluation protocol, and the main performance measure is Average Precision computed across IoU thresholds from 0.5 to 0.95. Secondary metrics include AP50, AP75, average recall, precision, recall, F1-Score, and FPS, which together provide a fair and reproducible evaluation. The official split contains 6471 training, 548 validation, and 1580 testing images.

B. Preprocessing and Data Pipeline

For robust detection and a fair evaluation pipeline, a consistent preprocessing strategy was applied to all models. The final detector input size was fixed at 640 × 640 pixels for both training and evaluation. Images were resized while preserving their original aspect ratio, and letterbox padding was used to minimize geometric distortion and retain contextual information. In the SR-enhanced pipeline, WDSR first reconstructed the image at an intermediate pre-detection resolution of 960 × 540, after which the image was converted to the same 640 × 640 detector input used by the baseline YOLOv8 model. This setting ensured that the comparison between SR-enhanced and non-SR models remained fair. In the VisDrone dataset, regions annotated as “ignored” were preserved during training so that predictions overlapping these areas were not unfairly penalized. In addition, label quality was carefully checked to remove duplicate annotations and invalid bounding boxes, and the dataset taxonomy strictly followed the official ten-category VisDrone protocol to maintain evaluation consistency.

To enhance generalization and mitigate class imbalance, several different augmentation strategies for objects were used. These comprised multi-scale training using scaling factors of between 0.7× and 1.3×, random cropping that ensured targets remained visible, and photometric transformations like HSV jitter. Spatial augmentations were also applied, such as horizontal flips, motion-blur simulations, and copy-paste methods that repeated targets into diverse environments to combat long-tail sparsity. Mosaic and mixup augmentations were employed sparingly, balancing between label accuracy and diversity of data so as not to corrupt the representation of objects.

In addition, the batch sampling method was formulated to prioritize images with high densities of targets, i.e., those whose sizes are less than 16 pixels. This bias

guaranteed that the network always received strong gradient signals from examples of objects without losing the general class distribution. Last but not least, Super-Resolution (SR)-considerate preprocessing was included by means of the Wide Activation Super-Resolution (WDSR) module. Low- and high-resolution pairs were created through bicubic downsampling at a factor of ×2, allowing the WDSR to pre-train as an effective reconstruction network. In coupled training, SR-enhanced outputs took the place of raw inputs, keeping the detector working all the time on sharpened imagery with added high-frequency details vital to distinguishing objects.

C. YOLOv8-SR Framework (Overall Architecture)

The proposed YOLOv8-SR framework extends the base YOLOv8 for target detection in aerial imagery. Three main enhancements form the core of architecture. First, Wide Activation Super-Resolution (WDSR) is added as a front-end module to recover fine visual details before detection. Second, InceptionNeXt blocks are inserted into selected backbone stages to enlarge the receptive field and improve multi-scale context extraction. Third, a Separate and Enhanced Attention Module (SEAM) is introduced in the feature-fusion neck to emphasize discriminative regions while suppressing background clutter. Apart from these targeted enhancements, the rest of the YOLOv8 design remains unchanged, including its anchor-free detection head with decoupled branches for classification, bounding-box regression, and objectness estimation, as well as the PANet-based neck for feature aggregation. This design preserves YOLOv8's real-time speed while improving detection precision in UAV scenes.

The training strategy was executed in two phases. First, the WDSR module was pre-trained for the super-resolution task while the YOLOv8 detector remained frozen. After the WDSR stage stabilized, end-to-end fine-tuning was performed so that the reconstructed outputs could directly support the detection pipeline. This progressive strategy allowed the SR module to learn robust reconstruction before being optimized jointly with detection.

D. Feature Extraction and Module Design

The YOLOv8-SR++ framework enhances target recognition with modules that strengthen the quality of the input, enrich the multi-scale feature representation, and refine semantic focus during fusion. It highlights three design emphases: visual detail enhanced by super-resolution, contextual learning expanded with a multi-path backbone, and discriminative feature emphasis reinforced through attention mechanisms. Together, these components constitute a coherent feature extraction pipeline optimized for the complexity of UAV imagery. Further, each module will be introduced in sequence.

E. Wide-Activation Super-Resolution (WDSR)

The Wide Activation Super-Resolution (WDSR) module was introduced to counter the challenge of insufficient discriminative features while detecting objects. Since tiny objects in aerial or UAV imagery often span only a few pixels, their recognition benefits significantly from the enhancement of high-frequency components such as edges and textures. To achieve this, each WDSR block is designed to widen feature channels before activation. The typical sequence involves a 1 × 1 convolution, nonlinear activation via the ReLU function, 3 × 3 convolution, and lastly a 1 × 1 projection with residual scaling. Of note, batch normalization is skipped to maintain spatial accuracy and weight normalization is used to stabilize the optimization. By expanding the activation width (6–9× channel expansion), WDSR achieves better accuracy compared to deeper but narrower residual structures. For integration, a compact WDSR-x2 network is used to upsample low-resolution input frames so they align with the detector's working resolution. During fine-tuning, the detection loss and the SR reconstruction loss are jointly optimized, with a smaller weight assigned to the latter, ensuring that the SR module strengthens feature representation without introducing artifacts [32].

$$I_{HR} = WDSR(I_{LR}) \tag{1}$$

where I_{LR} is the low-resolution image, and I_{HR} is the high-resolution image.

