Enhanced YOLOv10 for Small-Object Rebar Detection in UAV Images on Construction Sites

Seunghyeon Wang[1] Sungkon Moon[2] Yuanzhe He[1] Rong-Lu Hong[1],

Ke-Ting Pan[3], and Ju-Hyung Kim[1]

[1] Department of Architectural Engineering, Hanyang University, Seoul, Republic of Korea
[2] Department of Civil Systems Engineering, Ajou University, Suwon, Republic of Korea
[3] Graduate Institute of Aerospace and Undersea Medicine, College of Biomedical Sciences, National Defense Medical University, Taipei, Taiwan
wsh1019@hanyang.ac.kr, skmoon@ajou.ac.kr, heyuanzhe1025@hanyang.ac.kr, abraham0814@hanyang.ac.kr, ke-ting.pan@mail.ndmctsgh.edu.tw, kcr97jhk@hanyang.ac.kr

Abstract

Unmanned Aerial Vehicles (UAV) images can streamline reinforced-concrete inspections, yet reliable automatic rebar counting is still challenging because rebars often occupy only a few pixels, appear in dense groups, and are frequently obscured or visually blended with background textures, shadows, and occlusions. To address this small-object setting, we develop an improved You Only Look Once (YOLO)v10-based detector that incorporates Omni-Dimensional Dynamic Convolution (ODConv) into the backbone, an Efficient Multi-scale Attention–guided Bidirectional Feature Pyramid Network (EMA-BiFPN) to strengthen multi-scale feature aggregation, a four-scale prediction design that adds an extra high-resolution detection head, and a Minimum Points Distance IoU (MPDIoU) regression loss to stabilize bounding-box learning under stringent IoU criteria. Experiments on a dataset collected from seven construction sites (1,328 images) show that the proposed approach achieves AP50 = 94.72 and AP50:95 = 71.86 at 35.93 FPS, outperforming the YOLOv10 baseline and delivering clearer gains for extremely small rebar instances. These results demonstrate the potential of the proposed framework as a practical offline analysis tool for UAV-based construction inspection, while also providing a foundation for future real-time deployment in field workflows.

Keywords –

Rebar inspection, Unmanned aerial vehicle, You only look once version 10, Omni-dimensional dynamic convolution

Introduction

Reinforced Concrete (RC) columns—structural members in which steel reinforcing bars (rebars) are embedded within concrete—are widely used as primary load-bearing elements in modern buildings. They carry gravity loads from slabs, floors, and roofs and transfer these actions through the structural system to the foundation [1,2]. While concrete provides high compressive and shear strength, its tensile resistance is limited; rebars are therefore required to carry tensile stresses and enhance the overall structural capacity and ductility.

Because the amount and arrangement of rebars directly affect safety and flexural performance, reinforcement detailing in RC columns must be carefully verified. Under-reinforcement can compromise load capacity and serviceability, whereas over-reinforcement increases material and labor costs and can create constructability issues. For this reason, design codes prescribe minimum and maximum reinforcement ratios, along with material and detailing requirements for both RC members and reinforcing steel [3]. Rebar verification is increasingly supported by Unmanned Aerial Vehicles (UAVs) equipped with Red–Green–Blue (RGB) cameras, which can acquire high-resolution images from viewpoints that are difficult or unsafe for inspectors to reach [4]. However, the interpretation of these images is still often performed manually, limiting throughput for frequent inspections and large-scale monitoring.

Deep learning–based object detection offers a practical way to automate rebar localization and counting from UAV imagery. In general, detectors identify objects by predicting class labels and regressing bounding boxes for instances of interest [5]. After training, such models can localize individual rebars and enable automated

counting with improved efficiency and repeatability. Although strong rebar-counting performance has been reported in more controlled settings (e.g., production lines, warehouses, and storage yards) [6–12], UAV images of RC columns introduce additional difficulties [2,4,13]. Rebars frequently appear as extremely small targets, occur in dense and repetitive patterns, and are captured under oblique viewpoints with shadows, occlusions, and cluttered backgrounds. In these conditions, a single bar may occupy only a few pixels, and closely spaced bars can be confused or merged, leading to missed detections and unstable counting—particularly when models fail to preserve high-resolution features or lack mechanisms suited to tiny, overlapping objects.

