Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MECA-Net: small traffic object detection in UAV imagery via multiscale edge–coordinate attention

Liping Wang
Zhengzhou University of Light Industry
Mingxin Han
Zhengzhou University of Light Industry
Yan Chen
Zhengzhou University of Light Industry
Heng Li
The Sixth Design and Research Institute of Mechanical Industry Co., Ltd.
Dongyao Zou
zdy@zzuli.edu.cn

Zhengzhou University of Light Industry

Article

Keywords: UAV imagery, small-object detection, edge-aware attention, coordinate attention, YOLO11n

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10651356/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The authors declare no competing interests.

MECA-Net: small traffic object detection in UAV imagery via multiscale edge–coordinate attention

Wang Liping[1], Han Mingxin[1], Chen Yan[1], Li Heng[2], and Zou Dongyao[1,*]

[1]School of Computer Science and Artificial Intelligence, Zhengzhou University of Light Industry, Zhengzhou 450000, China
[2]The Sixth Design and Research Institute of Mechanical Industry Co., Ltd., Zhengzhou, 450000, China
[*]zdy@zzuli.edu.cn

ABSTRACT

Small traffic objects in unmanned aerial vehicle (UAV) imagery often have limited pixel representation and weak boundary cues, making accurate real-time detection challenging. We present MECA-Net, a YOLO11n-based detector that integrates multi-scale contextual, edge, coordinate and high-resolution information. A multi-scale edge-aware attention module (MSEAA) combines depthwise dilated convolutions with fixed Sobel-gradient features through channel recalibration and residual learning. A coordinate-aware dual-branch attention module (CADBA) encodes positional dependencies separately along the height and width dimensions, while a stride-4 P2 branch extends the prediction pyramid to four scales. On the VisDrone2019-DET validation set, MECA-Net achieved 38.8% mAP_{50}, 22.4% $\mathrm{mAP}_{50:95}$, 8.5% AP_S and 38.6% recall, improving mAP_{50} and $\mathrm{mAP}_{50:95}$ by 6.0 and 3.2 percentage points over YOLO11n. The model contains 2.95 million parameters, requires 7.30 GFLOPs and achieves 158 FPS on an NVIDIA RTX 4090. These results indicate that integrating boundary, coordinate and high-resolution information is associated with improved small-object detection under the evaluated UAV traffic conditions while retaining real-time throughput on the tested desktop GPU.

Keywords: UAV imagery; small-object detection; edge-aware attention; coordinate attention; YOLO11n

Unmanned aerial vehicles (UAVs) provide flexible, low-cost platforms for acquiring high-resolution visual data and are increasingly used in intelligent transportation, urban surveillance, emergency response, infrastructure inspection and environmental monitoring. The elevated imaging altitude and variable viewpoint, however, produce many small and densely distributed objects. UAV imagery is also affected by substantial scale variation, occlusion, camera motion and complex urban backgrounds, as represented in the VisDrone benchmark[1].

Deep learning has substantially advanced object detection. Two-stage methods such as Faster R-CNN first generate candidate regions and then refine their categories and bounding boxes, providing reliable localisation at the cost of a more complex inference pipeline[2]. One-stage detectors formulate detection as direct dense prediction: SSD predicts default boxes on multiple feature maps[3], whereas the YOLO framework performs unified prediction in a single forward pass and is widely used because of its favourable speed–accuracy trade-off[4].Across successive YOLO variants, advances have involved backbone and neck design, multi-scale feature fusion, training and label-assignment strategies, and detection-head architectures[5–13]. Transformer-based detectors offer an alternative set-prediction formulation[14], and RT-DETR improves the accuracy–speed balance through an efficient hybrid encoder and query selection[15], although computational cost remains an important consideration for UAV deployment. Multi-scale representation is a common response to the severe scale variation of aerial objects. Feature Pyramid Networks propagate high-level semantics to high-resolution features through top-down and lateral connections[16], RetinaNet addresses foreground–background imbalance with focal loss[17], PANet adds a bottom-up path to shorten the information route between shallow localisation cues and deep semantics[18], and BiFPN performs repeated bidirectional fusion with learnable weights[19]. Recent UAV-oriented detectors increasingly preserve shallow high-resolution information while improving cross-scale interaction. LRDS-YOLO combines adaptive downsampling with a recalibrated feature pyramid[20], whereas SMA-YOLO uses bidirectional multi-branch fusion and an adaptive multi-scale prediction head[21].

Attention mechanisms can improve feature discrimination without substantially increasing model size. SENet recalibrates channel responses through global aggregation[22], CBAM combines channel and spatial attention[23], and Coordinate Attention preserves directional positional information through factorised pooling[24]. Recent aerial detectors further combine attention with multi-resolution fusion; Aero-LiteNet integrates multi-scale feature aggregation and cross-dimensional attention for cluttered UAV scenes[25]. Closely related YOLO11n-based work has also explored multi-scale edge-feature selection together with a high-resolution $P2$ prediction head[26]. The joint integration of explicit fixed-gradient priors, directional coordinate encoding and high-resolution prediction nevertheless remains comparatively underexplored.

Recent aerial studies have demonstrated the benefits of dedicated high-resolution branches and cross-scale interaction. CF-YOLO constructs a cross-scale feature pyramid on YOLO11 and reports a 12.7-percentage-point improvement in mAP_{50} over its baseline on VisDrone[27], while He et al. examine the broader potential of YOLO11 for remote-sensing detection[28]. These studies support the preservation of shallow spatial detail, but adding feature paths alone may also retain background noise. MECA-Net therefore combines semantic context, fixed edge gradients, directional coordinate encoding and a stride-4 prediction branch. The main contributions of this study are as follows:

- MSEAA replaces three selected C3k2 blocks associated with the $P2$, $P3$ and $P4$ stages of the YOLO11n backbone and combines parallel depthwise dilated convolutions with fixed x- and y-direction Sobel gradients. Adaptive channel recalibration and residual projection integrate contextual and boundary information while maintaining a compact backbone representation.
- CADBA models local context and encodes positional information independently along the height and width dimensions. CADBA is applied to each fused $P2$–$P5$ neck feature before prediction.
- SDL adds a stride-4, 160×160 $P2$ prediction branch to the original $P3$–$P5$ pyramid, providing denser spatial sampling for very small road users.
- Experiments on VisDrone2019-DET indicate that MECA-Net improves small-object detection under the evaluated conditions while remaining compact. The reported results are based on a single training run and do not estimate training variability.

Related work

Real-Time Object Detectors

Deep-learning-based object detectors can be grouped into two-stage, one-stage and Transformer-based frameworks. Faster R-CNN integrates region proposal into the detection network[2], whereas Cascade R-CNN progressively refines predictions using increasing intersection-over-union thresholds[29]. Their region-wise processing is effective but adds inference overhead. One-stage detectors directly predict object categories and bounding boxes from dense feature maps. The YOLO family has progressively improved the computational and representational efficiency of this formulation[4–13], whereas anchor-free approaches such as FCOS and CenterNet simplify the prediction pipeline[30,31]. DETR and RT-DETR remove several hand-designed components through set prediction and efficient query selection[14,15]. YOLO11n was selected as the baseline in this study because its official implementation and reported architecture provide a compact speed–accuracy starting point for UAV small-object detection[12,13].

Attention mechanisms in detection

Attention modules reweight informative features across channels, spatial positions or coordinate directions. Squeeze-and-excitation networks learn channel gates from global descriptors[22], whereas CBAM applies channel and spatial attention sequentially[23]. Coordinate Attention factorises two-dimensional pooling into height- and width-direction encodings, retaining location cues that conventional global pooling may remove[24]. Recent aerial work extends these ideas through multi-resolution and cross-dimensional attention[25], while a closely related YOLO11n-based detector combines multi-scale edge-feature selection with a high-resolution $P2$ prediction head[26]. MSEAA and CADBA differ from these approaches by explicitly introducing fixed channel-wise Sobel gradients and applying directional coordinate recalibration to the fused $P2$–$P5$ features.

Small-object detection

Small-object detection is difficult in aerial and traffic imagery because objects occupy few pixels and are easily confused with cluttered backgrounds. Feature-pyramid methods have evolved from FPN and RetinaNet to PANet and BiFPN[16–19]. Recent UAV detectors increasingly coordinate high-resolution feature preservation, adaptive downsampling and multi-scale fusion[20,21]. Regression objectives also influence localisation; GIoU improves bounding-box optimisation in non-overlapping cases[32].Recent UAV-oriented studies have explored multi-resolution attention and feature fusion[25]., edge-aware feature selection with high-resolution prediction[26]., cross-scale feature-pyramid design[27]., and broader YOLO11 adaptation to remote-sensing imagery[28]. MECA-Net adds fixed-gradient boundary priors and directional coordinate cues to this design space.

Methods

Overview of MECA-Net

The YOLO11n baseline follows the conventional backbone–neck–head paradigm and performs prediction at the $P3$, $P4$ and $P5$ feature levels, with output strides of 8, 16 and 32, respectively[12,13]. MECA-Net retains the overall single-stage detection framework while introducing three modifications for UAV small-object detection, as illustrated in Figure 1.

First, three selected C3k2 blocks associated with the $P2$, $P3$ and $P4$ stages of the YOLO11n backbone are replaced by the proposed multi-scale edge-aware attention module (MSEAA). The replacement combines multi-scale contextual modelling with fixed-gradient boundary information while maintaining a compact backbone. Second, a coordinate-aware dual-branch attention module (CADBA) recalibrates each fused $P2$–$P5$ neck feature before prediction. Third, a small-object detection layer (SDL) extends the original $P3$–$P5$ prediction pyramid with an additional stride-4 $P2$ branch.