Each of these can be represented in Eqs. (2)–(7).

Conv2D:

$$[Y = W \times X + b] \tag{2}$$

where X represents the input, Y is the output, W is the filter, b is the bias term.

Batch_Norm_2D:

$$\mu_B = \frac{1}{m}\sum_{i=1}^{m} X_i \tag{3}$$

where μ_B is the mean, m is the batch size, and X_i the normalized input.

$$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \mu_B\right)^2 \tag{4}$$

where σ_B^2 is the variance of the batch.

$$\hat{x}_i = \frac{\left(X_i - \mu_B\right)}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{5}$$

where ϵ is a small constant to avoid division by zero in batch normalization.

ReLU:

$$f(X) = \max(0, X) \tag{6}$$

Maxpool2D:

$$Y = \max(X) \tag{7}$$

This corresponds to MaxPooling, which selects the maximum value within a feature window.

F. InceptionNeXt in the YOLOv8 Backbone

Traditional convolutional backbones often struggle to perceive the global context required for robust small-target recognition, since their limited receptive fields restrict feature diversity. In order to overcome such a limitation, the InceptionNeXt architecture provides an effective way of replacing large-kernel convolutions with parallel processing branches that factorize them. Every InceptionNeXt block combines four separate paths: (i) a standard square convolution kernel, (ii) a vertical band kernel, (iii) a horizontal band kernel, and (iv) an identity shortcut. The multi-branch design improves directional feature extraction while keeping computation efficient, hence enhancing the receptive field coverage without imposing excessive memory or processing overhead. For incorporation into YOLOv8, particular C2f stages in the backbone were substituted with InceptionNeXt blocks. This was a change that enabled the model to pick up more contextual hints, and these are especially important for their ability to detect objects embedded within complex backgrounds. Mathematically, the output of an InceptionNeXt (Eq. (8)) block, denoted as O_{IN}, can be expressed as:

$$O_{IN} = \sum_{i=1}^{n} W_i \times F_i \tag{8}$$

where W_i are the weights associated with different convolutional filters, and F_i represents the features extracted at various scales.

Separate and Enhanced Attention Module (SEAM) in the Neck Feature fusion in the PAN-FPN neck tends to dilute discriminative cues, which is especially detrimental for object detection. To overcome this drawback, we added a SEAM to the design. The SEAM approach is derived from a dual-pathway concept that processes features through channel-attention and spatial-attention mechanisms. The channel-attention pathway focuses on the most informative feature channels, while the spatial-attention pathway identifies important regions in the feature maps. Their outputs are then boosted and combined to re-weight the feature representations, allowing the network to concentrate more effectively on fine-grained object details while suppressing redundant and noisy background information. This selective refinement is particularly valuable in UAV imagery, where occlusion, clutter, and high object density are typical issues. SEAM modules were positioned after every fusion node in the neck (e.g., pyramid levels P3–P5). Positioning SEAM at these critical junctures ensures that multi-scale features are refined before progressing to the next layers, thus enhancing localization accuracy and the detection of densely packed targets.

G. Bounding-Box Regression Loss: CIoU + NWD Hybrid

Bounding-box regression for objects is often unstable under IoU-based metrics, since even a shift of a few pixels can significantly reduce the IoU for tiny bounding boxes. To overcome this challenge, we integrate Complete IoU (CIoU) with NWD into the detection loss. CIoU enhances traditional IoU by incorporating three factors: overlap, center distance, and aspect ratio alignment. Its formulation is given as Eq. (9) [29, 33]:

$$LCIoU = 1 - IoU(b, b^*) + \frac{\rho^2(c, c^*)}{d^2} + \alpha v. \tag{9}$$

where ρ represents the Euclidean distance between the predicted and ground-truth box centers, d is the diagonal length of the minimum enclosing box, v measures the aspect-ratio mismatch, and α balances the penalty term.

On the other hand, NWD as Eq. (10) treats each bounding box as a two-dimensional Gaussian distribution parameterized by its center and dimensions. The distance between two boxes is then measured using the 2-Wasserstein metric, normalized by a scale factor τ.

$$NWD(b, b^*) = \frac{W_2(N_b, N_{b^*})}{\tau} \tag{10}$$

The corresponding regression loss is expressed as Eq. (11):

$$L_{NWD} = 1 - e^{-NWD} \tag{11}$$

This formulation is less sensitive to small localization shifts, making it particularly suitable for object regression. Finally, the hybrid detection loss integrates both components along with classification, objectness, and super-resolution terms as Eq. (12):

$$L = \lambda_{box}\left(\beta L_{CIoU} + (1-\beta) L_{NWD}\right) + \lambda_{cls} L_{cls} + \lambda_{obj} L_{obj} + \lambda_{SR} L_{SR} \tag{12}$$

This combined formulation ensures robust localization for objects while balancing detection and reconstruction objectives.