To address these challenges, this study improves small-object rebar detection in UAV imagery using a You Only Look Once (YOLO)v10-based framework. The main contributions are:

1) Omni-Dimensional Dynamic Convolution (ODConv) is integrated into the YOLOv10 backbone to strengthen local feature extraction for tiny rebars in complex scenes.

2) An Efficient Multi-scale Attention–guided Bidirectional Feature Pyramid Network (EMA-BiFPN) is introduced to aggregate information across scales while retaining fine-grained spatial cues.

3) Minimum Points Distance Intersection over Union (MPDIoU) is adopted as the bounding-box regression loss to improve fitting robustness for small and heavily overlapped rebars.

4) A four-head multi-scale detection design is investigated to increase sensitivity to extremely small rebars and improve generalization across object sizes.

In practical construction settings, the proposed framework can support site inspection workflows by enabling automatic identification of rebars from UAV imagery for progress monitoring, quality assessment, and site documentation. In the current study, the framework is evaluated in an offline analysis setting, where UAV images are first collected and then processed by the trained detection model. This setup establishes the technical feasibility of the method, while future extensions may consider real-time deployment through onboard inference or live transmission to an external server.

2. Proposed methodology

2.1 Overview of YOLOv10 model

YOLOv10 is a recently proposed single-stage detector that supports end-to-end inference without Non-Maximum Suppression (NMS) [14]. Its architecture is organized into three main parts: backbone, neck, and head. The backbone builds multi-level feature representations from the input image using building blocks such as CBS, SCDown, C2f, C2fCIB, and PSA-based self-attention. Next, the Neck performs multi-scale feature aggregation via an FPN–PAN fusion pipeline, merging high-resolution cues from early layers with deeper semantic information to preserve fine details. Finally, the head uses a lightweight decoupled prediction design together with a consistent dual-assignment scheme.

2.2. Architectural enhancements of YOLOv10

An improved YOLOv10 variant is developed (Figure 1), with its primary architectural changes from the original model summarized in Figure 2.

The proposed design includes four major upgrades. First, the standard C2f module is augmented with ODConv to improve adaptive feature extraction, which is particularly beneficial for tiny targets. Second, EMA-BiFPN replaces/extends the original fusion scheme; by applying multi-scale attention during feature aggregation, it better preserves small-object cues while limiting unnecessary computation. Third, MPDIoU is used for bounding-box regression to strengthen localization when overlap is limited and object shapes are ambiguous, improving both precision and recall. Finally, an extra prediction head is added to better handle very small rebars, yielding four detection heads dedicated to extremely small, small, medium, and large objects. Together, these modifications expand the effective scale coverage and improve robustness across diverse rebar sizes.

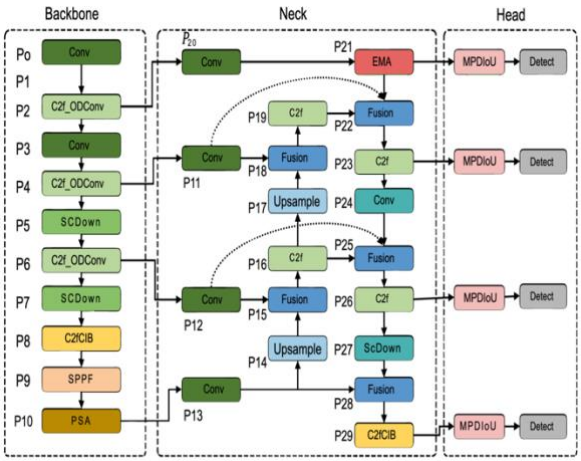


Figure 1. Architecture of modified YOLOv10

2.2.1. Enhanced backbone for local features

In UAV imagery of construction sites, rebars are typically thin, small, and densely clustered, and their visual appearance can resemble surrounding textures,

which makes precise localization difficult. Although the original C2f module in YOLOv10 is effective, its representational flexibility can be limited when convolutional adaptation occurs along only a subset of possible dimensions. This limitation can reduce feature separability when object scale and background complexity vary widely.