MSEAA processes multi-scale semantic context and fixed-gradient edge information through two parallel branches. The semantic branch employs depthwise dilated convolutions with different dilation rates to enlarge the effective receptive field without substantially increasing the parameter count. In parallel, the edge branch applies fixed Sobel operators independently to each channel to extract gradients along the x- and y-directions. The semantic and edge features are subsequently concatenated, batch-normalised, adaptively recalibrated through channel attention and projected to the required channel dimension before residual addition. This design enables the backbone to retain weak contour cues while incorporating complementary contextual information.

CADBA operates on the fused features generated by the multi-scale neck. It first models local context using layer normalisation and depthwise convolution. Global average pooling is then performed separately along the height and width dimensions, preserving positional information in one direction while aggregating long-range dependencies in the other. The resulting directional descriptors are transformed into height- and width-aware attention maps, whose joint response is applied to the local feature. A residual feed-forward block further enhances the recalibrated representation while maintaining stable feature propagation.

The original $P3$–$P5$ prediction pyramid is extended to $P2$–$P5$ by introducing SDL. For an input resolution of 640×640, the additional $P2$ feature map has a spatial resolution of 160×160 and an output stride of 4. This high-resolution branch combines shallow spatial detail with upsampled semantic information and provides denser prediction locations for small traffic objects. The anchor-free decoupled detection formulation of YOLO11n is retained, while the neck and prediction head are extended to incorporate the additional $P2$ feature.

Multi-scale edge-aware attention (MSEAA)

Small traffic objects may contain limited texture and weak boundary responses after repeated downsampling. To improve the representation of these objects while maintaining a compact backbone, MSEAA replaces three selected C3k2 blocks associated with the $P2$, $P3$ and $P4$ stages of the YOLO11n backbone, as illustrated in Fig. 1. The module combines a multi-scale semantic branch, a fixed-gradient edge branch, adaptive channel recalibration and a residual projection, as shown in Fig. 2(a). Let $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times H \times W}$ denote the input feature map, where C_{in}, H and W represent the input-channel number, height and width, respectively. The output channel dimension of the replaced backbone stage is denoted by C_{out}.

Multi-scale semantic branch

A 3×3 max-pooling operation with stride 1 first aggregates local responses without changing the spatial resolution:

$$\mathbf{X}_p = \text{MaxPool}_{3\times3, s=1, p=1}(\mathbf{X}), \tag{1}$$

Two parallel 3×3 depthwise convolutions with dilation rates of 2 and 3 are subsequently applied to capture complementary contextual information:

$$\mathbf{F}_s = \text{Concat}\left[\text{DWConv}_{3\times3, d=2}(\mathbf{X}_p), \text{DWConv}_{3\times3, d=3}(\mathbf{X}_p)\right], \tag{2}$$

where $\mathbf{F}_s \in \mathbb{R}^{2C_{\text{in}} \times H \times W}$. The padding of each dilated convolution is selected to preserve the spatial resolution.
For a convolutional kernel of size k and dilation rate d, the effective receptive-field size is

$$k_{\text{eff}} = k + (k-1)(d-1). \tag{3}$$

Accordingly, the two 3×3 branches correspond to effective receptive fields of 5×5 and 7×7, respectively. Because depthwise convolution operates independently on each channel, the receptive field is enlarged with substantially lower computational cost than that of standard large-kernel convolution.

Fixed-gradient edge branch

In parallel with the semantic branch, fixed Sobel kernels are applied independently to each input channel to extract gradients along the horizontal and vertical directions:

$$\mathbf{K}_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad \mathbf{K}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{4}$$

The fixed Sobel operators are applied channel-wise with unit padding so that the spatial resolution is preserved.

$$\mathbf{G}_x = \mathbf{X} *_{\mathrm{dw}} \mathbf{K}_x, \qquad \mathbf{G}_y = \mathbf{X} *_{\mathrm{dw}} \mathbf{K}_y, \qquad \mathbf{F}_e = \mathrm{Concat}[\mathbf{G}_x, \mathbf{G}_y], \tag{5}$$

where $*_{\mathrm{dw}}$ denotes channel-wise two-dimensional convolution with padding of 1. Consequently, $\mathbf{G}_x, \mathbf{G}_y \in \mathbb{R}^{C_{\mathrm{in}} \times H \times W}$ and $\mathbf{F}_e \in \mathbb{R}^{2C_{\mathrm{in}} \times H \times W}$.The Sobel kernels remain fixed during training and therefore do not introduce additional learnable parameters.

Adaptive channel recalibration and residual projection

The semantic and edge features are concatenated along the channel dimension and batch-normalised:

$$\mathbf{F} = \mathrm{BN}\left(\mathrm{Concat}\left[\mathbf{F}_s, \mathbf{F}_e\right]\right), \qquad \mathbf{F} \in \mathbb{R}^{4C_{\mathrm{in}} \times H \times W}. \tag{6}$$

Global average pooling is then used to obtain a channel descriptor. Following the channel-recalibration principle of SENet[22], two 1×1 pointwise convolutions and a sigmoid function generate an attention vector for the $4C_{\mathrm{in}}$ fused channels:

$$\mathbf{A}_c = \sigma\left[\mathbf{W}_2\, \delta\left(\mathbf{W}_1\, \mathrm{GAP}(\mathbf{F})\right)\right], \qquad \widehat{\mathbf{F}} = \mathbf{A}_c \odot \mathbf{F}. \tag{7}$$

Here, $\mathrm{GAP}(\mathbf{F}) \in \mathbb{R}^{4C_{\mathrm{in}} \times 1 \times 1}$. The first pointwise convolution reduces the channel dimension from $4C_{\mathrm{in}}$ to an intermediate dimension C_a, and the second restores it to $4C_{\mathrm{in}}$, where

$$C_a = \max\left(1, \left\lfloor \frac{4C_{\mathrm{in}}}{r} \right\rfloor\right), \tag{8}$$

and r denotes the channel-reduction ratio used in the implementation. In all experiments, the channel-reduction ratio was set to $r = 16$. The function $\delta(\cdot)$ denotes the SiLU activation and $\sigma(\cdot)$ denotes the sigmoid function.

Because MSEAA replaces C3k2 blocks whose input and output channel dimensions may differ, the recalibrated feature and the residual branch are projected to C_{out} channels before element-wise addition:

$$\begin{aligned} \mathscr{R}(\mathbf{X}) &= \begin{cases} \mathbf{X}, & C_{\mathrm{in}} = C_{\mathrm{out}}, \\ \mathrm{Conv}_{1\times1}(\mathbf{X}), & C_{\mathrm{in}} \neq C_{\mathrm{out}}, \end{cases} \\ \mathbf{Y}_m &= \mathscr{R}(\mathbf{X}) + \mathrm{CBS}_{1\times1}^{4C_{\mathrm{in}} \to C_{\mathrm{out}}}\left(\widehat{\mathbf{F}}\right). \end{aligned} \tag{9}$$

Here, $\mathrm{CBS}_{1\times1}$ denotes a 1×1 convolution followed by batch normalisation and SiLU activation. The resulting feature satisfies $\mathbf{Y}_m \in \mathbb{R}^{C_{\mathrm{out}} \times H \times W}$, allowing MSEAA to directly replace the corresponding C3k2 block while preserving the required stage output dimensions.

Coordinate-aware dual-branch attention

Conventional two-dimensional global pooling compresses the spatial dimensions into a single channel descriptor and may therefore discard the positional information of small objects. Inspired by Coordinate Attention[24], the coordinate-aware dual-branch attention module (CADBA) combines local-context modelling, directional coordinate encoding and residual feed-forward enhancement, as illustrated in Fig. 2(b). Given the input feature $\mathbf{Y}_m \in \mathbb{R}^{C \times H \times W}$, layer normalisation followed by a 3×3 depthwise convolution first produces the local context feature

$$\mathbf{Z} = \mathrm{DWConv}_{3\times3}\left(\mathrm{LN}(\mathbf{Y}_m)\right), \tag{10}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalisation. LN denotes the layer-normalisation operation used in the implementation and does not alter the feature dimensions.