H. Training Protocol

Training was conducted under a carefully designed protocol aimed at balancing detection performance with computational efficiency. AdamW served as the primary optimizer, while all ablation variants were trained under the same schedule to ensure reproducibility [34]. The initial learning rate was 0.001 and was gradually reduced using a cosine decay schedule across 300 epochs. A weight decay of 0.0005 was applied to reduce overfitting, and a short warm-up stage was used to stabilize early training. Mixed-precision training (FP16/AMP) was employed to reduce memory usage and accelerate computation. To

evaluate the impact of the proposed enhancements in a controlled manner, ablation studies were conducted progressively: (i) baseline YOLOv8, (ii) YOLOv8 with WDSR, (iii) addition of InceptionNeXt blocks, (iv) incorporation of SEAM, and (v) introduction of the hybrid CIoU + NWD loss. For each configuration, we report standard detection metrics together with model size, FLOPs, and inference FPS.

I. Implementation

All experiments were conducted in PyTorch using Ultralytics YOLOv8 as the base detector. Selected C2f modules in the backbone were replaced with InceptionNeXt blocks, SEAM was added after feature fusion in the neck, and a lightweight WDSR module was placed before the detector to enhance input detail. Pre-trained COCO weights were used to initialize the YOLOv8 backbone. The WDSR module was trained through a reconstruction warm-up stage before joint fine-tuning with the detector. All comparisons were carried out at the same final detector input resolution so that the extra computational cost introduced by the proposed modules could be measured fairly.

J. Justification of Design Choices for Target Detection

The integration of WDSR, InceptionNeXt, SEAM, and the CIoU+NWD hybrid loss was designed specifically to address the challenges of UAV target detection. WDSR restores high-frequency details in low-resolution inputs, which is especially important for separating small objects from complex aerial backgrounds. InceptionNeXt broadens the receptive field and strengthens multi-scale context modeling without imposing high computational cost. SEAM refines both spatial and channel attention so that the detector focuses more effectively on informative regions while suppressing irrelevant clutter. Finally, the CIoU+NWD hybrid loss stabilizes bounding-box regression by combining geometric alignment with a distribution-aware similarity term, which is particularly useful for small and densely packed objects.

IV. Experiments and Results

We evaluated YOLOv8-SR++ with a focused set of tests to measure accuracy, speed, and robustness for object detection in UAV imagery. This section describes the experimental setup and metrics, compares against strong baselines, reports ablation results, and offers qualitative examples that show what each module adds.

A. Experimental Setup

Our experimental assessment was performed on the VisDrone 2019 dataset, a popular and challenging benchmark for aerial image processing. The dataset includes over 10,000 drone-shot images captured in various urban and suburban settings, with annotations for 10 object classes. It has additional challenges posed by diverse lighting, occlusions, and complex backgrounds, making it an excellent testbed for object detection research [35].

Our model was implemented in PyTorch and trained on an NVIDIA RTX 3090 GPU with 24 GB of VRAM. AdamW was used as the optimizer because of its stable convergence and effective weight-decay regularization. In the SR-enhanced setting, low-resolution inputs were first reconstructed by the WDSR module to recover fine details before being passed to the detector.

The main hyperparameters were selected to balance performance and efficiency. Training was carried out with a batch size of 16 for 300 epochs, using an initial learning rate of 0.001 with cosine annealing. In the SR branch, images were first reconstructed at 960 × 540 and then converted through standard letterbox preprocessing to the final detector input size of 640 × 640. A warm-up stage of 3 epochs was used to stabilize early training, and a weight decay of 0.0005 was applied to reduce overfitting. The AdamW optimizer used $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

For a fair and consistent evaluation, we followed the standard training and validation splits of the VisDrone benchmark. All baseline models and all proposed variants were trained under the same conditions, with the same final detector input size, optimization schedule, and evaluation protocol.

B. Evaluation Metrics

The performance of YOLOv8-SR++ is evaluated using a set of complementary metrics that reflect detection quality, localization quality, and real-time efficiency in complex UAV environments. The primary metrics are mean Average Precision under the COCO protocol, including mAP@0.5 and mAP@0.5:0.95. In addition, we report precision, recall, F1-Score, FPS, parameter count, and FLOPs. These measures together provide a balanced view of detector robustness, localization behavior, and deployment efficiency.

TABLE I. Performance Metrics

Metric Equation	Equation No.
$Precision = \frac{TP}{TP + FP}$	(13)
$Recall = \frac{TP}{TP + FN}$	(14)
$F1 - Score = 2 \times \left(\frac{P \times R}{P + R} \right)$	(15)
$AP = \sum_{i=0}^{n-1} (P(i) \times (R(i) - R(i-1)))$	(16)
$mAP = \left(\frac{\sum_{i=1}^{k} AP_i}{k} \right)$	(17)

Legends: *TP*: True Positive, *TN*: True Negative, *FP*: False Positive, *FN*: False Negative, *P*: Precision, *R*: Recall.

