LCEA-YOLO: Improving Small Object Detection in Aerial Imagery Using Local Contrast Enhancement Attention and Inner-CIoU Loss

Sara Ennaama
SIGL LAB, ENSA of Tetouan, Abdelmalek Essaadi University Tetouan, Morocco
sara.ennaama@etu.uae.ac.ma (corresponding author)

Hassan Silkan
Department of Computer Science, Laboratory LAROSERI, Faculty of Sciences, Chouaib Doukkali University, El Jadida, Morocco
silkan.h@ucd.ac.ma

Abderrahim Tahiri
SIGL LAB, ENSA of Tetouan, Abdelmalek Essaadi University Tetouan, Morocco
t.abderrahim@uae.ac.ma

Faouzia Ennaama
LAMIGEP, Moroccan School of Engineering Sciences, Marrakech, Morocco
F.ennaama@emsi.ma

Received: 15 February 2026 | Revised: 5 March 2026 | Accepted: 20 March 2026

 DOI: https://doi.org/10.48084/etasr.18190

ABSTRACT

This study proposed the Local Contrast Enhancement Attention - You Only Look Once (LCEA-YOLO) model to address the challenge of detecting small objects in high-altitude Unmanned Aerial Vehicle (UAV) images. LCEA-YOLO is a customized detector developed using the YOLOv10n architecture that embeds an LCEA module to emphasize local subtle contrast variations, as well as an Inner-CIoU loss function that enhances the regression ability for smaller objects. Experiments on the UAVDT dataset were conducted, with the results demonstrating that LCEA-YOLO exceeded other baseline algorithms with an overall mAP@0.5 of 47.2% and mAP@0.5:0.95 of 29.9%. Additionally, the model presented remarkable resilience on challenging under-represented categories, yielding improvements of 15.0% and 7.6% in detection accuracy for trucks and buses, respectively. These findings validated the benefit of targeted local contrast enhancement and scale-aware regression for real-time UAV detection.

Keywords-object detection; Unmanned Aerial Vehicles (UAV); YOLOv10; attention mechanism; Inner-CIoU loss; small object detection; deep learning

I. INTRODUCTION

The growing deployment of Unmanned Aerial Vehicles (UAVs) has enabled various applications, such as urban traffic surveillance, agriculture inspection, and public safety services. A requirement of these applications is the automated identification of objects of interest from aerial images. However, their detection from UAV platforms presents significant challenges. Specifically, due to the high-altitude perspectives, objects like vehicles or pedestrians may occupy few pixels and exhibit low contrast against complex backgrounds, impairing the discriminability of features as well as affecting the stable detection performance in practical UAV applications [1, 2].

To comply with the processing speed requirements of drone systems, one-stage detector networks like those models from the You Only Look Once (YOLO) family have dominated the field due to their competitive performance-complexity trade-off [3]. Iterations of the YOLO framework have further refined this efficiency through structural innovations. For example, an anchor-free detection mechanism was adopted in YOLOv8,

while the feature aggregation was improved by the C2f module [4]. In [5], YOLOv10 introduced an end-to-end detection formula that removes the necessity for the conventional filtering phase referred to as Non-Maximum Suppression (NMS). This was achieved by implementing an integrated dual label assignment method, which resulted in streamlining the inference pipeline and improving deployment speed. Authors in [6] indicated that YOLOv11, combined with lightweight embedded systems, achieved a high-performance real-time ground vehicle detection for Advanced Driver Assistance Systems (ADAS).

While Convolutional Neural Network (CNN)-based architectures have long defined the state-of-the-art, Transformer-based detectors have emerged as competitive alternatives. Authors in [7] proposed the Real-Time DEtection TRansformer (RT-DETR), the first real-time end-to-end Detection Transformer to match or surpass YOLO models in both speed and accuracy on standard benchmarks. This paradigm has been extended to aerial scenarios, where the RT-DETR framework was adapted for UAV small object detection [8]. Meanwhile, YOLO-based methods continue to evolve with targeted modifications for aerial imagery, such as SRTSOD-YOLO [9], which enhanced YOLO11 with multi-scale feature fusion for UAV targets. However, none of these approaches explicitly leverage local contrast cues to improve the separability of small targets from cluttered backgrounds, which is the specific focus of this paper.