To retain directional positional information, average pooling is performed separately along the width and height dimensions. The resulting descriptors preserve the height and width coordinates, respectively:

$$\begin{aligned} \mathbf{z}_h(c,h) &= \frac{1}{W}\sum_{w=1}^{W}\mathbf{Z}(c,h,w), \\ \mathbf{z}_w(c,w) &= \frac{1}{H}\sum_{h=1}^{H}\mathbf{Z}(c,h,w). \end{aligned} \tag{11}$$

Here, $\mathbf{z}_h \in \mathbb{R}^{C\times H\times 1}$ retains the height coordinate, whereas $\mathbf{z}_w \in \mathbb{R}^{C\times 1\times W}$ retains the width coordinate. After transposing $\mathbf{z}_w$, the two descriptors are concatenated along the spatial dimension and transformed using a 1×1 convolution followed by batch normalisation. The resulting feature is then split into height- and width-direction branches, from which the corresponding attention maps are generated:

$$\begin{aligned} \mathbf{Q} &= \mathrm{BN}\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{Concat}[\mathbf{z}_h,\mathbf{z}_w^T]\right)\right), \\ (\mathbf{Q}_h,\widetilde{\mathbf{Q}}_w) &= \mathrm{Split}(\mathbf{Q}), \\ \mathbf{A}_h &= \sigma(\mathrm{DWConv}(\mathbf{Q}_h)), \\ \mathbf{A}_w &= \left[\sigma\left(\mathrm{DWConv}(\widetilde{\mathbf{Q}}_w)\right)\right]^T. \end{aligned} \tag{12}$$

where $\mathbf{Q}_h \in \mathbb{R}^{C\times H\times 1}$ and $\widetilde{\mathbf{Q}}_w \in \mathbb{R}^{C\times W\times 1}$. The width-direction response is transposed back after depthwise convolution, yielding $\mathbf{A}_w \in \mathbb{R}^{C\times 1\times W}$, whereas $\mathbf{A}_h \in \mathbb{R}^{C\times H\times 1}$. The attention maps $\mathbf{A}_h \in \mathbb{R}^{C\times H\times 1}$ and $\mathbf{A}_w \in \mathbb{R}^{C\times 1\times W}$ are broadcast to $H\times W$ and jointly applied to the local feature. The 1×1 convolution preserves the channel dimension C. A residual connection then preserves the original input representation:

$$\mathbf{Y}_c = \mathbf{Y}_m + \mathbf{Z}\odot\mathbf{A}_h\odot\mathbf{A}_w, \tag{13}$$

where $\odot$ denotes element-wise multiplication with broadcasting.
Finally, a pre-normalisation feed-forward network enhances nonlinear inter-channel interactions:

$$\mathbf{Y}_{\mathrm{out}} = \mathbf{Y}_c + \mathrm{MLP}\left(\mathrm{LN}(\mathbf{Y}_c)\right). \tag{14}$$

The MLP comprises two 1×1 convolutions separated by a nonlinear activation. The first residual path preserves the input feature while adding joint height- and width-direction recalibration, and the second residual path enhances channel-wise representation. This dual-residual formulation allows CADBA to incorporate directional positional information while maintaining stable propagation of fine-grained features.

Small-object detection layer

The baseline YOLO11n detector performs prediction at the $P3$, $P4$ and $P5$ feature levels, with output strides of 8, 16 and 32, respectively. For an input resolution of 640×640, the corresponding feature maps have spatial resolutions of 80×80, 40×40 and 20×20. Objects whose width or height is smaller than the stride-8 sampling interval of $P3$ may be represented by only a few spatial locations, which can weaken classification responses and reduce bounding-box regression stability.

To preserve finer spatial information, the small-object detection layer (SDL) introduces the stride-4 backbone feature $\mathbf{C}_2$ into the neck and fuses it with the upsampled $P3$ feature:

$$\mathbf{P}_2 = \Phi\left(\mathrm{Up}(\mathbf{P}_3),\mathbf{C}_2\right), \tag{15}$$

where $\mathrm{Up}(\cdot)$ denotes twofold upsampling and $\Phi(\cdot)$ denotes the implemented feature-fusion operation. For a 640×640 input, the resulting $\mathbf{P}_2$ feature map has a spatial resolution of 160×160 and an output stride of 4. This high-resolution feature is supplied to an additional prediction branch, so the detector performs prediction at four feature levels:

$$\mathscr{Y} = \{\mathrm{Head}(\mathbf{P}_s) \mid s\in\{2,3,4,5\}\}. \tag{16}$$

The anchor-free decoupled prediction formulation of YOLO11n is retained, while the neck and prediction pyramid are extended to incorporate the additional high-resolution feature. The $P2$ branch provides denser prediction locations for small traffic objects and preserves more shallow spatial detail than the original $P3$–$P5$ pyramid. Because the submitted ablation study evaluates SDL only after MSEAA and CADBA have been introduced, the reported change should be interpreted as the cumulative effect of adding SDL to that configuration rather than as an isolated standalone effect.

Experiments

Datasets

All experiments are conducted on the VisDrone2019-DET dataset, a widely used benchmark for UAV-based object detection. The complete benchmark contains 10,209 static images, comprising 6,471 training images, 548 validation images, 1,610 test-dev images and 1,580 test-challenge images, with ten annotated object categories (pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle and tricycle). Because the images are captured from elevated viewpoints, the dataset contains many small and densely distributed objects; frequent occlusion and complex urban backgrounds make it well suited for evaluating small-object detection methods. All models are trained on the training set and evaluated on the validation set.

Implementation Details

All experiments for the YOLO11n baseline and MECA-Net were conducted on the same workstation equipped with an NVIDIA GeForce RTX 4090 GPU (24 GB) and an Intel Core i9-13900K CPU, running Ubuntu 22.04 with PyTorch 2.1 and CUDA 12.1. Input images were resized to 640×640. The YOLO11n baseline and MECA-Net were trained for 300 epochs using SGD with an initial learning rate of 0.01, a momentum of 0.937 and a weight decay of 5×10^{-4}. The learning rate followed cosine annealing with a 3-epoch warm-up, and the batch size was set to 32. Data augmentation included random scaling in the range of 0.5–1.5, random horizontal flipping, Mosaic and MixUp. A fixed random seed of 42 was used for the reported experiments, and data augmentation was disabled during validation and inference.

For the baseline detectors reproduced in this study, the same VisDrone2019-DET training and validation splits, input resolution and evaluation protocol were used. Where results from previously published VisDrone2019 experiments are included for comparison, they are explicitly identified in Table 1. Model-specific components, including detection losses, label-assignment strategies and prediction mechanisms, were retained from their corresponding implementations unless otherwise stated. This avoids altering the original optimisation objectives of different detector architectures while maintaining a common dataset and evaluation setting.

Inference speed of the YOLO11n baseline and MECA-Net was evaluated on the same NVIDIA RTX 4090 GPU using an input resolution of 640×640, a batch size of 1 and FP32 precision. Before timing, 50 warm-up iterations were performed to stabilise GPU execution. Latency was measured over all 548 images in the VisDrone2019-DET validation set with CUDA synchronisation. The complete timing procedure was repeated three times, and the mean throughput was reported as FPS. The reported timing includes image preprocessing, model inference and post-processing, including non-maximum suppression.

Evaluation Metrics

Detection performance is evaluated using precision (P), recall (R), mAP_{50}, $\mathrm{mAP}_{50:95}$, and small-object average precision (AP_S). Model complexity is evaluated using the number of trainable parameters (Params) and floating-point operations (GFLOPs), while inference efficiency is measured in frames per second (FPS).

Precision and recall are defined as

$$P = \frac{TP}{TP + FP}, \tag{17}$$

$$R = \frac{TP}{TP + FN}, \tag{18}$$

where TP, FP, and FN denote the numbers of true-positive, false-positive, and false-negative detections, respectively. During validation, predictions with confidence scores below 0.001 were discarded before constructing the precision–recall curves. For detectors requiring non-maximum suppression (NMS), an IoU threshold of 0.7 was used during post-processing. The same validation settings were applied to the YOLO11n baseline and MECA-Net.

The mAP_{50} metric denotes mean average precision evaluated at an intersection-over-union (IoU) threshold of 0.50. In contrast, $\mathrm{mAP}_{50:95}$ is averaged over IoU thresholds from 0.50 to 0.95 with an interval of 0.05, providing a more stringent assessment of both classification and localisation performance.

For locally evaluated models, small-object performance is additionally evaluated using AP_S following the COCO evaluation protocol. Objects with a ground-truth bounding-box area smaller than 32^2 pixels are treated as small objects. Before evaluation, predicted boxes are mapped back to the native resolution of each image. Small-object categories are determined from the ground-truth annotation areas in the original-image coordinate system. This metric is reported separately because the primary objective of MECA-Net is to improve the detection of small traffic objects in UAV imagery.

The number of trainable parameters is calculated directly from the model parameters. GFLOPs are measured using the `get_flops` utility in Ultralytics 8.3.0 with `ultralytics-thop` 2.0.0 and an input tensor of size $1 \times 3 \times 640 \times 640$. The profiler counts a multiplication and an addition as two floating-point operations. The same profiling procedure is used for all models reproduced locally in this study. For models whose results are taken from previously published studies, the reported parameter counts and GFLOPs are retained from the corresponding sources.

Inference speed of the YOLO11n baseline and MECA-Net configurations is evaluated on the same NVIDIA GeForce RTX 4090 GPU using an input resolution of 640×640, a batch size of 1, and FP32 precision. Before timing, 50 warm-up iterations are performed to stabilise GPU execution. Latency is then measured over all 548 images in the VisDrone2019-DET validation set with CUDA synchronisation. The reported FPS includes image preprocessing, model inference, and post-processing. For models requiring non-maximum suppression, the NMS operation is also included in the measured latency. The same timing procedure is applied to all configurations reported in the ablation study.

Comparison with Representative Object Detectors

Table 1 compares MECA-Net with representative lightweight and larger-scale detectors on the VisDrone2019-DET validation set. Under the same evaluation setting as the YOLO11n baseline, MECA-Net increases mAP_{50} from 32.8% to 38.8%, $mAP_{50:95}$ from 19.2% to 22.4%, and AP_S from 5.6% to 8.5%, corresponding to gains of 6.0, 3.2 and 2.9 percentage points, respectively. Among the reproduced detectors with approximately 3 million parameters or fewer, MECA-Net achieves the highest accuracy in the three reported AP metrics. The complete model contains 2.95 M parameters and requires 7.30 GFLOPs, compared with 2.58 M parameters and 6.50 GFLOPs for YOLO11n. Under the common RTX 4090 timing protocol, MECA-Net achieves 158 FPS.