Precision refers to the proportion of correctly predicted positive detections, defined in Eq. (13), while recall reflects the portion of all ground-truth instances retrieved successfully, as given in Eq. (14). To have a well-balanced indication of recognition performance, the harmonic mean of precision and recall, expressed by the F1-Score, is

calculated from Eq. (15). As the aerial detection tasks face large background areas and sparsely distributed targets, it is normally impractical to compute true negatives directly; hence, the overall detection accuracy is estimated through an established approximation based on the relationship between precision, recall, and F1-Score, represented in Eqs. (16)–(17). To make sure the proposed model can meet the real-time requirements during UAV deployment, inference speed is measured in terms of frames per second, or FPS, whose reciprocal value corresponds to latency. Finally, the computational complexity is reported using the number of trainable parameters and FLOPs, which provide insight into the model's deployment capability on resource-constrained UAV platforms Table I.

The consolidated performance metrics for YOLOv8-SR++ and its variants are presented in Table II, showing consistent improvements in detection quality as the proposed modules are progressively integrated.

TABLE II. OVERALL ACCURACY, PRECISION, RECALL, AND MAP FOR YOLOV8-SR++ AND VARIANTS

Metric	Base	+WDSR	+IncNx	+SEAM	SR++
Accuracy (%)	74.6	77.9	79.5	80.8	82.1
Precision (%)	67.2	70.4	72.5	74.0	76.2
Recall (%)	62.5	64.1	66.0	67.8	69.4
F1-Score (%)	64.8	67.1	69.0	70.8	72.6
mAP@0.5 (%)	43.0	45.3	47.2	48.6	49.6
mAP@0.5:0.95 (%)	26.7	28.6	30.4	31.4	32.1
FPS	101	62	48	45	43

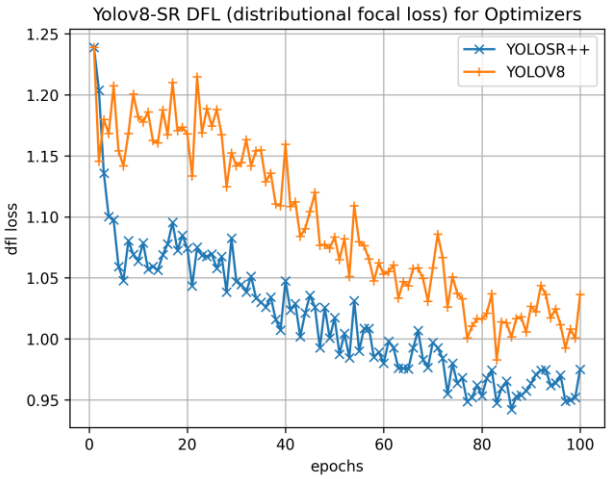
Base = YOLOv8-s (Baseline); +WDSR = YOLOv8-s + WDSR; +IncNx = YOLOv8-s + WDSR + InceptionNeXt; +SEAM = YOLOv8-s + WDSR; + InceptionNeXt+ SEAM; SR++ = (YOLOv8-s + WDSR + InceptionNeXt+ SEAM+ CIoU + NWD).

C. *Ablation Studies*

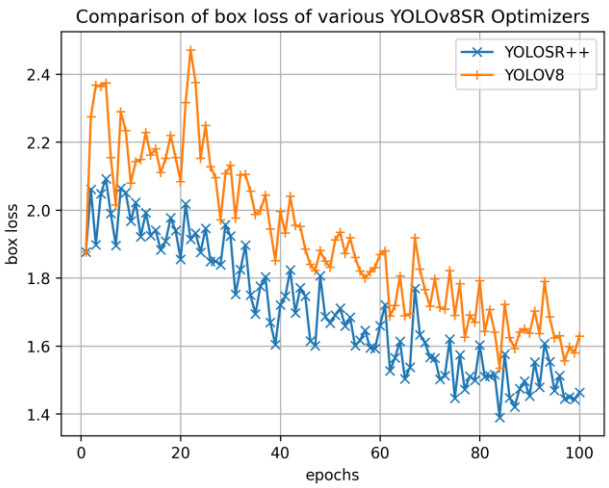
Ablation experiments were conducted to quantify the contribution of each module in YOLOv8-SR++. All experiments were performed under identical training conditions and evaluated on the VisDrone benchmark to ensure a controlled comparison. The results, summarized in Table II, highlight the progressive impact of WDSR, InceptionNeXt, SEAM, and the CIoU + NWD hybrid loss.

First, the introduction of the WDSR super-resolution module led to a substantial improvement by enhancing low-quality UAV images through the recovery of high-frequency details and clearer structural patterns, which allowed the backbone to extract more discriminative representations. This resulted in marked improvements in both accuracy and precision. Incorporating InceptionNeXt into the backbone further improved recognition robustness and contextual understanding through multi-branch convolutional diversity and expanded receptive fields. Finally, SEAM was added at the neck stage, where it improved feature-map importance by applying spatial and channel-wise attention simultaneously. This reduced background noise common in aerial scenes and enabled the detector to focus more effectively on the most relevant target regions.