Despite these developments, identifying small and ambiguous targets within overhead photographs is an ongoing research problem. There are two main reasons for the unsatisfactory performance in small object detection of aerial images by previous YOLO-based detectors. The first challenge relates to the fact that mainstream backbone and neck architectures are primarily tailored to generic object detection tasks without an explicit mechanism for reinforcing fine-grained local cues. In particular, the global attention mechanisms can be designed targeting long-range dependencies, but are not tailored to represent subtle local contrast changes that are essential in separating tiny targets from intricate backgrounds. Second, frequently used bounding box regression losses treat objects of all sizes equally and may suffer from unstable optimization for very small objects, where small perturbations cause large changes in overlap measures.

To solve these issues, several studies have focused on two complementary directions: features of networks and loss functions. Attention modules, such as channel-wise recalibration using Squeeze-and-Excitation (SE) mechanisms and combined spatial and channel refinement using the Convolutional Block Attention Module (CBAM), have been effective in enhancing the discriminative ability of features by accentuating relevant regions [10, 11]. In [12], the incorporation of distinct visual characteristics based on specific visual features, such as local contrast, is important for enhancing small object detection in cluttered images. Simultaneously, the regression process has been refined through several IoU variants that surround geometric penalties, namely GIoU, DIoU, and CIoU, leading to better localization of predicted boxes [13, 14]. Building on these ideas, the Inner-IoU concept was introduced to increase regression sensitivity for small bounding boxes by focusing on the inner geometry of predicted and ground-truth boxes [15].

Despite these developments, there is still a gap in explicitly leveraging local contrast cues within lightweight YOLO architectures for aerial small object detection, while simultaneously refining the regression loss to be more sensitive to small bounding boxes. To address this, the present paper introduces LCEA-YOLO, a modified YOLOv10-nano (YOLOv10n) architecture in order to optimize small object detection in aerial images. This method integrates a Local Contrast Enhancement Attention (LCEA) module into the detection head to improve discriminative local representations, while employing an Inner-CIoU loss to stabilize bounding box optimization for diminutive instances.

II. PROPOSED METHOD

This study introduces LCEA-YOLO, a specialized variant of YOLOv10n designed for aerial images, where objects are small and visually uncertain due to the altitude, clutter, and limited pixel resolution [1, 2]. The direct and real-time recognition mode in YOLOv10 are preserved in this version [5] based on two targeted modifications: (i) a LCEA module is plugged into the detection head to enrich multi-scale predictions; and (ii) an Inner-CIoU regression loss is employed to increase the sensitivity toward localization errors for small boxes beyond IoU-based losses [13, 14], as well as the inner concept proposed in Inner-IoU [15].

A. YOLOv10n Baseline Architecture

LCEA-YOLO is developed based on the YOLOv10n architecture [5]. YOLOv10 significantly pushes the YOLO series forward by proposing an end-to-end NMS-free training framework. It obtains it by following a unified dual label assignment scheme, which blends singular matches with multiple positive allocations. This also eliminates the requirement of NMS during inference, which greatly simplifies the pipeline and decreases deployment latency [5]. The 'nano' variant was chosen as a baseline since it delivers an optimal balance of processing overhead against detection precision.

B. Overall Architecture of LCEA-YOLO

LCEA-YOLO follows the standard YOLO pipeline [3] and maintains the three core functional blocks: a feature extractor (backbone), an aggregator (neck), and a predictor (head). The feature extractor extracts hierarchical feature maps at multiple spatial resolutions, while the PAN-FPN neck fuses these features through top-down and bottom-up paths to form rich multi-scale representations. The detection head then produces predictions at three scales (P3, P4, P5), enabling the detection of objects of different sizes, as commonly adopted in YOLO-style detectors [3, 5]. Figure 1 depicts the general structure of the LCEA-YOLO architecture, with the key modification being the integration of the proposed LCEA module into the detection head at each prediction scale (P3, P4, and P5). It is applied after the corresponding head branch computation, while the backbone and PAN-FPN neck remain unchanged, and the end-to-end inference pipeline of YOLOv10 is preserved [5].