Model	mAP_{50} (%)	$mAP_{50:95}$ (%)	AP_S (%)	Params (M)	GFLOPs
YOLOv5n[7†]	31.7	18.2	–	2.50	7.10
YOLOv8n[9†]	32.1	18.5	–	3.00	8.10
YOLOv9t[10]	33.0	18.0	6.5	2.00	7.70
YOLOv10n[11†]	32.2	18.6	–	2.26	6.50
YOLO11n[12]	32.8	19.2	5.6	2.58	6.50
YOLOv8s[9]	43.0	26.0	9.2	11.10	28.50
YOLOv9s[10]	41.0	25.0	8.5	7.10	26.40
SL-YOLO[33]	46.9	28.9	10.1	9.60	36.70
YOLOv10s[11‡]	40.0	24.0	–	8.10	24.60
YOLO11s[12]	44.6	29.4	9.5	9.42	21.30
MASF-YOLO-s[34]	49.2	32.9	11.3	12.05	44.30
YOLO11m[12]	47.8	32.2	10.8	20.04	67.70
RT-DETR-R18[15]	36.8	26.4	7.2	20.09	77.00
MECA-Net (Ours)	**38.8**	**22.4**	**8.5**	**2.95**	**7.30**

Table 1. Performance comparison with representative object detectors on the VisDrone2019-DET dataset. AP_S denotes average precision for small objects according to the COCO definition. "–" indicates that AP_S was not reported. Values marked with † and ‡ were obtained from previous VisDrone2019 comparison experiments. † Values for YOLOv5n, YOLOv8n and YOLOv10n are from Ref.[35]. ‡ Values for YOLOv10s are from Ref.[36].

Ablation Study

Table 2 presents the incremental ablation results for MSEAA, CADBA and SDL. Replacing the three selected C3k2 backbone blocks with MSEAA increases precision from 43.3% to 45.2%, mAP_{50} from 32.8% to 34.3%, and $mAP_{50:95}$ from 19.2% to 20.0%. At the same time, the parameter count decreases from 2.58 M to 2.45 M and the computational cost decreases from 6.50 to 6.40 GFLOPs, indicating that the first-stage accuracy improvement is achieved while slightly reducing the overall model complexity.

Baseline	MSEAA	CADBA	SDL	Precision (%)	Recall (%)	mAP50 (%)	mAP50–95 (%)	Params (M)	GFLOPs	FPS
✓	–	–	–	43.3	33.1	32.8	19.2	2.58	6.50	340
✓	✓	–	–	45.2	33.8	34.3	20.0	2.45	6.40	333
✓	✓	✓	–	48.1	35.8	36.4	20.9	2.87	7.05	190
✓	✓	✓	✓	**48.2**	**38.6**	**38.8**	**22.4**	**2.95**	**7.30**	**158**

Table 2. Ablation study of MSEAA, CADBA and SDL on the VisDrone2019-DET validation set. The proposed components are progressively introduced into the YOLO11n baseline. The results represent cumulative effects obtained by adding each module sequentially.

Adding CADBA to the MSEAA-based configuration further increases precision from 45.2% to 48.1%, recall from 33.8% to 35.8%, and mAP_{50} from 34.3% to 36.4%, while $mAP_{50:95}$ increases from 20.0% to 20.9%. This step increases the parameter count to 2.87 M and the computational cost to 7.05 GFLOPs, while the measured throughput decreases from 333 to 190 FPS.

With SDL subsequently incorporated, recall increases from 35.8% to 38.6%, while mAP_{50} and $mAP_{50:95}$ increase to 38.8% and 22.4%, respectively. The complete MECA-Net configuration therefore improves mAP_{50} and $mAP_{50:95}$ by 6.0 and 3.2 percentage points, respectively, relative to the YOLO11n baseline. The final model contains 2.95 M parameters and requires 7.30 GFLOPs, corresponding to increases of 0.37 M parameters and 0.80 GFLOPs over the baseline. Throughput decreases from 340 to 158 FPS across the complete incremental sequence, with the largest measured reduction occurring when CADBA is added.

Because the modules are evaluated sequentially, the changes associated with CADBA and SDL represent their incremental contributions to the preceding configurations rather than their standalone effects. The present ablation therefore supports the cumulative effectiveness of the complete design but does not independently quantify module interactions.

Qualitative Analysis

To qualitatively compare MECA-Net with the YOLO11n baseline, representative daytime and nighttime traffic scenes were selected from the VisDrone2019 validation set, and the corresponding detection results are presented in Figs. 3 and 4.

As shown in Fig. 3(a), the overpass scene contains complex road geometry and several small vehicles distributed along curved ramps. Compared with YOLO11n, MECA-Net produces additional detections in several distant and densely populated regions, suggesting improved sensitivity to small objects under complex spatial layouts. In the elevated-road scene in Fig. 3(b), substantial scale variation is present between nearby and distant vehicles. MECA-Net yields a denser set of predictions for small distant vehicles, which is consistent with the use of the high-resolution 160×160 $P2$ prediction branch. In the urban arterial scene in Fig. 3(c), where vehicles are densely distributed around the intersection, the prediction patterns of the two models differ in several crowded regions, with MECA-Net producing more separated detections for adjacent objects. These qualitative observations are consistent with the intended roles of multi-scale boundary enhancement, directional coordinate encoding and high-resolution prediction. However, because the displayed panels do not include ground-truth overlays or matched error annotations, they should be interpreted as illustrative examples rather than as independent evidence of reduced false negatives or improved localisation accuracy.

In the nighttime intersection scene shown in Fig. 4(a), YOLO11n produces fewer predictions for several small and distant pedestrians, whereas MECA-Net produces additional predictions in these low-contrast regions. In the commercial-street scene in Fig. 4(b), strong illumination from headlights, street lamps and illuminated signs introduces substantial background interference. The two models exhibit different prediction patterns around these high-intensity regions, with MECA-Net producing a denser set of predictions for roadside traffic objects. In the wide multi-lane scene in Fig. 4(c), distant vehicles occupy only a small number of pixels and are densely distributed across the roadway. MECA-Net produces more predictions for small vehicles in several distant lanes, which is qualitatively consistent with the intended roles of multi-scale context modelling and the additional high-resolution $P2$ prediction branch.

Taken together with the daytime examples, these nighttime scenes illustrate differences in prediction behaviour between MECA-Net and YOLO11n under small object scales, dense traffic and complex illumination. The observations are consistent with the design motivation of combining multi-scale contextual information, explicit boundary cues, directional coordinate encoding and high-resolution prediction. Nevertheless, because the qualitative panels do not include ground-truth overlays, matched error annotations or statistical summaries, they should not be interpreted as independent evidence of improved recall, reduced false-positive rates or robustness across these conditions.

To further examine the spatial response patterns of YOLO11n and MECA-Net at the feature level, gradient-weighted class activation mapping (Grad-CAM) is applied to representative UAV traffic scenes, as shown in Fig. 5. The visualisations provide qualitative information on the image regions associated with the responses of the two detectors. Under the same visualisation

settings, warmer colours denote higher activation values, whereas cooler colours denote lower activation values.

To further examine the spatial response patterns of YOLO11n and MECA-Net at the feature level, gradient-weighted class activation mapping (Grad-CAM)[?] is applied to representative UAV traffic scenes, as shown in Fig. 5. The same target-layer selection and visualisation procedure are used for both models. The resulting maps provide qualitative information about the image regions associated with model responses. Warmer and cooler colours indicate relatively higher and lower activation values within the adopted visualisation procedure, respectively; activation magnitudes are not interpreted as directly comparable quantitative scores across models.

As illustrated in Fig. 5, YOLO11n exhibits relatively diffuse activation patterns in several scenes, with responses extending beyond traffic objects to structured background regions such as road markings, building boundaries, zebra crossings and shadows. These spatially distributed responses suggest that high-contrast background structures may contribute to the baseline feature representation, particularly in scenes containing small or densely distributed objects.

In comparison, MECA-Net exhibits a different spatial response distribution, with several high-activation regions appearing visually more concentrated around traffic-object regions and their surrounding structural context. Reduced activation in parts of the non-target background is also observed in some examples. These patterns are qualitatively consistent with the design motivations of MSEAA, which introduces multi-scale contextual and explicit gradient information, and CADBA, which incorporates directional coordinate information into feature recalibration. However, Grad-CAM reflects model sensitivity rather than localisation accuracy and cannot establish the causal contribution of an individual module. The visualisations should therefore be considered complementary to, rather than a substitute for, the quantitative results reported in Tables **??** and 2.

To examine the remaining failure modes of MECA-Net, representative challenging cases are analysed in Fig. 6. The examples illustrate three recurring sources of difficulty: background clutter and object–background similarity [Fig. 6(a)], severe occlusion [Fig. 6(b)], and adverse illumination combined with extremely small object scale [Fig. 6(c)]. These factors are not mutually exclusive and may occur simultaneously in UAV imagery. Because the examples are selected for qualitative illustration, the observations should not be interpreted as estimates of error frequency.