To address this issue, ODConv [15] is incorporated into the C2f module. ODConv introduces dynamic modulation across four dimensions—spatial positions, input channels, output channels, and kernel indices—which enables fine-grained, content-adaptive filtering. Compared with earlier dynamic convolution variants such as CondConv and DyConv, which typically modulate fewer components of the convolution process [16], ODConv provides richer adaptation while remaining computationally feasible.

The ODConv operation can be written as:

$$y = (a_{w1}W_1 + \cdots + a_{wn}W_n) * x \quad (1)$$

where x and y are the input and output feature maps, W_n is the n -th kernel, and a_{wn} is its corresponding weighting coefficient. With omni-dimensional attention, each kernel is further modulated by multiple attention terms:

$$y = \begin{pmatrix} a_{w1} \odot a_{f1} \odot a_{c1} \odot a_{s1} \odot W_1 + \cdots \\ + a_{wn} \odot a_{fn} \odot a_{cn} \odot a_{sn} \odot W_n \end{pmatrix} * x \quad (2)$$

where $\odot$ denotes element-wise multiplication, and a_{fn}, a_{cn}, and a_{sn} represent attention factors associated with spatial- and channel-related modulation. This omni-dimensional control improves discrimination for tiny rebars that are easily confused with background patterns, while maintaining efficiency suitable for real-time inference.

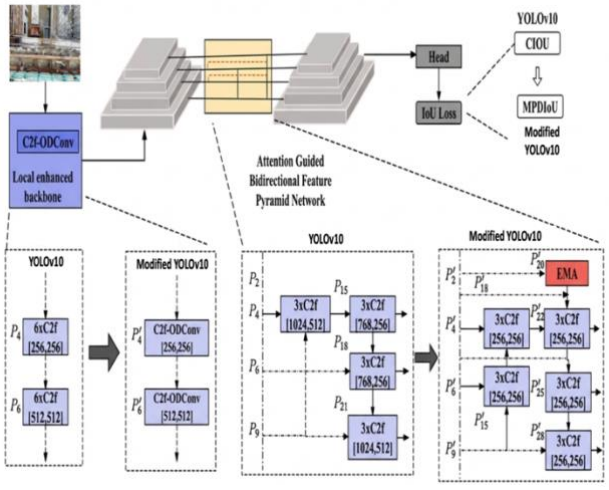


Figure 2. Comparison of original YOLOv10 (baseline), and modified YOLOv10

2.2.2. Attention guided bidirectional feature pyramid network

Accurate rebar detection requires strong feature fusion across scales, because rebars can appear at very different sizes depending on camera distance, view angle, and illumination. However, the PANet-style fusion used in YOLOv10 largely relies on relatively shallow bidirectional aggregation, and repeated convolution-based merging can weaken high-level semantic cues—an effect that is particularly harmful when detecting extremely small objects.

To improve cross-scale integration, an EMA-BiFPN is designed by incorporating Efficient Multi-scale Attention (EMA) [16]. EMA refines features by modeling both spatial and channel relationships, helping the network emphasize small-object regions. In the proposed neck, a top-down pathway propagates semantic context to guide the fusion of low-level details, while skip connections between intermediate pyramid nodes help preserve informative features and strengthen gradient flow.

Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, EMA splits channels into G groups X_g to reduce cost. Two lightweight branches (e.g., 1×1 and 3×3 convolutions) extract complementary context. Channel attention terms are produced using global average pooling:

$$G_{avg}^{1\times1} = \mathrm{GAP}\left(X_g^{1\times1}\right), \qquad G_{avg}^{3\times3} = \mathrm{GAP}(X_g^{3\times3}) \quad (3)$$

and the group-wise refined output is computed as:

$$X_g^{EMA} = G_{avg}^{1\times1} \odot X_g^{1\times1} + G_{avg}^{3\times3} \odot X_g^{3\times3} \quad (4)$$

All X_g^{EMA} are then concatenated along the channel dimension to form the final refined representation. This attention-guided fusion improves multi-scale context modeling while keeping computation lightweight, which is advantageous for dense small-object detection in complex scenes.