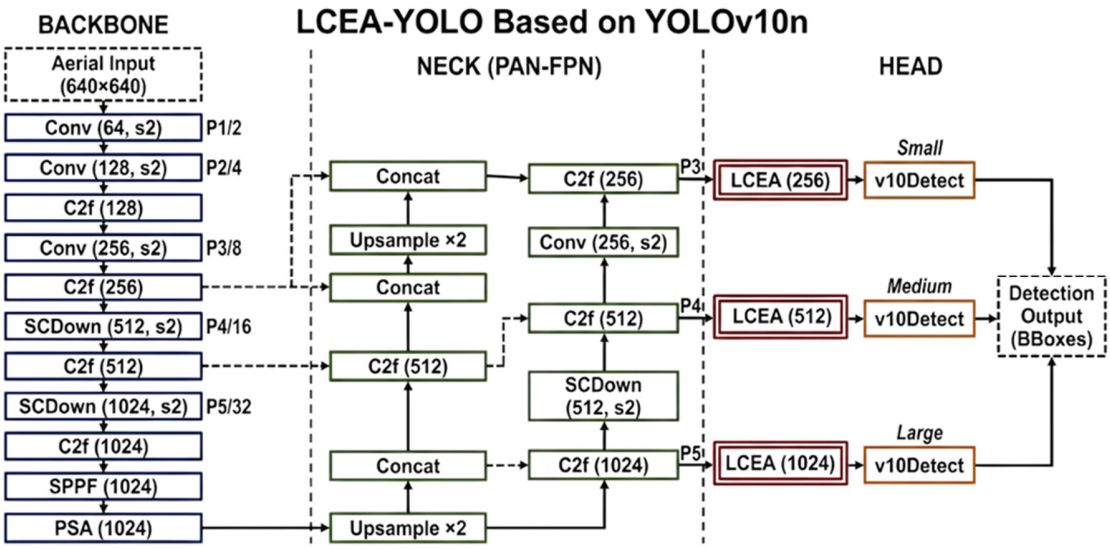

Fig. 1. General structure of the introduced LCEA-YOLO.

C. *Local Contrast Enhancement Attention Module*

Small objects in aerial imagery often exhibit poor visual distinction from the surrounding background [1, 2]. Conventional attention mechanisms (e.g., SE and CBAM) improve representation by reweighting features, but they are not specifically engineered to amplify local contrast cues that are especially relevant for small targets [10-12]. Motivated by local contrast-aware attention principles [12], the LCEA module is proposed, as illustrated in Figure 2.

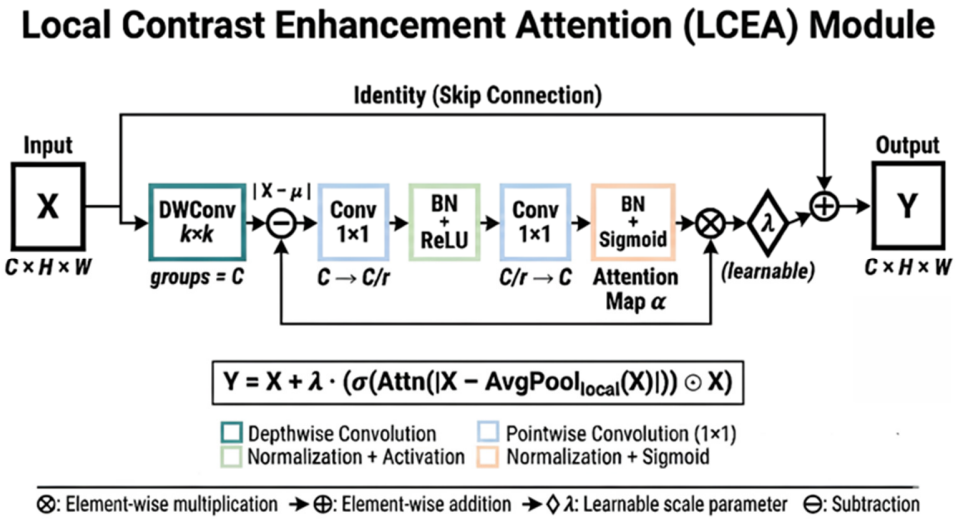

Fig. 2. LCEA module.

Let $X \in \mathbb{R}^{C\times H\times W}$ denote the input feature map, in which C specifies the channel depth, while H and W indicate the spatial dimensions. The module performs local contrast modeling and feature refinement as follows:

1) *Local Mean Estimation via Depthwise Convolution*

A depthwise convolution with a $k \times k$ kernel and groups $= C$ is used to approximate local averaging (i.e., local pooling) in a channel-wise manner:

$$\mu = DWConv_{k\times k}(X),\ groups = C \quad (1)$$

where $\mu \in \mathbb{R}^{C\times H\times W}$ represents the locally averaged response.

2) *Local Contrast Map Computation*

A local contrast map is obtained by measuring the absolute deviation from the local mean:

$$X_{contrast} = |X - \mu| \quad (2)$$

This operation highlights local variations that are often indicative of object boundaries and subtle structures.