Finally, the adoption of the hybrid CIoU + NWD loss provided a substantial enhancement in localization stability. While the CIoU component ensured geometric alignment in bounding-box prediction, the NWD term delivered smooth gradient behavior for distributions with minimal overlap, an essential requirement in dense scenes with irregular perspectives. Together, these components yielded a cumulative improvement of 7.5% in overall accuracy, 9% in precision, and 6.6% in mAP@0.5 compared with the baseline YOLOv8 detector while maintaining real-time performance at 43 FPS. These results confirm that each architectural element contributes to the high-performance UAV-oriented target recognition skill.



(a)



(b)

Fig. 3. Comparison of box loss and dfl loss of various YOLOv8SR++ optimizers: (a) dfl loss; (b) box loss.

This progression clearly demonstrates that each module makes complementary contributions. WDSR enhances input fidelity, InceptionNeXt broadens context awareness,

SEAM sharpens attention to critical regions, and CIoU+NWD stabilizes bounding-box regression (Fig. 3) for objects [36]. While FPS declines as modules are added, the final system achieves a well-balanced trade-off between accuracy and efficiency, confirming its suitability for UAV-based object detection.

D. Component-Wise Analysis

A closer look at each component reveals how they individually strengthen the detection pipeline. The WDSR module functions as the foundation by improving the visual quality of UAV inputs. By enhancing edges and restoring fine-grained structures, it substantially benefits the detection of very distant targets such as distant vehicles and pedestrians. The InceptionNeXt backbone expands the network's receptive field and captures richer multi-scale features, which improve recognition across varying distances and object sizes. Its inclusion, when stacked on top of WDSR, delivered nearly a two percentage point boost in mAP@0.5, demonstrating its effectiveness for contextual feature aggregation. The SEAM module plays a critical role in complex UAV environments. By applying both channel-wise and spatial attention, SEAM enables the network to emphasize target regions while suppressing irrelevant or noisy background signals [27]. This capability is especially useful in cluttered urban scenes where occlusion and background patterns often confuse detectors. Finally, the CIoU + NWD hybrid loss enhances the stability of bounding-box regression. While CIoU ensures alignment in overlap, distance, and aspect ratio, NWD provides robustness for very small bounding boxes by reducing sensitivity to pixel-level shifts and improving convergence under noise or occlusion. Together, these modules address complementary weaknesses of YOLOv8, and their integration produces a synergistic effect that elevates overall performance in UAV-based object detection [37].

E. Performance-Efficiency Trade-Off

We analyzed the trade-off between detection performance and real-time efficiency by jointly evaluating accuracy-related metrics and inference speed. According to Table II, each architectural enhancement improves detection performance while introducing a measured computational cost. The baseline YOLOv8 achieved the highest throughput at 101 FPS but the lowest detection performance. Integrating WDSR improved visual quality and recognition performance, although FPS decreased to 62 because of the added super-resolution computation (Fig. 4). The inclusion of InceptionNeXt further strengthened multi-scale feature extraction, and SEAM delivered additional gains in precision and localization stability without a severe speed penalty.

Overall, the complete YOLOv8-SR++ model performed best, achieving 82.1% accuracy, 76.2% precision, and real-time inference at 43 FPS. The results show that each added module gradually reduces inference speed while steadily improving detection quality. This balance makes the proposed architecture suitable for UAV target-recognition tasks that require both accuracy and real-time processing.