2.2.3 Loss function with MPDIoU

Bounding box regression performance depends strongly on the loss function. For small objects, IoU-based losses can be unstable because small coordinate shifts may sharply reduce IoU, weakening the learning signal for accurate localization. YOLOv10 commonly employs Complete IoU (CIoU) [17], which extends IoU by adding penalties for center distance and aspect-ratio mismatch:

$$L_{CIOU} = 1 - \mathrm{IoU}\left(B_{gt}, B_{prd}\right) + \frac{\rho^2\left(c_{gt}, c_{prd}\right)}{c^2} + \alpha v \quad (5)$$

where $\mathrm{IoU}\left(B_{gt}, B_{prd}\right) = \frac{B_{gt} \cap B_{prd}}{B_{gt} \cup B_{prd}}$, $\rho\left(c_{gt}, c_{prd}\right)$ is the Euclidean distance between box centers, and $c^2 = w_c^2 + h_c^2$ is the squared diagonal length of the smallest enclosing box covering both boxes. The aspect-ratio term is:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w_{prd}}{h_{prd}}\right)^2,$$
$$+\alpha = \frac{v}{\left(1 - \mathrm{IoU}(B_{gt}, B_{prd})\right) + v} \quad (6)$$

Although CIoU improves convergence, it can be less sensitive when predicted and ground-truth boxes have similar aspect ratios but differ in scale, which is common for thin objects such as rebars.

To strengthen localization for tiny, densely distributed targets, MPDIoU (Minimum Points Distance IoU) [18] is adopted. MPDIoU introduces additional constraints based on distances between corresponding corner points. For two shapes A and B, the squared distances for two paired corners are:

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \quad (7)$$
$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \quad (8)$$

and MPDIoU is computed as:

$$\mathrm{MPDIoU} = \frac{A \cap B}{A \cup B} - \frac{w^2 + h^2}{d_1^2 + w^2 + h^2} - \frac{w^2 + h^2}{d_2^2 + w^2 + h^2} \quad (9)$$

where w and h are the image width and height. By explicitly penalizing corner mismatches, MPDIoU provides a stronger regression signal for small objects, improving box positioning and scale estimation under crowded and low-overlap conditions.

2.2.4 Multi-prediction head

To improve coverage of tiny rebars, the detector is extended to a four-head, multi-scale prediction scheme. Specifically, an additional high-resolution pyramid level is introduced by leveraging shallower Neck features that retain richer local details. This feature level is refined through the bidirectional fusion pathway and connected to an extra detection head. As a result, the model produces predictions at four scales, designed to cover extremely small, small, medium, and large objects, respectively.

During training, the total objective is computed by summing the losses from all four heads:

$$L_{Total} = \sum_{s=1}^{4} (L_{Cls}^{(s)} + L_{Obj}^{(s)} + \lambda L_{Reg}^{(s)}) \quad (10)$$

where $s \in \{1, 2, 3, 4\}$ indexes the prediction head (from the highest-resolution head for extremely small targets to the lowest-resolution head for large targets), $L_{Cls}^{(s)}$ and $L_{Obj}^{(s)}$ denote the classification and objectness losses for head s, and λ balances the regression term. For bounding box regression, MPDIoU is used as:

$$L_{Reg}^{(s)} = 1 - \mathrm{MPDIoU}(B_{gt}^{(s)}, B_{prd}^{(s)}) \quad (11)$$

where $B_{gt}^{(s)}$ and $B_{prd}^{(s)}$ represent the matched ground-truth and predicted bounding boxes assigned to head s. By introducing a dedicated high-resolution head and supervising it with MPDIoU-based regression, the model receives a stronger learning signal for extremely small rebars, which reduces missed detections in crowded scenes while maintaining robust performance across diverse object scales.