3) Attention Map Generation

The contrast map is passed through an efficient attention sub-network composed of two 1×1 convolutions arranged in a bottleneck form $C \to C/r \to C$, $BN + ReLU$ after the first 1×1 convolution, and $BN + Sigmoid$ after the second one (as shown in Figure 2). The resulting attention map is:

$$\alpha = Attn(X_{contrast}) \quad (3)$$

where $\alpha \in \mathbb{R}^{C \times H \times W}$ and its values are constrained to [0, 1] by the Sigmoid inside the attention block.

4) Feature Refinement with Residual Scaling

Finally, the resulting attention mask scales the initial feature map X via point-wise multiplication and is added back through a residual connection:

$$Y = X + \lambda \cdot (\alpha \odot X) \quad (4)$$

where $Y \in \mathbb{R}^{C \times H \times W}$, $\odot$ denotes element-wise multiplication, and λ is a learnable scaling parameter initialized to zero to ensure stable training.

The depthwise convolution kernel size is set to $k = 7$, and the bottleneck reduction ratio is $r = 4$. In practice, λ can be initialized to zero to make the module behave close to an identity mapping at the beginning of training, improving optimization stability.

D. Inner-CIoU Loss Function

Precise bounding box regression is essential for small objects, as minor pixel-level errors can result in significant quality degradation. IoU-based metrics have progressed from GIoU [13] to DIoU [14], incorporating geometric constraints beyond overlap. Nevertheless, small bounding boxes are susceptible to reduced tolerance to localization errors. To tackle this issue, the Inner-IoU approach was incorporated [15] and applied within the CIoU formulation [14], producing an Inner-CIoU loss that computes the overlap using scaled inner boxes, increasing regression sensitivity for small objects. Formally, given a network-generated box B_{pred} and the actual annotated reference B_{gt}, Inner-IoU defines auxiliary boxes B'_{pred} and B'_{gt} by scaling them with a ratio $s \in (0, 1)$ around their centers [15]. Following the original Inner-IoU formulation, a value of $s = 0.7$ was set in all experiments.

The boundaries of the inner bounding box edges are calculated by:

$$b_l^{inner} = x_c - \frac{w \times s}{2}, \; b_r^{inner} = x_c + \frac{w \times s}{2} \quad (5)$$

$$b_t^{inner} = y_c - \frac{h \times s}{2}, \; b_b^{inner} = y_c + \frac{h \times s}{2} \quad (6)$$

where (x_c, y_c) indicate the midpoint locations of the localization box, while w and h stand for its horizontal and vertical dimensions.

The Inner-IoU (IoU_{inner}) is calculated as the Intersection over Union of these scaled inner prediction and ground-truth boxes:

$$IoU_{inner} = \frac{Area\,(B'_{pred} \cap B'_{gt})}{Area\,(B'_{pred} \cup B'_{gt})} \quad (7)$$

The IoU component in the regression loss is evaluated using these inner boxes, while the distance and aspect ratio constraints follow the CIoU formulation derived from DIoU-based components [14]. The final Inner-CIoU loss is defined as:

$$Loss_{inner-CIoU} = 1 - IoU_{inner} + \frac{\rho^2\,(c_{pred}, c_{gt})}{c^2} + \alpha v \quad (8)$$

where $\rho^2\,(c_{pred}, c_{gt})$ denotes the straight-line distance squared between the core coordinates of the proposed and target boxes, while c represents the longest cross-section of their smallest enclosing frame, v evaluates the aspect ratio consistency, and α acts as a balancing weight [14].

This modification provides stronger supervision when objects occupy few pixels in aerial views [1, 2].

III. EXPERIMENTS AND RESULTS

A. Experimental Setup

1) Dataset

For this study, the 'Aerial Vehicles Dataset' was utilized, published by UAVDT via Roboflow Universe [16]. This material was used under the Creative Commons Attribution 4.0 International license. To be in line with the intrinsic challenges mentioned in Section I, this database was particularly well-suited as a testing platform because it captures high-altitude viewpoints, where vehicles are represented by very few pixels and often have weak appearance contrast against a cluttered background [1, 2]. The dataset includes 8,626 high resolution annotated video frames of four different vehicle classes: bus, car, truck, and van. It was divided into 75% for training (6,469 images), 6.3% for validation (547 images), and 18.7% for testing (1,610 images).

2) Implementation Details

All models were implemented in the Ultralytics framework and optimized over 100 epochs utilizing a single NVIDIA RTX 5090 GPU. Following the standard training protocol of the YOLO family [3], the batch size was set at auto mode (batch = -1). This configuration allowed for the dynamic allocation of the largest feasible batch size supported by VRAM, making the training process more efficient. Model weights were updated via the AdamW optimizer [17] using an initial learning rate of 0.001 and a cosine annealing warmup schedule. The input resolution was fixed at 640 × 640 pixels for all experiments.