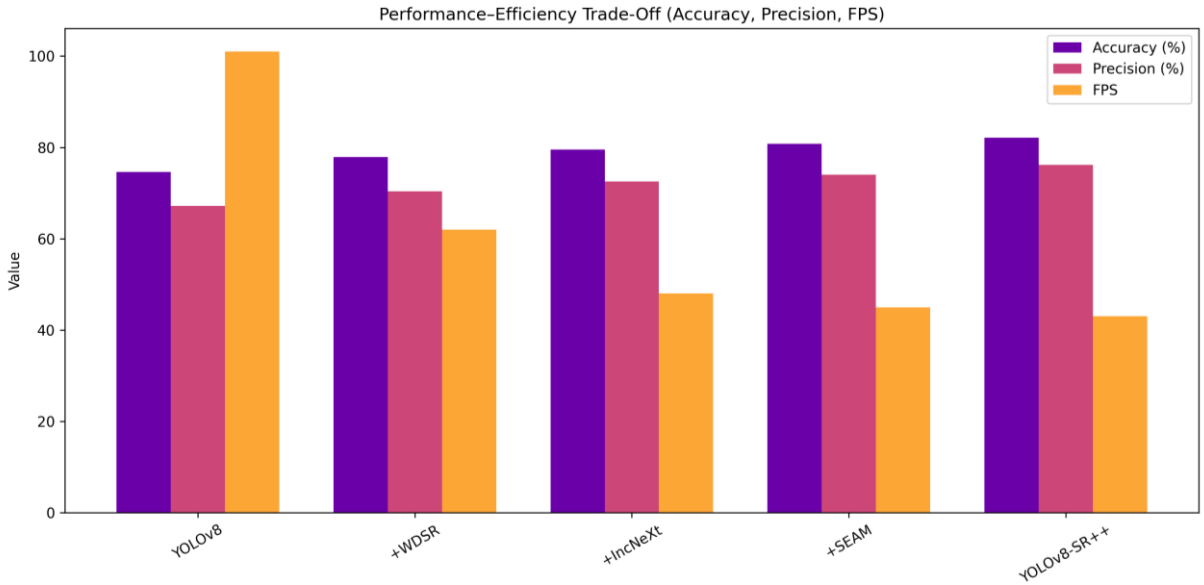


Fig. 4. Comparison of accuracy and speed between YOLOv8 and YOLOv8-SR++.

F. Qualitative Results

To better illustrate the strengths of YOLOv8-SR++, we visualized detection outcomes across several challenging scenarios in UAV imagery. In scenes with occluded pedestrians, baseline YOLOv8s often fail to recognize individuals partially hidden by vehicles or other objects. In contrast, YOLOv8-SR++ successfully identifies these partially visible pedestrians, owing to the improved spatial resolution provided by WDSR and the refined focus offered by SEAM [38]. Similarly, in high-density urban environments where vehicles and bikes are packed closely together under difficult lighting conditions such as shadows or strong illumination changes, standard detectors typically struggle with poor contrast and overlapping objects. YOLOv8-SR++, however,

demonstrates greater robustness, reliably detecting closely spaced targets. In long-range drone shots, distant vehicles and pedestrians that appear indistinct in the baseline model become more distinguishable after WDSR enhances input resolution and SEAM selectively emphasizes informative features.

These qualitative examples highlight the practical advantages of the proposed framework. By combining super-resolution, enhanced multi-scale representation, and attention-driven refinement, YOLOv8-SR++ consistently outperforms the baseline in complex real-world aerial scenes where occluded, or distant targets are otherwise missed.

V. Discussion

The experimental results show that the proposed YOLOv8-SR++ framework provides a meaningful improvement in UAV-based target recognition while preserving practical inference speed. The framework combines four complementary enhancements—WDSR, InceptionNeXt, SEAM, and the CIoU+NWD hybrid loss—to address several limitations of the baseline YOLOv8 model. Quantitative results indicate that each module contributes to stronger detection performance, and their combined effect leads to clear gains in accuracy, precision, and localization consistency. These benefits are especially visible in complex aerial environments characterized by occlusion, illumination variation, background clutter, and dense target distributions [39, 40].

The introduction of the WDSR module is especially beneficial in cases where aerial images suffer from resolution degradation due to high-altitude capture or camera motion. By reconstructing sharper textures and strengthening edge continuity, WDSR provides the backbone with richer visual cues, thus elevating the quality of feature extraction. This is particularly beneficial in UAV imagery, where targets frequently appear with limited visual detail. InceptionNeXt further amplifies this performance by enabling multi-path convolutional processing, improving context modeling across varying orientations and spatial patterns. The resulting receptive-field diversity ensures that target structures are captured more reliably even when scene dynamics introduce unusual geometric configurations.

SEAM further sharpens the representational focus of the network by separating and enhancing spatial and channel attention, which improves emphasis on informative regions while suppressing irrelevant background noise. This is particularly important in aerial scenes, where rooftops, roads, trees, and shadows can mislead conventional detectors. The hybrid CIoU+NWD loss complements these improvements by reinforcing the stability of bounding-box regression through a combination of geometric alignment and distribution-aware similarity. Although each component introduces some additional computation, the resulting model still operates in real time. The reduction in FPS from 101 in YOLOv8 to 43 in YOLOv8-SR++ is offset by clear gains in detection quality. The performance–efficiency comparison therefore shows that the model still meets the speed requirements of many UAV applications.

The trend of Table II is clear: accuracy and precision are increasing smoothly with each architectural improvement, while FPS is decreasing moderately. The final YOLOv8-SR++ model will exhibit an optimal tradeoff: high recognition performance coupled with real-time inference capability. For scenarios in UAV applications where decisions need to be made very fast, like live traffic analytics or emergency monitoring, or even autonomous navigation, this is the required balance.

Despite this, several considerations remain for future research. Although advantageous, the reliance on super-resolution introduces additional computational overhead that could hinder deployment on micro-UAVs or battery-constrained platforms. Lightweight approximations of WDSR or adaptive SR methods can be explored to alleviate this challenge. In a similar vein, although InceptionNeXt and SEAM provide strong feature enhancement, their intrinsic complexity suggests the need for further optimization of inference speed through model compression methods such as pruning, quantization, or knowledge distillation. Other promising directions include exploring attention mechanisms with a lower computational footprint or integrating transformer-like global reasoning modules in hybrid form optimized for UAV hardware.