(a) Background clutter and object–background similarity. In Fig. 6(a), numerous small traffic objects are distributed across a visually complex urban environment containing vegetation, building structures and other high-contrast background patterns. When object boundaries have visual characteristics similar to those of surrounding structures, background gradients may compete with responses associated with true object boundaries. Although MSEAA introduces explicit gradient information, such structured background responses can still make small objects difficult to distinguish reliably. This example therefore highlights a limitation of the current edge-aware representation under cluttered and low-contrast object–background conditions.

(b) Severe occlusion. Figure 6(b) illustrates a scene in which pedestrians, non-motorised road users and vehicles are partially obscured by dense vegetation and neighbouring objects. Under severe occlusion, only a limited portion of the target remains visible, reducing the information available for classification and localisation. Multi-scale feature fusion can provide complementary contextual cues, but it cannot reconstruct target information that is absent from the visible image. Missed or uncertain detections may therefore remain, particularly when severe occlusion coincides with small object size.

(c) Adverse illumination and extremely small objects. The nighttime scene in Fig. 6(c) combines low illumination, strong local light sources and distant objects occupying only a few pixels. Reduced contrast weakens appearance and boundary cues, whereas illumination-induced edges may compete with gradients associated with true object boundaries. Extremely small objects also contain inherently limited spatial information. Although the additional stride-4 $P2$ prediction branch provides denser spatial sampling than the original $P3$–$P5$ pyramid, it cannot recover discriminative information that is absent from the input or substantially degraded during feature extraction. Detection therefore remains challenging when adverse illumination and extreme object scale occur simultaneously.

Overall, Fig. 6 indicates that MECA-Net remains challenged by background clutter, severe occlusion, adverse illumination and extremely small object scale. These observations suggest several directions for further improvement. Adaptive boundary modelling may help suppress responses to structured background gradients, while stronger context-aware feature fusion may improve representation under partial occlusion. For extremely small or poorly illuminated targets, further preservation of high-resolution features and temporal information from adjacent video frames may provide complementary cues when single-frame appearance information is insufficient.

Discussion

Fixed Sobel gradients provide explicit boundary cues but may also respond strongly to structured background edges, such as lane markings, shadows and building boundaries, while becoming less informative when object contours are weak or fragmented. Furthermore, the high-resolution $P2$ prediction branch increases spatial sampling density but cannot recover visual information that is absent from extremely small or heavily occluded targets. These observations suggest that further improvements may require more adaptive boundary modelling and stronger context-aware feature representations rather than simply increasing feature-map resolution.

Several additional limitations should be considered when interpreting the reported results. First, the quantitative results are based on a single training run with random seed 42; consequently, training variability has not been estimated. Repeated experiments with multiple random seeds are needed to report mean performance and dispersion and to determine whether the observed gains remain stable across training runs. Second, evaluation is restricted to the VisDrone2019-DET validation set, which primarily represents visible-light UAV imagery. Generalisation across datasets, sensors, geographic environments and flight conditions therefore remains to be established. Third, although MECA-Net achieves real-time throughput under the reported desktop-GPU setting, its deployment efficiency has not been evaluated on representative embedded UAV hardware.

Future studies should therefore evaluate MECA-Net on additional aerial benchmarks and representative edge platforms, investigate more adaptive mechanisms for low-contrast and occluded objects, and explore model compression or lightweight feature-recalibration designs to reduce the overall computational overhead. For UAV video applications, temporal modelling or multi-frame feature aggregation may also provide complementary information when single-frame visual cues are insufficient.

Conclusion

MECA-Net extends the YOLO11n baseline for small traffic-object detection in UAV imagery through three complementary components: MSEAA, CADBA and a stride-4 small-object detection layer (SDL). MSEAA combines multi-scale depthwise dilated convolutions with fixed Sobel-gradient features to model contextual and boundary information, while CADBA introduces directional coordinate recalibration to the fused multi-scale features. SDL extends the prediction pyramid from $P3$–$P5$ to $P2$–$P5$, providing an additional high-resolution prediction branch for small objects.

On the VisDrone2019-DET validation set, the complete model achieved 38.8% mAP_{50}, 22.4% $\text{mAP}_{50:95}$, 8.5% AP_S and 38.6% recall, with 2.95 M parameters and 7.30 GFLOPs. Under the reported NVIDIA RTX 4090 timing protocol, MECA-Net achieved a throughput of 158 FPS. Compared with the YOLO11n baseline, mAP_{50}, $\text{mAP}_{50:95}$ and AP_S increased by 6.0, 3.2 and 2.9 percentage points, respectively.

These results suggest that jointly incorporating multi-scale contextual information, explicit boundary cues, directional coordinate information and high-resolution prediction can improve small-object detection under the evaluated UAV traffic conditions while retaining real-time throughput under the reported desktop-GPU setting. However, the quantitative results are based on a single training run, and the current incremental ablation does not independently isolate the standalone contribution or interactions of all proposed components. Further evaluation across multiple random seeds, additional aerial datasets and representative embedded UAV platforms is therefore required to assess training stability, generalisation and deployment efficiency. Future work will also investigate more adaptive feature enhancement for low-contrast and occluded objects and lightweight model designs for reducing computational overhead.

Data availability

The VisDrone2019 dataset analysed in this study is publicly available from the VisDrone benchmark. The experimental results generated during the current study are included in this article. Additional information required to reproduce the reported experiments is available from the corresponding author upon reasonable request.

Code availability

The source code used to implement and evaluate MECA-Net is available from the corresponding author upon reasonable request.

References

1. Zhu, P. *et al.* Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 7380–7399, DOI: 10.1109/TPAMI.2021.3119563 (2022).
2. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, vol. 28, 91–99 (2015).
3. Liu, W. *et al.* SSD: Single shot multibox detector. In *Computer Vision – ECCV 2016*, vol. 9905 of *Lecture Notes in Computer Science*, 21–37, DOI: 10.1007/978-3-319-46448-0_2 (Springer, 2016).
4. Redmon, J., Divvala, S., Girshick, R. & Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788 (2016).
5. Redmon, J. & Farhadi, A. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767* DOI: 10.48550/arXiv.1804.02767 (2018).

6. Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* DOI: 10.48550/arXiv.2004.10934 (2020).

7. Jocher, G. YOLOv5 by ultralytics. Software, version 7.0, DOI: 10.5281/zenodo.3908559 (2022). Released 22 November 2022.

8. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430* DOI: 10.48550/arXiv.2107.08430 (2021).

9. Jocher, G., Chaurasia, A. & Qiu, J. Ultralytics YOLOv8. Software, version 8.0.0 (2023). Released 10 January 2023; accessed 6 August 2026.

10. Wang, C.-Y., Yeh, I.-H. & Liao, H.-Y. M. YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision – ECCV 2024*, vol. 15089 of *Lecture Notes in Computer Science*, 1–21, DOI: 10.1007/978-3-031-72751-1_1 (Springer, 2024).

11. Wang, A. *et al.* YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*, vol. 37, 107984–108011, DOI: 10.52202/079017-3429 (2024).

12. Jocher, G. & Qiu, J. Ultralytics YOLO11. Software, version 11.0.0 (2024). Released 10 September 2024; accessed 6 August 2026.

13. Khanam, R. & Hussain, M. YOLOv11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* DOI: 10.48550/arXiv.2410.17725 (2024).

14. Carion, N. *et al.* End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, vol. 12346 of *Lecture Notes in Computer Science*, 213–229, DOI: 10.1007/978-3-030-58452-8_13 (Springer, 2020).

15. Zhao, Y. *et al.* DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16965–16974 (2024).

16. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–2125 (2017).

17. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988 (2017).

18. Liu, S., Qi, L., Qin, H., Shi, J. & Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8759–8768 (2018).

19. Tan, M., Pang, R. & Le, Q. V. EfficientDet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790 (2020).

20. Han, Y. *et al.* LRDS-YOLO enhances small object detection in UAV aerial images with a lightweight and efficient design. *Sci. Reports* **15**, 22627, DOI: 10.1038/s41598-025-07021-6 (2025).

21. Zhou, S., Zhou, H. & Qian, L. A multi-scale small object detection algorithm SMA-YOLO for UAV remote sensing images. *Sci. Reports* **15**, 9255, DOI: 10.1038/s41598-025-92344-7 (2025).

22. Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018).

23. Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018*, 3–19, DOI: 10.1007/978-3-030-01234-2_1 (Springer, 2018).

24. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13713–13722 (2021).

25. Liu, H. & Su, T. Aero-LiteNet: Robust aerial small object detection via multi-scale fusion and neighborhood-aware optimization. *Sci. Reports* **16**, 17389, DOI: 10.1038/s41598-026-46022-x (2026).

26. Zhu, H. & Xie, X. Enhanced YOLOv11n for small object detection in UAV imagery: Higher accuracy with fewer parameters. *Sci. Reports* **16**, 5536, DOI: 10.1038/s41598-026-35301-2 (2026).

27. Wang, C. *et al.* CF-YOLO for small target detection in drone imagery based on YOLOv11 algorithm. *Sci. Reports* **15**, 16741, DOI: 10.1038/s41598-025-99634-0 (2025).

28. He, L.-h., Zhou, Y.-z., Liu, L., Cao, W. & Ma, J.-h. Research on object detection and recognition in remote sensing images based on YOLOv11. *Sci. Reports* **15**, 14032, DOI: 10.1038/s41598-025-96314-x (2025).