3) Evaluation Metrics

Performance was evaluated using COCO metrics [18], specifically Mean Average Precision (mAP), Recall (R), and Precision (P). For general detection assessment, mAP@0.5 was recorded, while mAP@0.5:0.95 was used for a rigorous evaluation of bounding box localization accuracy. To ensure the reliability of the reported results, all experiments were conducted over three independent runs with different random seeds (0, 42, and 123). The observed standard deviation of mAP@0.5 across runs was ± 0.3%, confirming the stability of the proposed improvements.

B. Ablation Study: Dissecting Component Contributions

An ablation study was carried out using the YOLOv10n baseline [5] to quantify the effect of the proposed architectural improvements and validate the hypothesis regarding local feature refinement. As illustrated in Table I, the off-the-shelf backbone and neck architecture lacked explicit mechanisms to improve fine-grained local cues, leading to a baseline mAP@0.5 of 44.2%.

TABLE I. ABLATION STUDY RESULTS ON UAVDT

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	mAP@0.5: 0.95 (%)
Baseline (YOLOv10n)	57.3	42.6	44.2	27.6
+ LCEA only	58.2	45.5	46.8	29.7
+ Inner-CIoU only	59.6	44.9	47.1	29.7
LCEA-YOLO (current)	59.2	45.2	47.2	29.9

The addition of the LCEA module itself brought a large improvement in mAP@0.5 by 2.6%. This validates the conjecture derived in Section II-C that by performing local pooling operation and calculating the absolute deviation between local mean and each pixel value, the LCEA module can capture informative local variations on or near object boundaries.

Moreover, it significantly contributed to the rise of Precision (from 57.3% to 59.6%) when using the Inner-CIoU loss. This further justifies the success of regressing with scaled inner boxes in producing more sensitive regression and stronger supervision, where objects appear at low resolution on aerial images [15]. The final LCEA-YOLO model, which unites both components, obtained the optimum results (47.2% in mAP@0.5 and 29.9% in mAP@0.5:0.95), revealing that the feature enhancement and loss function refinement yielded a synergistic effect.

C. Comparative Analysis with State-of-the-Art Models

To guarantee an unbiased and thorough evaluation, state-of-the-art baseline models (i.e., YOLOv8n, YOLOv10n, and YOLOv11n) were re-implemented on the dataset split [16]. Given that detection performance is extremely dependent on the training mode selected, presenting direct comparisons to literature figures obtained on setups of varying specifications (number of epochs and data augmentation pipeline) would be misleading. Thus, all reported comparative results in this work were trained from scratch following the exact same conditions as listed in Section III-A. This enabled the architectural contribution of the proposed method to be disentangled from external parameters for training and yields a very strict comparison baseline.

The comparative analysis in Table II demonstrates that LCEA-YOLO consistently outperforms standard general-purpose architectures. While YOLOv8n improved feature aggregation through the C2f module [4] and YOLOv10n simplified the inference pipeline [5], they still struggle with small, low-contrast targets. LCEA-YOLO surpassed the recent YOLOv11n by 1.9% in mAP@0.5 and achieved a significantly higher Precision of 59.2%. Furthermore, LCEA-YOLO added only 80,868 parameters (+3.5%) over YOLOv10n with a marginal increase of 0.2 GFLOPs, confirming that the accuracy gains are achieved without any additional computational burden.

TABLE II. COMPARISON WITH THE STATE-OF-THE-ART ON THE DATASET [16]

Model	Params (M)	GFLOP	Precision (%)	Recall (%)	mAP@0.5 (%)
YOLOv8n	3.01	8.1	56.7	42.7	44.2
YOLOv10n	2.27	6.5	57.3	42.6	44.2
YOLOv11n	2.58	6.3	57.9	43.5	45.3
LCEA-YOLO (current)	2.35	6.7	59.2	45.2	47.2

While recent adaptations of the YOLO framework have proven highly effective for real-time ground vehicle detection and autonomous driving applications [6], these comparative results validate that the targeted local contrast enhancements provide a distinct and necessary advantage for the unique challenges of specialized UAV deployments.

D. Per-Class Performance and Discussion

The advantage of LCEA-YOLO is more remarkable in individual object categories, such as the ones that are poorly visible against the background. The mAP@0.5 performance for each class is presented in Table III among the different classes of objects.