2.3 Evaluation metrics

To explicitly assess small-object detection performance, Average Precision (AP)50 and AP50:95 are reported. In addition, size-specific AP is also provided. All images are resized to 640 × 640 using aspect-ratio–preserving letterbox resizing, and bounding box sizes are measured in the resized coordinate system. For each ground-truth rebar box with width w and height h (in pixels), the size indicator is defined as:

$$s = \min(w, h) \quad (11)$$

Each instance is then assigned to one of four size bins:

- Extremely Small (ES): $s < 8$ pixel
- Small (S): $8 \leq s < 16$ pixel
- Medium (M): $16 \leq s < 32$ pixel
- Large (L): $s \geq 32$ pixel

3. Experimental setting

3.1 Data construction

Rebar images were collected using a DJI Phantom 4 Pro at seven active construction sites in South Korea. The UAV was manually flown to capture high-resolution RGB images from a near-vertical viewpoint at approximately 1–2 m above the rebar layout. RC columns were positioned near the image center, and data were acquired during working hours to reflect realistic site conditions, including changes in illumination, background clutter, and scene complexity. In total, 1,328 images were collected and resized to 640×640 pixels. The dataset was randomly split into training/validation/test sets using a 60/20/20 ratio (796/266/266 images).

Because augmentation outcomes depend on parameter settings, augmentation ranges were selected through iterative trial-and-error to maintain visually plausible samples. Eight single-augmentation training sets were produced from the 796 training images—brightness, contrast, scaling, perspective, rotation, translation, shearing, and Gaussian blur (Table 1). A "sum-of-techniques" setting was also constructed by combining the original training set with all augmented versions, resulting in 7,164 training images (796 × 9).

Bounding boxes were manually annotated using LabelImg for each visible rebar. All instances were assigned a single class label ("rebar"), and all other regions were treated as background. Labels were stored

in PASCAL VOC XML format. The number of annotated rebars in each subset is reported in Table 1.

3.2 Comparative method

Within a unified YOLOv10 framework, multiple backbones were evaluated, including CNN-based designs (CSPNet, ConvNeXt, EfficientNet) and transformer-based models (Swin Transformer, ViT, PVT, MobileViT, and Axial Transformer). For fair comparison, all models were trained under the same hyperparameter search space. For each architecture, 162 configurations were explored, and the best-performing setting was selected. Hyperparameters included batch size {1,2,4}, learning rate {0.00025, 0.0005, 0.001}, weight decay {0.0001, 0.0003, 0.0005}, momentum {0.5, 0.7, 0.9}, and maximum iterations {20k, 30k}.

All detectors were initialized from COCO-pretrained weights and fine-tuned on the rebar dataset to accelerate convergence and improve robustness for small, densely clustered targets under complex backgrounds.

Table 1. Augmentation settings and dataset statistics

Configuration	Range of parameter	Images	Rebars
Original (training)	–	796	19,576
Brightness	[-50, 50]	796	19,576
Contrast	[0.5, 3.0]	796	19,576
Scale	[x: 0.8, 1.5], [y: 0.8, 1.5]	796	19,576
Perspective	[0.01, 0.18]	796	19,576
Rotation	[−35°, 35°]	796	19,576
Translation	[x: -0.3, 0.3], [y: -0.3, 0.3]	796	19,576
Shearing	[− 20°, 20°]	796	19,576
Blurring	[0.5, 1.5]	796	19,576
Validation		266	6,812
Test	–	266	6,737

4. Results and discussion

4.1 Sensitivity to weight initialization

To quantify stochastic variation due to random initialization (and the resulting optimization trajectory), the best hyperparameter configuration for each architecture was retrained using five different random seeds. Overall, seed-to-seed variance is small across all models. Standard deviations remain low for both AP50 (approximately 0.15–0.34) and AP50:95 (approximately 0.16–0.41), indicating that the observed differences

Table 3. Model performance by image conditions

not primarily driven by randomness and that model rankings are largely consistent across runs (Table 2).