There is also room for refinement on the combination of the CIoU+NWD loss. Though NWD largely stabilizes regressions, its interaction with CIoU when there are extreme occlusions or non-overlapping bounding boxes needs a deeper investigation. Moreover, an advanced assignment strategy like Wasserstein-guided matching or task-aligned label assignment may further strengthen the convergence.

In conclusion, the proposed YOLOv8-SR++ framework has immense potential as a practical and real-time solution for target recognition in UAV-based applications. It realizes higher accuracy and precision with real-time capabilities through the synergistic integration of super-resolution, backbone enhancement, attention refinement, and hybrid regression loss. These findings indicate its applicability to mission-critical aerial tasks, such as surveillance, search-and-rescue analytics, agricultural inspection, urban monitoring, and autonomous UAV navigation. Further refinement of efficiency, lightweight design, and adaptive learning mechanisms will enhance the model's suitability for a wide range of real-world deployments.

VI. Conclusion

This research introduced YOLOv8-SR++, an advanced deep learning framework for improving object detection in UAV imagery through the coordinated integration of super-resolution, multi-scale representation learning, attention refinement, and hybrid loss optimization. The model addresses one of the most persistent challenges in aerial perception: reliable detection of tiny, occluded, or distant objects under varying lighting conditions and cluttered backgrounds.

The framework combines four complementary components that improve detection performance from different angles. First, the Wide Activation Super-Resolution (WDSR) module restores high-frequency details in low-resolution UAV images, improving the visual quality of the detector input and helping the network capture small and distant objects. Second, InceptionNeXt blocks in the backbone strengthen multi-scale contextual representation while preserving computational efficiency. Third, the Separate and Enhanced Attention Module (SEAM) in the neck refines both channel and spatial feature importance, allowing the detector to focus more effectively on salient object regions while suppressing background noise. Finally, a hybrid CIoU+NWD loss is used to improve bounding-box regression stability, especially for crowded targets commonly found in aerial scenes.

Extensive experiments on the VisDrone dataset show that YOLOv8-SR++ achieves a mean average precision of 49.6% at mAP@0.5, outperforming baseline YOLOv8s (43.0%) and YOLOv8-SR (45.3%) while still maintaining real-time inference at 43 FPS. These findings confirm that the integrated modules contribute positively to both detection accuracy and robustness, making YOLOv8-SR++ an efficient and effective detector for UAV-based vision tasks.

Ablation experiments confirmed the contribution of each enhancement: WDSR improved input fidelity, InceptionNeXt broadened feature diversity, SEAM strengthened discriminative attention, and the hybrid loss reduced localization instability. Qualitative evaluation further showed that YOLOv8-SR++ can successfully detect partially occluded and heavily overlapping objects that are often missed by standard YOLO models.

From a practical perspective, YOLOv8-SR++ is well suited for UAV applications such as disaster response, border patrol, traffic monitoring, and crop inspection, where real-time object detection is essential. At the same time, the framework still faces some limitations, particularly increased computational cost, dependence on the training dataset, and longer training time, all of which may affect deployment on lightweight or edge-based UAV systems.

Future work will explore lightweight SR architectures such as FSRCNN and Fast-SRGAN, transformer-based attention mechanisms such as MobileViT, and knowledge distillation to compress the model for low-power platforms. In addition, cross-dataset fine-tuning on UAVDT, DOTA, and xView could further evaluate its generalization ability.

Overall, YOLOv8-SR++ demonstrates that combining image enhancement, feature extraction, attention refinement, and improved localization within a unified framework can deliver stronger UAV object-detection performance without sacrificing real-time feasibility. The proposed framework provides a solid foundation for future research in UAV-based computer vision.

CODE AND DATA AVAILABILITY

The VisDrone dataset used in this study is publicly available. The implementation details, training settings, and ablation protocol are described in the manuscript to support reproducibility. The source code and trained weights can be shared by the corresponding author upon reasonable request.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Gangeshwar Mishra conceptualized and designed the study, conducted data collection, developed the framework and participated in data analysis and interpretation. Rohit Tanwar contributed to the development of the framework, supervised the implementation of the intervention, and assisted in manuscript writing and revisions. Prinima Gupta supported data analysis and interpretation and provided critical feedback on the manuscript. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work to ensure its integrity and accuracy.

ACKNOWLEDGMENT

We acknowledge the technical and library staff of Manav Rachna University and Shri Mata Vaishno Devi University for providing timely support and resources whenever required.

REFERENCES

[1] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv Preprint, arXiv:1804.02767, 2018. https://arxiv.org/abs/1804.02767

[2] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv Preprint, arXiv:2004.10934, 2020. https://arxiv.org/abs/2004.10934

[3] J. Yu *et al*., "Wide activation for efficient and accurate image super-resolution (WDSR)," in *Proc. the ECCV Workshops (ECCVW)*, 2018.