The biggest relative gain was registered for the truck (+15.0%). For the aerial surveillance, both bus and truck may be highly visually ambiguous due to diverse geometries and heavy occlusions. The bus class also exhibits a significant increase of +7.6%. Conversely, the car class, typically having more stable pixel footprints and contrast profiles, marginally benefits from this (+2.7%).

TABLE III. PER-CLASS MAP@0.5 (%) COMPARISON

Class	Baseline (YOLOv10n)	LCEA-YOLO (current)	Relative improvement (%)
Bus	47.6	51.2	+7.6%
Car	67.8	69.6	+2.7%
Truck	30.0	34.5	+15.0%
Van	31.5	33.7	+7.0%

This discrepancy explains the underlying advantage of the proposed approach: standard regression losses do not distinguish objects across scales, resulting in unstable optimization for challenging and low-frequency targets. LCEA-YOLO pushes the network to focus on local contrast and enhances the sensitivity of regression for small objects, which effectively overcomes the performance degradation associated with small size and visually confusing objects.

E. Qualitative Evaluation

To verify the numerical progress, a qualitative comparison was employed between the YOLOv10n baseline and the proposed LCEA-YOLO model in a challenging high-altitude aerial scene. As depicted in Figure 3, the reference architecture encountered difficulties when identifying miniature instances located at different distances to the camera, leading to a high

number of missed (13 False Negatives) and poorly localized objects (8 False Positives). The simple baseline often fails to confidently separate the subtle car parts from urban infrastructure and roads.

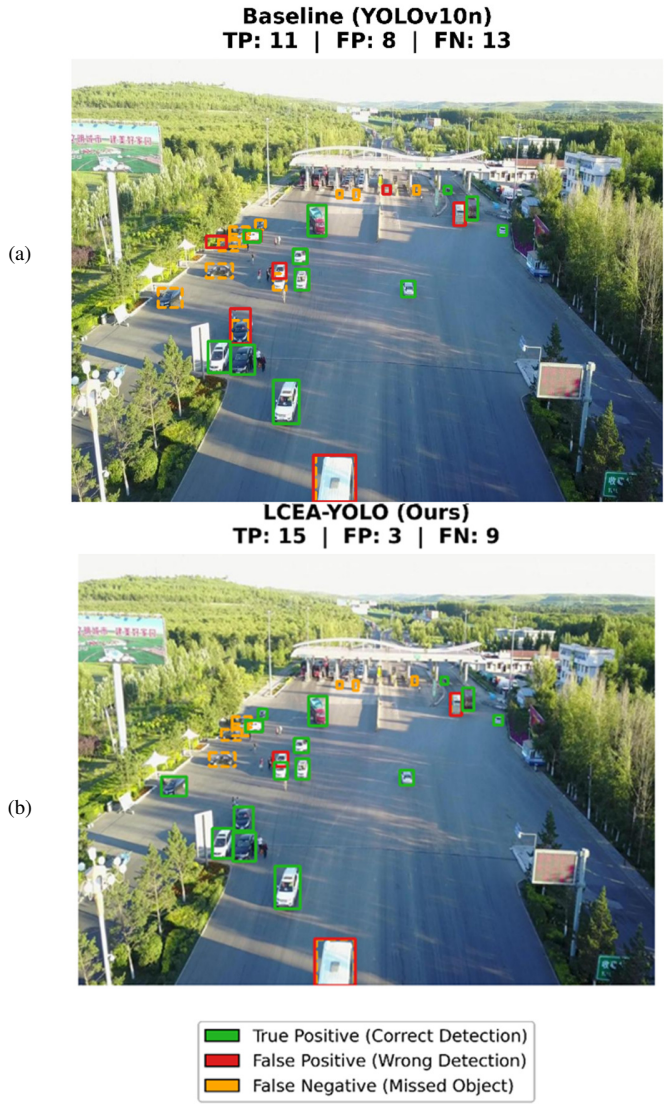

Fig. 3. Qualitative detection comparison between: (a) the YOLOv10n baseline and (b) the proposed LCEA-YOLO model.

On the other hand, LCEA-YOLO demonstrated much stronger robustness in this corresponding scene. The resulting model raised the True Positives count higher up to 15, which means that there is less loss of relevant items (down from 11), and meanwhile decreased False Negatives down to 9, drastically lowering False Positives by more than the half (8 dropped to 3). Qualitatively, this can be seen as much more correct looking bounding boxes (green) on far or hard to localize vehicles, and fewer ground truth localization failures in red. These qualitative results explain the quantitative findings and visually demonstrate that the integration of feature enhancement with the LCEA module and strict regression bounds in Inner-CIoU loss can effectively alleviate detection degradation against small aerial targets.