29. Cai, Z. & Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162 (2018).

30. Tian, Z., Shen, C., Chen, H. & He, T. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9627–9636 (2019).

31. Zhou, X., Wang, D. & Krähenbühl, P. Objects as points. *arXiv preprint arXiv:1904.07850* DOI: 10.48550/arXiv.1904.07850 (2019).

32. Rezatofighi, H. *et al.* Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 658–666 (2019).

33. Chen, D. & Zhang, L. SL-YOLO: A stronger and lighter drone target detection model. *arXiv preprint arXiv:2411.11477* DOI: 10.48550/arXiv.2411.11477 (2024).

34. Lu, L., He, D., Liu, C. & Deng, Z. MASF-YOLO: An improved YOLOv11 network for small object detection on drone view. *arXiv preprint arXiv:2504.18136* DOI: 10.48550/arXiv.2504.18136 (2025).

35. Liu, J. & Wang, Y. MSG-YOLO: A multi-scale statistical-guided network for drone object detection. *Appl. Sci.* **16**, 6050, DOI: 10.3390/app16126050 (2026).

36. Li, H. *et al.* RLRD-YOLO: An improved YOLOv8 algorithm for small object detection from an unmanned aerial vehicle (UAV) perspective. *Drones* **9**, 293, DOI: 10.3390/drones9040293 (2025).

Acknowledgements

This work was supported by the Science and Technology Tackling Project of Henan Province under Grant Nos. 252102211052 and 262102211047.

Author contributions statement

L.W. conceived the study, developed the proposed method, conducted the experiments, analysed the results and prepared the original manuscript. M.H. contributed to software implementation, experimental validation and data analysis. Y.C. contributed to data preparation and experimental validation. H.L. contributed to result interpretation and manuscript revision. D.Z. supervised the study, contributed to the methodological design, acquired funding and reviewed the manuscript. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Dongyao Zou (zdy@zzuli.edu.cn).

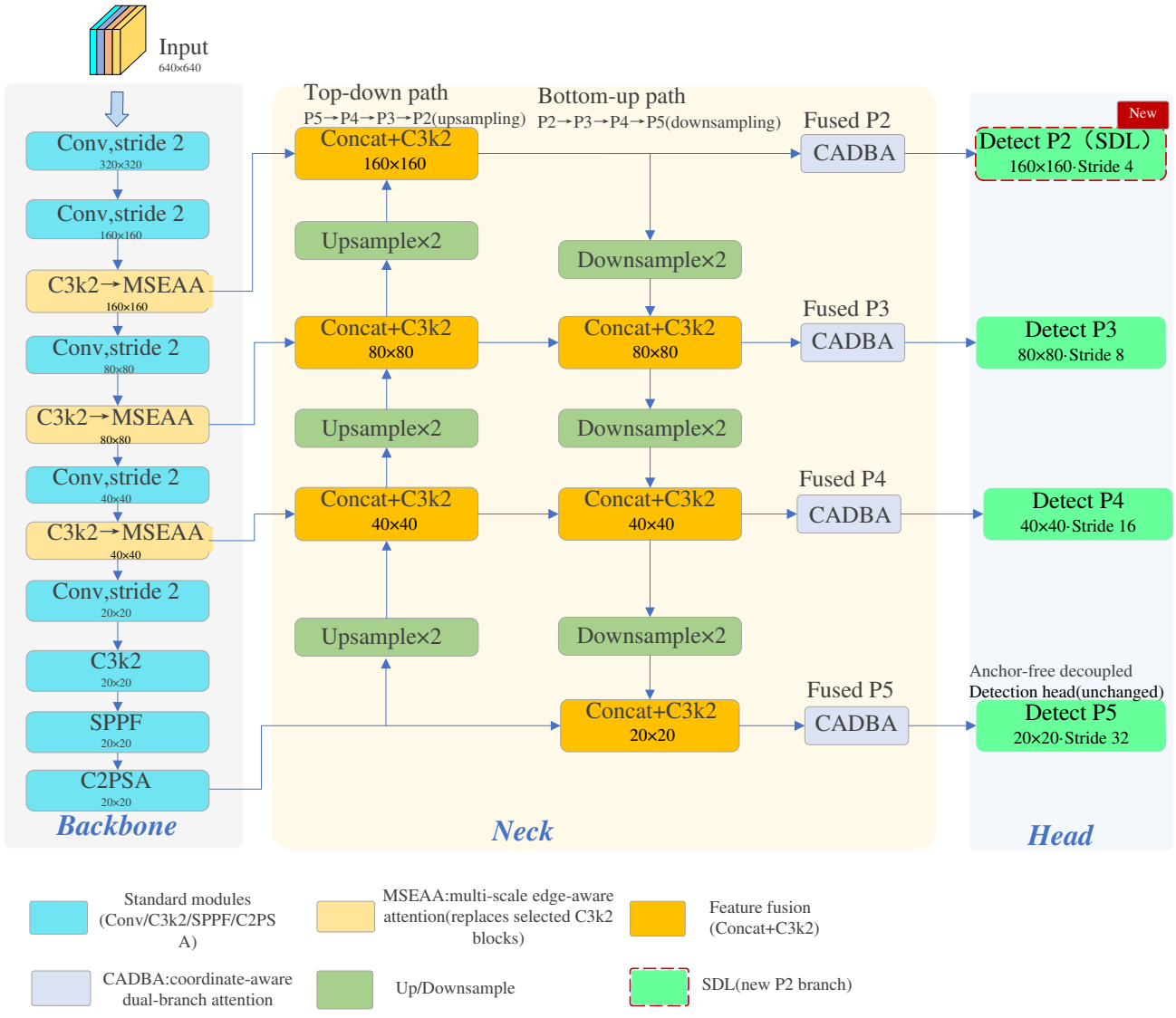
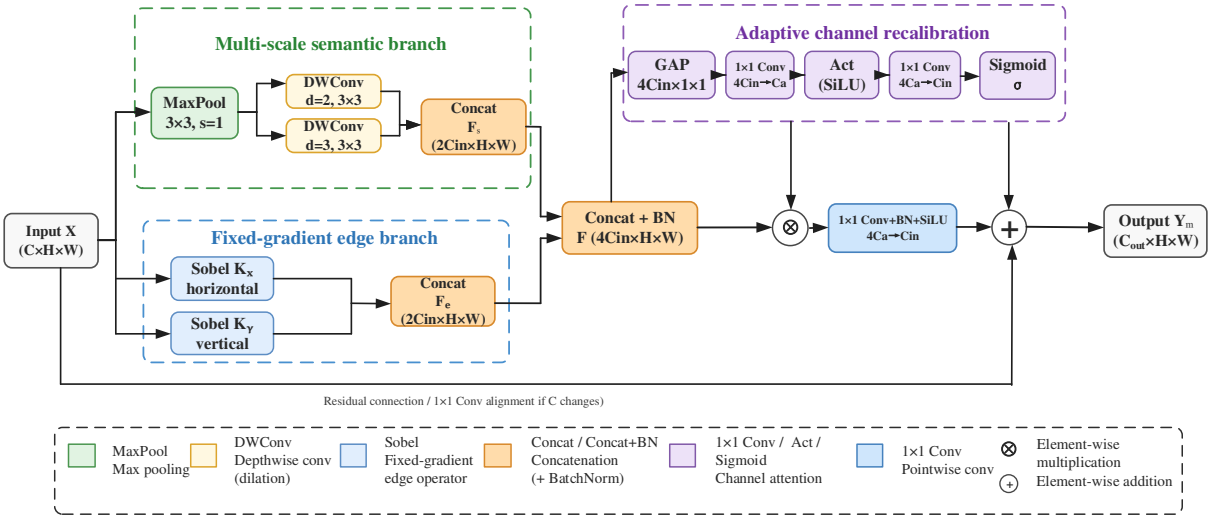
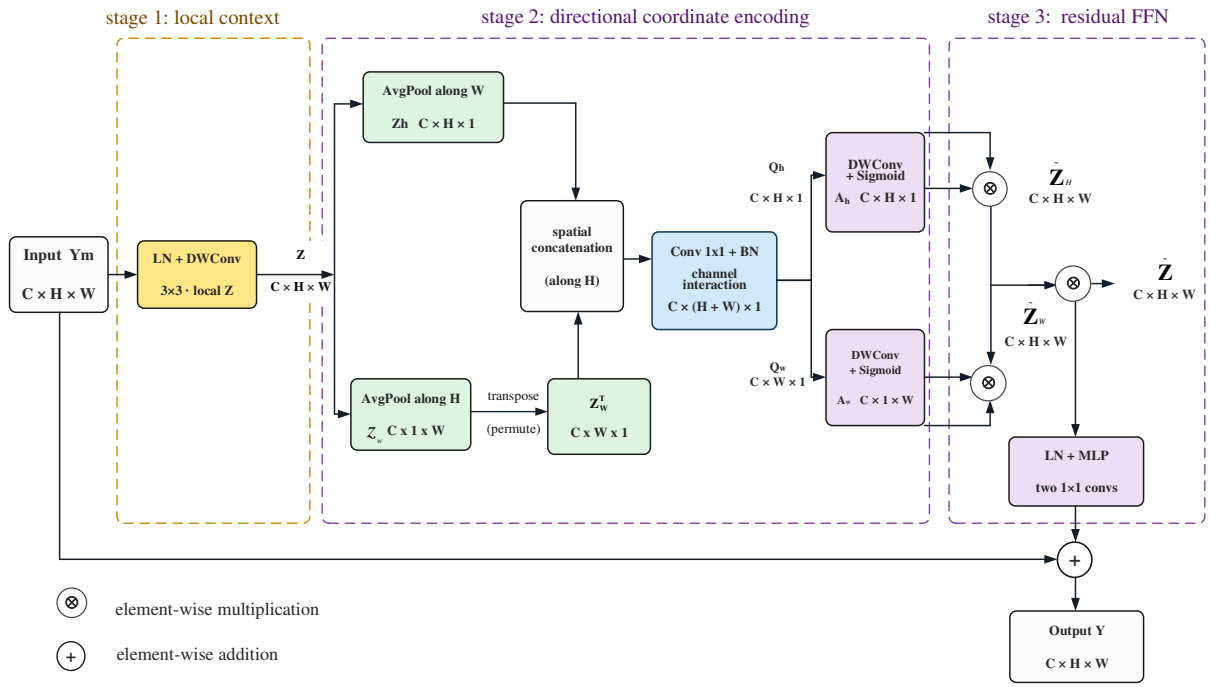