[4] W. Yu, P. Zhou, S. Yan, and X. Wang, "InceptionNeXT: When inception meets ConvNeXt," arXiv Preprint, arXiv:2303.16900, 2023. doi: 10.48550/arXiv.2303.16900

[5] VisDrone Team. (2023). VisDrone-Dataset: The dataset for drone-based detection and tracking. [Online]. Available: https://github.com/VisDrone/VisDrone-Dataset

[6] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv Preprint, arXiv:1804.02767, 2018.

[7] H. Xu, W. Zheng, F. Liu, P. Li, and R. Wang, "Unmanned aerial vehicle perspective small target recognition algorithm based on improved YOLOv5," *Remote Sens*. (Basel)., vol. 15, 2023. doi: 10.3390/rs15143583

[8] G. Jocher and Ultralytics, *YOLOv5 by Ultralytics*, 2020.

[9] Ultralytics, *YOLOv8: Next-Generation YOLO Object Detector*, 2023.

[10] H. Zhang, C. Wu, and Z. Shen, "YOLO-SR: Object detection with super-resolution," *IEEE Access*, vol. 8, pp. 1–12, 2020.

[11] D. Gupta and S. Rathore, "An improved YOLOv8 for multiclass object detection in aerial imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, 5500210, 2023. doi: 10.1109/TGRS.2023.5500210

[12] C. Dong, *et al.*, "PG-YOLO: A novel lightweight object detection method for edge devices in industrial internet of things," *IEEE*

Internet Things J., vol. 10, no. 3, pp. 2339–2350, 2022. doi: 10.1109/JIOT.2022.3217070

[13] G. S. Xia *et al*., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3974–3983.

[14] C. Chen *et al*., "YOLO-based UAV technology: A review of the research and its applications," *Drones*, vol. 7, no. 3, 190, 2023. doi: 10.3390/drones7030190

[15] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 38, no. 2, pp. 295–307, 2016.

[16] C. Ledig *et al*., "Photo-realistic single image super-resolution using a GAN," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] J. Liang *et al*., "SwinIR: Image restoration using swin transformer," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 1833–1844.

[18] Y. Zhang *et al*., "Residual dense network for image super-resolution," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2472–2481.

[19] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF ICCV Workshops*, 2021, pp. 1905–1914. doi: 10.1109/ICCVW54120.2021.00217

[20] J. Zhang, J. Lei, W. Xie, Z. Fang, Y. Li, and Q. Du, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 1–15, 2023. doi: 10.1109/TGRS.2023.3258666

[21] M. Xu, Z. Wang, J. Zhu, X. Jia, and S. Jia, "Multi-attention generative adversarial network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 1–15, 2022. doi: 10.1109/TGRS.2022.3180068

[22] M. Tan, R. Pang, and Q. V Le, "EfficientDet: Scalable and efficient object detection," in Proc. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] A. Dosovitskiy *et al*., "An Image is Worth 16×16 Words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[24] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[25] A. Vaswani *et al*., "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] W. Zhu and K. Chen, "Real-time object detection for unmanned aerial vehicles based on vision transformer and edge computing," *Sci. Rep.*, vol. 16, 2026. doi: 10.1038/s41598-026-37938-5

[27] J. Liang *et al.*, "Aerial small target detection algorithm based on cross-scale separated attention," *Plos One*, vol. 20, no. 11, e0337318, 2025. doi: 10.1371/journal.pone.0337318

[28] H. Rezatofighi *et al*., "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.

[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[30] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.

[31] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized gaussian wasserstein distance for tiny object detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 190, pp. 87–113, 2022. doi: 10.1016/j.isprsjprs.2022.06.002

[32] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. CVPR Workshops (CVPRW)*, 2017. https://arxiv.org/abs/1707.02921

[33] Z. Zheng *et al*., "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020, vol. 34, no. 07, pp. 12993–13000.

[34] P. Zhou, X. Xie, Z. Lin, and S. Yan, "Towards understanding convergence and generalization of AdamW," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 46, no. 6, pp. 6486–6493, 2024. doi: 10.1109/TPAMI.2024.3382294

[35] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1921–1930. doi: 10.1109/ICCV.2019.00201

[36] J. Yu, Y. Fan, J. Yang, N. Xu, X. Wang, and T. S. Huang, "Wide activation for efficient and accurate image super-resolution," arXiv Preprint, arXiv:1808.08718, 2018.

[37] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord, "xView: Objects in Context in Overhead Imagery," arXiv Preprint, arXiv:1802.07856, 2018.

[38] H. Xu, W. Zheng, F. Liu, P. Li, and R. Wang, "Unmanned aerial vehicle perspective small target recognition algorithm based on improved YOLOv5," *Remote Sens. (Basel)*., vol. 15, no. 14, 3583, 2023. doi: 10.3390/rs15143583

[39] L. Wang and H. Li, "Small object classification using YOLOv8-SR in complex urban landscapes," *J. Vis. Commun. Image Represent.*, vol. 89, 103759, 2023. doi: 10.1016/j.jvcir.2023.103759

[40] M. Wang, Y. Liu, and Z. Zhang, "YOLOv8: Next generation of YOLO object detection," arXiv Preprint, arXiv:2302.04356, 2023.