The proposed method achieves the best mean performance while maintaining stable results across seeds (AP50 = 94.52±0.22; AP50:95 = 71.52±0.21), suggesting that the gains are reproducible rather than dependent on a single favorable run. Among the baselines, Swin Transformer is the strongest competitor and also shows high stability (AP50 = 94.02±0.17; AP50:95 = 70.64±0.18), whereas PVT remains competitive but trails both Swin Transformer and the proposed method on AP50 and AP50:95.

Notably, the improvement is larger for AP50:95 than for AP50. Relative to Swin Transformer, the proposed approach increases AP50 by about 0.50 points (94.52 vs. 94.02) and improves AP50:95 by about 0.88 points (71.52 vs. 70.64), implying a stronger benefit for localization quality under stricter IoU thresholds.

Table 2. Results across random initializations (μ±σ)

Random factor	Architecture	AP50	AP50:95
Weight initialization	CSPNet	92.36±0.16	69.61±0.17
	ConvNeXt	92.53±0.18	70.22±0.19
	EfficientNet	92.29±0.15	69.08±0.18
	Swin Transformer	94.02±0.17	70.64±0.18
	ViT	89.19±0.34	62.81±0.41
	PVT	92.55±0.19	70.10±0.17
	MobileViT	89.77±0.22	64.71±0.16
	Axial Transformer	92.14±0.21	69.24±0.22
	Proposed	94.52±0.22	71.52±0.21
	ConvNeXt	92.34±0.48	69.87±0.67
	EfficientNet	92.04±0.44	68.79±0.58
	Swin Transformer	93.95±0.45	71.10±0.56
	ViT	88.86±0.69	62.30±0.88
	PVT	92.45±0.51	70.65±0.61
	MobileViT	89.43±0.57	64.33±0.69
	Axial Transformer	91.93±0.50	68.96±0.64
	Proposed method	94.45±0.41	72.30±0.49

4.2 Impact of image conditions

Table 3 reports performance under different image-condition groups, created by partitioning the test set into two subsets per factor (examples shown in Figure 3). Because AP is computed from the full precision–recall curve, it is a dataset-level metric and is not additive across subsets; therefore, subgroup AP values are not expected to average to the overall test AP.

Performance consistently drops under more

Image condition	Range	AP50	AP50:95	AP_{ES}	AP_S	AP_M	AP_L
Illumination	High	95.82	73.41	89.64	93.86	97.12	97.84
	Low	94.21	71.32	86.92	92.31	96.53	97.41
Visual range	Near-field view	96.43	74.26	91.12	94.71	97.38	97.96
	Far-field view	93.02	69.44	83.41	90.52	95.87	97.12
Similar object	Absent	95.74	73.24	89.76	93.93	97.15	97.86
	Present	94.08	71.12	86.31	92.05	96.42	97.35
Rebar overlap	Non-overlap	95.94	73.76	89.98	94.09	97.19	97.88
	Overlap	94.11	71.03	86.44	92.12	96.41	97.33
Shadow	Absent	95.63	73.12	89.22	93.58	97.03	97.79
	Present	94.36	71.48	86.91	92.37	96.62	97.46
Number of rebars	> 25	94.18	71.14	86.73	92.09	96.48	97.39
	≤ 25	95.88	73.69	90.05	94.1	97.22	97.91

challenging conditions. Across all six factors—low illumination, far-field views, presence of similar objects, rebar overlap, shadows, and scenes with more than 25 rebars—the "hard" subset yields lower AP50 and AP50:95 than the corresponding "easier" subset. The degradation is more pronounced for AP50:95, suggesting that difficult conditions primarily harm localization precision (tight box fitting) rather than only coarse detection.