Figure 1. Overall architecture of MECA-Net. Three selected C3k2 blocks associated with the *P*2, *P*3 and *P*4 backbone stages are replaced by the proposed multi-scale edge-aware attention module (MSEAA), which integrates multi-scale semantic context and fixed-gradient boundary information. CADBA recalibrates each fused *P*2–*P*5 neck feature using height- and width-aware attention. SDL extends the original *P*3–*P*5 prediction pyramid by introducing an additional stride-4 *P*2 detection branch.



(a) multi-scale edge-aware attention module (MSEAA)



(b) coordinate-aware dual-branch attention module(CADBA)

Figure 2. Internal structures of MSEAA and CADBA. (a) MSEAA replaces selected C3k2 blocks and combines multi-scale contextual features with fixed Sobel-gradient information and channel recalibration. (b) CADBA recalibrates fused features using local-context modelling, directional coordinate attention and a residual feed-forward block.

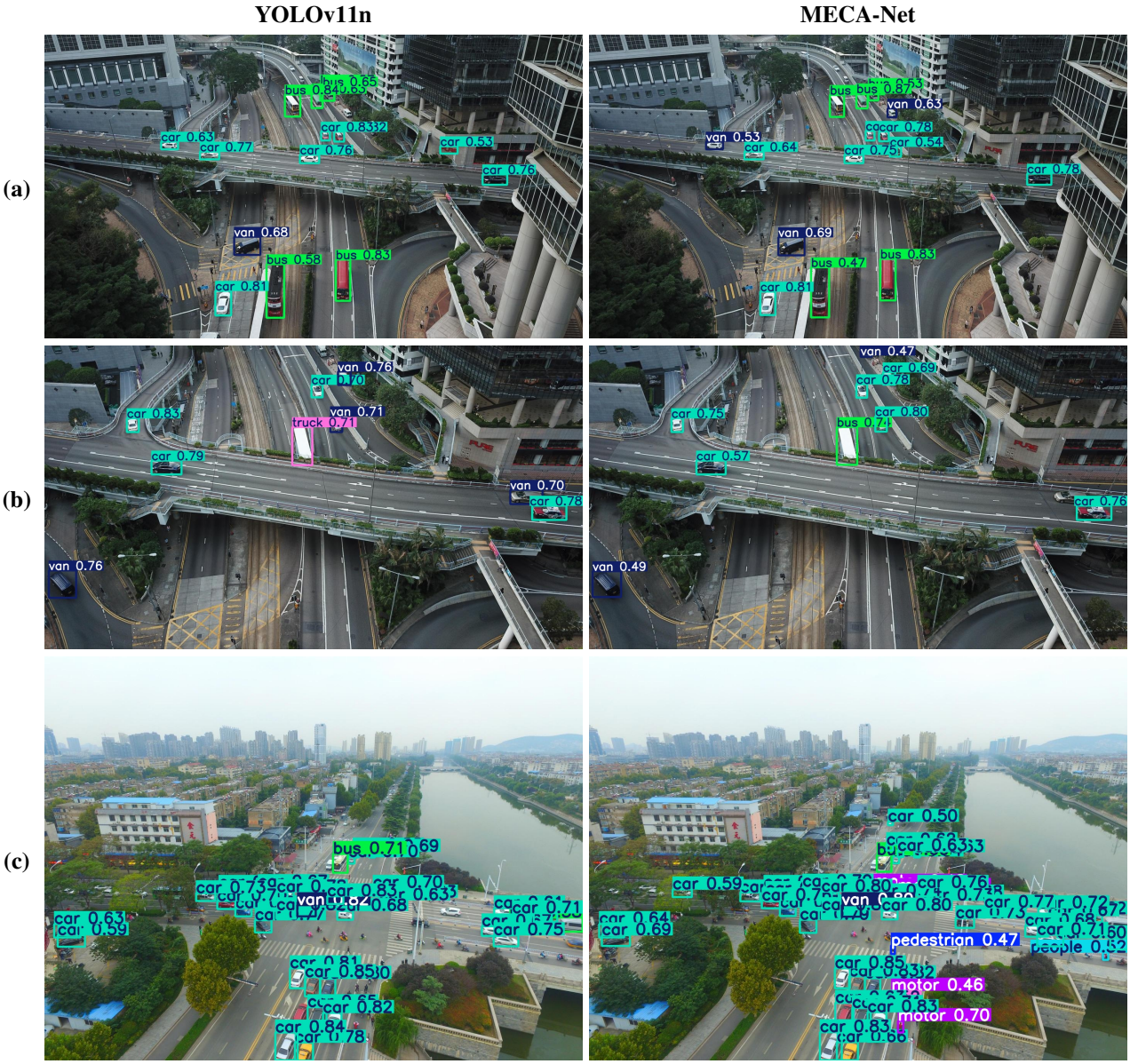


Figure 3. Qualitative comparison of detection results obtained by YOLO11n and MECA-Net on daytime urban traffic scenes from the VisDrone2019 validation set. (a) Bird's-eye view of an urban overpass; (b) elevated road scene; and (c) bird's-eye view of an urban arterial road. The left and right columns show the results of YOLO11n and MECA-Net, respectively.

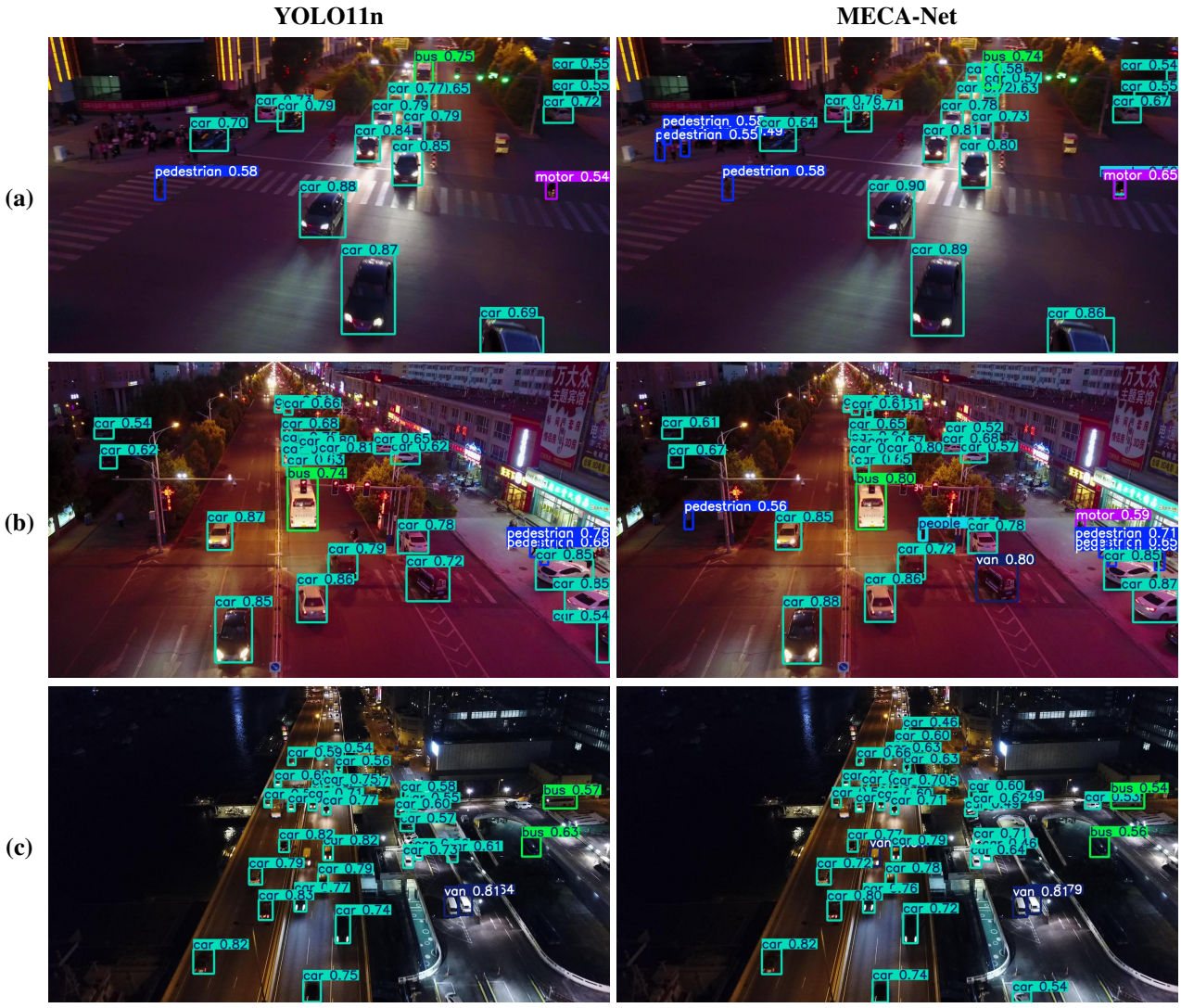


Figure 4. Qualitative comparison of detection results obtained by YOLO11n and MECA-Net on nighttime traffic scenes from the VisDrone2019-DET validation set. (a) Vehicles and pedestrians at an intersection; (b) nighttime traffic flow on a commercial street; and (c) a nighttime multi-lane road. The left and right columns present the results obtained by YOLO11n and MECA-Net, respectively.