Figure 4 presents the normalized confusion matrices for both the YOLOv10n baseline and LCEA-YOLO, evaluated on the testing set. LCEA-YOLO demonstrated a consistent improvement in correctly identifying all four vehicle types, with the most notable gains observed for buses and trucks. Furthermore, the proportion of objects misclassified as background is notably reduced, particularly for trucks, confirming the quantitative improvements reported in Table III. These results validate that the LCEA module and Inner-CIoU loss effectively enhance the model's ability to distinguish small and visually similar targets from cluttered backgrounds.

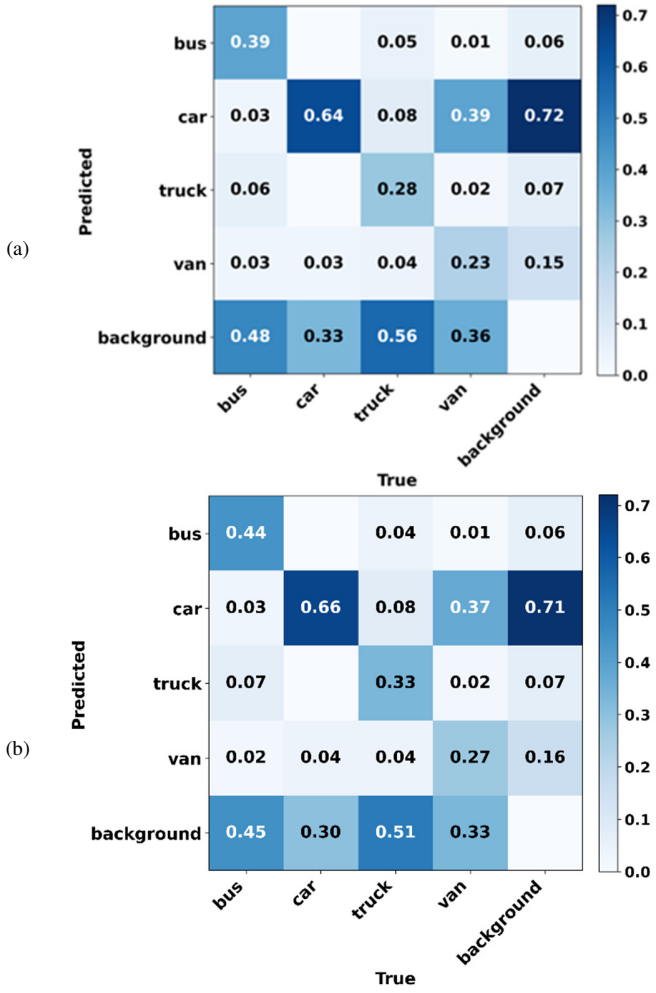

Fig. 4. Normalized confusion matrix on the testing set: (a) YOLOv10n baseline, (b) LCEA-YOLO.

In summary, the experimental results consistently verify the effectiveness of the proposed LCEA-YOLO across all evaluation dimensions. Quantitatively, the model achieves a +3.0% improvement over the YOLOv10n baseline and +1.9% over YOLOv11n in mAP@0.5, while adding only 3.5% more parameters. The per-class analysis reveals that the largest gains occur on the most challenging categories, with +15.0% for

trucks and +7.6% for buses. The qualitative evaluation further validates these findings by showing a substantial reduction in both missed detections and false alarms.

IV. CONCLUSION

To tackle the difficulties of small and visually uncertain object detection in UAV images, a powerful detector, called Local Contrast Enhancement Attention - You Only Look Once (LCEA-YOLO), which is based on YOLOv10n, was introduced. By incorporating the proposed LCEA module, the network effectively emphasized subtle local contrast changes, which are indispensable while separating small targets from cluttered backgrounds. Additionally, by using the Inner-CIoU loss function, this model could make a fast regression for small-sized bounding boxes, addressing the issues of localizing errors in conventional detection models.

The UAVDT dataset was utilized to confirm the viability of the proposed strategy [16]. LCEA-YOLO achieved an overall mAP@0.5 of 47.2%, exceeding the YOLOv10n baseline by 3.0% and the recent YOLOv11n by 1.9%. The ablation study indicated that the two proposed components jointly enhanced feature representation and localization accuracy. The model was also robust on underrepresented and difficult classes, with a 15.0% improvement for truck detection and a 7.6% gain in bus detection performance. These results confirmed that targeted architectural changes in the areas of local contrast and scale-sensitive regression can bring clear benefits beyond general-purpose YOLO architectures for aerial surveillance.