The most influential factor is visual range. Moving from near-field to far-field imagery causes the largest decline (AP50: 96.43→93.02, −3.41; AP50:95: 74.26→69.44, −4.82). This reduction is driven mainly by the smallest objects: AP_ES drops from 91.12 to 83.41 (−7.71), far exceeding the decreases for medium and large objects (AP_M −1.51; AP_L −0.84). This indicates that the dominant failure mode in far-field scenes is reduced small-object detectability due to weaker pixel evidence, which also explains the sharper drop in AP50:95.

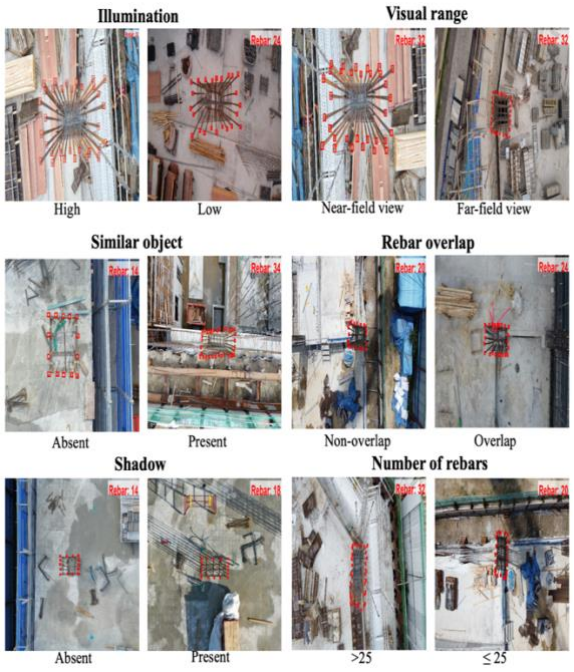
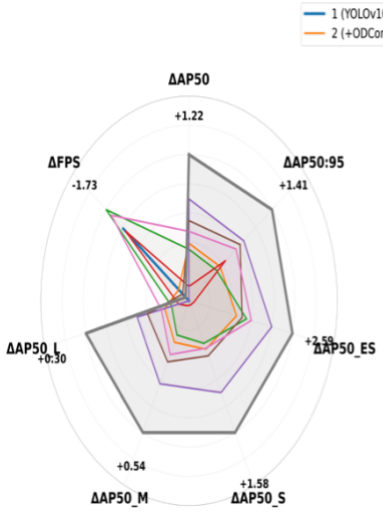


Figure 3. Examples of detected results by image conditions

4.3 Ablation results

Figure 4 summarizes the ablation of ODConv, EMA-BiFPN, and MPDIoU using both relative changes (radar chart) and absolute values (table). Using the YOLOv10 baseline as reference (AP50 = 93.50; AP50:95 = 70.45;



Variant	ODConv	EMA-BiFPN	MPDIoU	AP50	AP50:95	AP50_ES	AP50_S	AP50_M	AP50_L	FPS
1	-	-	-	93.50	70.45	84.64	90.60	95.87	97.05	37.66
2	✓	-	-	93.98	70.98	85.83	91.18	96.04	97.12	36.12
3	-	✓	-	93.93	70.92	86.08	91.11	96.01	97.10	38.11
4	-	-	✓	93.62	71.07	84.76	90.65	95.89	97.08	37.58
5	✓	✓	-	94.35	71.38	86.71	91.70	96.21	97.20	35.84
6	✓	-	✓	94.17	71.32	85.97	91.26	96.12	97.17	36.01
7	-	✓	✓	94.08	71.25	86.20	91.17	96.09	97.13	37.97
8	✓	✓	✓	94.72	71.86	87.23	92.18	96.41	97.35	35.93

Figure 4. Ablation results of proposed method

37.66 FPS), each component provides a distinct benefit.

ODConv improves overall accuracy (AP50 = 93.98; AP50:95 = 70.98) and strengthens small-object metrics (AP_ES = 85.83; AP_S = 91.18), with a modest throughput decrease (36.12 FPS). EMA-BiFPN yields comparable accuracy gains (AP50 = 93.93; AP50:95 = 70.92) and provides the strongest single-module improvements for small targets (AP_ES = 86.08; AP_S = 91.11), while slightly increasing speed (38.11 FPS), indicating a favorable accuracy–efficiency balance. MPDIoU produces the largest single-module improvement on AP50:95 (71.07) with only a small change in AP50 (93.62), suggesting its primary effect is tighter localization under stricter IoU thresholds.