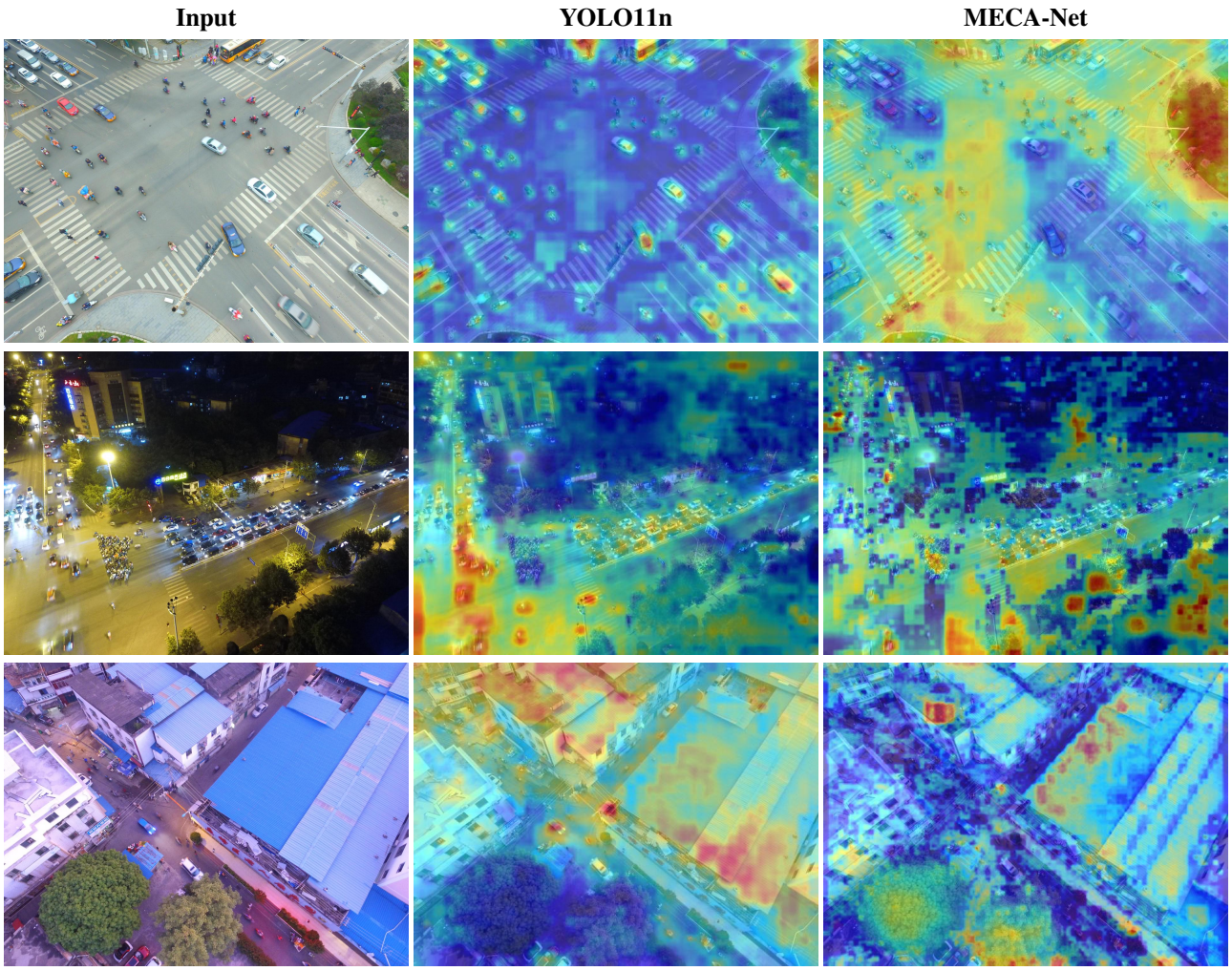


Figure 5. Grad-CAM visualisations for representative UAV traffic scenes. The columns show the input images, YOLO11n activation maps and MECA-Net activation maps, respectively. Warmer and cooler colours denote relatively higher and lower activation values, respectively, under the adopted visualisation procedure. The maps areused to compare spatial response patterns qualitatively and should not be interpreted as quantitative measurements of detection performance.

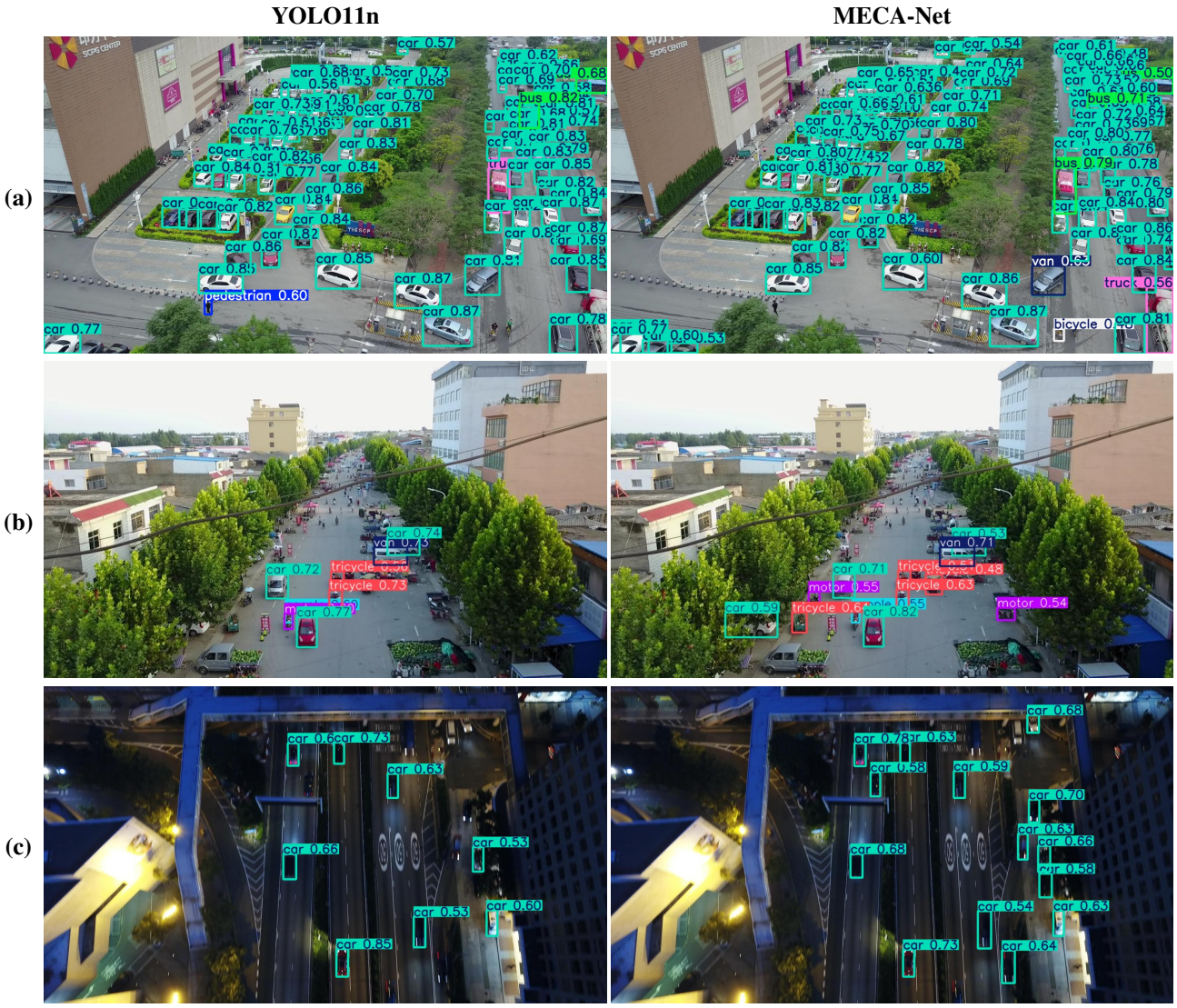


Figure 6. Representative challenging cases of YOLO11n and MECA-Net on the VisDrone2019 validation set. The left and right image columns show predictions produced by YOLO11n and MECA-Net, respectively. **(a)** A cluttered daytime scene containing small objects and complex background structures; **(b)** a scene with severe vegetation-induced occlusion; and **(c)** a low-illumination nighttime scene containing distant and extremely small objects. The examples illustrate representative failure conditions and are intended for qualitative analysis rather than statistical comparison of error rates.