Despite the fact that LCEA-YOLO can achieve a competitive performance for the detection of objects in aerial imagery, there are still challenges, such as severe weather and heavily occluded cases. Future work will entail the deployment of LCEA-YOLO on lightweight embedded edge devices, such as NVIDIA Jetson Nano and Jetson Orin, to measure actual inference latency, throughput, and energy consumption under real-world UAV operating conditions. In addition, this line of research will be extended on other large-scale aerial datasets in order to obtain more evidence that the model performs well across various environmental domains.

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest regarding the publication of this paper.

DATA AVAILABILITY

This study utilized the "Aerial Vehicles Dataset" created by UAVDT and published via Roboflow [16] (available at https://universe.roboflow.com/uavdt/aerial-vehicles-hjarh).

ACKNOWLEDGMENT

The authors would like to thank UAVDT for making the Aerial Vehicles Dataset publicly available. This research received no external funding.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Google Gemini in order to polish the English language and improve readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] P. Zhu, L. Wen, X. Bian, H. Ling, and Q. Hu, "Vision Meets Drones: A Challenge." arXiv, Apr. 23, 2018, https://doi.org/10.48550/arXiv.1804.07437.

[2] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *15th European Conference on Computer Vision (ECCV 2018)*, Sept. 2018, pp. 370–386, https://doi.org/10.1007/978-3-030-01249-6_23.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, https://doi.org/10.1109/CVPR.2016.91.

[4] *Ultralytics YOLO*v8, G. Jocher, J. Qiu, and A. Chaurasia, 2023, https://github.com/ultralytics/ultralytics.

[5] A. Wang *et al.*, "YOLOv10: Real-Time End-to-End Object Detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107984–108011, Dec. 2024, https://doi.org/10.52202/079017-3429.

[6] M. Chaman *et al.*, "A Real-Time Vehicle Detection System for ADAS in Autonomous Vehicles Using YOLOv11 Deep Neural Network on Embedded Edge Platforms," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 28077–28082, Oct. 2025, https://doi.org/10.48084/etasr.12138.

[7] Y. Zhao *et al.*, "Detrs beat yolos on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Sept. 2024, https://doi.org/10.1109/CVPR52733.2024.01605.

[8] L. Tan, C. Zhang, H. Bai, Z. Liu, and Y. Li, "A real-time and efficient detector for small object in UAV aerial images," *Scientific Reports*, vol. 15, no. 1, Nov. 2025, Art. no. 39233, https://doi.org/10.1038/s41598-025-22855-w.

[9] Z. Xu, H. Zhao, P. Liu, L. Wang, G. Zhang, and Y. Chai, "SRTSOD-YOLO: Stronger Real-Time Small Object Detection Algorithm Based on Improved YOLO11 for UAV Imageries," *Remote Sensing*, vol. 17, no. 20, Dec. 2025, Art. no. 3414, https://doi.org/10.3390/rs17203414.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, https://doi.org/10.1109/CVPR.2018.00745.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *15th European Conference on Computer Vision (ECCV 2018)*, Sept. 2018, https://doi.org/10.1007/978-3-030-01234-2_1.

[12] Y. Dai, Y. Wu, F. Zhou, and K. Barnard, "Attentional Local Contrast Networks for Infrared Small Target Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 11, pp. 9813–9824, Nov. 2021, https://doi.org/10.1109/TGRS.2020.3044958.

[13] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, https://doi.org/10.1109/CVPR.2019.00075.

[14] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 12993–13000, Apr. 2020, https://doi.org/10.1609/aaai.v34i07.6999.

[15] H. Zhang, C. Xu, and S. Zhang, "Inner-IoU: More Effective Intersection over Union Loss with Auxiliary Bounding Box." arXiv, Nov. 14, 2023, https://doi.org/10.48550/arXiv.2311.02877.

[16] *Aerial Vehicles Computer Vision Dataset*, Roboflow, https://universe.roboflow.com/uavdt/aerial-vehicles-hjarh.

[17] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." arXiv, Jan. 04, 2019, https://doi.org/10.48550/arXiv.1711.05101.

[18] T. Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *13th European Conference of Computer Vision – ECCV*, Sept. 2014, https://doi.org/10.1007/978-3-319-10602-1_48.