Combining modules yields consistent additional gains, and the full configuration achieves the best overall results (AP50 = 94.72; AP50:95 = 71.86), with strong improvements for extremely small and small objects (AP_ES = 87.23; AP_S = 92.18). Relative to the baseline, the complete model increases AP50 by +1.22 and AP50:95 by +1.41. The gains are largest for small objects (AP_ES +2.59; AP_S +1.58) and remain positive for medium and large objects (AP_M +0.54; AP_L +0.30), while inference speed decreases slightly by 1.73 FPS. Overall, ODConv mainly trades speed for accuracy, EMA-BiFPN improves both accuracy and efficiency, and MPDIoU primarily enhances localization quality; together, they provide complementary improvements.

5. Conclusion

This work investigated UAV-based rebar detection for reinforced-concrete column inspection, where rebars commonly appear as extremely small, densely distributed targets under cluttered backgrounds, shadows, and varying illumination. To improve detection in this small-object regime, an enhanced YOLOv10 framework was developed by (i) integrating ODConv into the backbone to strengthen local feature extraction, (ii) adopting an attention-guided EMA-BiFPN neck for more effective multi-scale fusion, (iii) replacing the regression loss with MPDIoU to improve box localization under crowded and low-overlap conditions, and (iv) introducing a four-head multi-scale prediction design that includes an additional high-resolution detection branch for very small rebars.

Across experiments, the proposed model achieved consistently higher detection accuracy than baseline backbones and showed limited performance variance across random initializations, supporting the reproducibility of the observed gains. Condition-based analysis further revealed that far-field imagery is the most challenging scenario, producing the largest performance drop due to the reduced pixel footprint of rebars and weaker visual evidence—especially for extremely small instances.

Although the present study focused on offline analysis of UAV images, an important direction for future research is the development of an end-to-end real-time system in which drone video streams are processed either on an onboard computing unit or through live transmission to a companion server for immediate rebar detection during flight. Such an extension would further strengthen the practical applicability of the proposed framework within construction inspection workflows.

Future work will also focus on improving cross-site generalization through site-wise data splits, developing adaptive-resolution inference strategies for far-field scenes, and integrating counting-oriented post-processing methods to reduce errors caused by dense overlap and ambiguous object separation.

References

[1] J. Liu, P. Liu, L. Feng, W. Wu, D. Li, F. Chen, Towards automated clash resolution of reinforcing steel design in reinforced concrete frames via Q-learning and building information modeling, Autom. Constr. (2020). https://doi.org/10.1016/j.autcon.2019.103062.

[2] S. Wang, I. Eum, S. Park, J. Kim, A labelled dataset for rebar counting inspection on construction sites using unmanned aerial vehicles, Data Br. (2024) 110720. https://doi.org/10.1016/j.dib.2024.110720.

[3] Z. Wan, C. Xu, X. Chen, H. Qi, J. Liu, L. Feng, Automated rebar arrangement in prefabricated concrete components using multi-agent transfer reinforcement learning, Autom. Constr. 181 (2026) 106590. https://doi.org/10.1016/j.autcon.2025.106590.

[4] S. Wang, M. Kim, H. Hae, M. Cao, J. Kim, The Development of a Rebar-Counting Model for Reinforced Concrete Columns: Using an Unmanned Aerial Vehicle and Deep-Learning Approach, J. Constr. Eng. Manag. 149 (2023) 1–13. https://doi.org/10.1061/JCEMD4.COENG-13686.

[5] T. Sun, B. Han, S. Rusinkiewicz, Y. Shao, Rebar grasp detection using a synthetic model generator and domain randomization, Autom. Constr. 176 (2025) 106252. https://doi.org/10.1016/j.autcon.2025.106252.

[6] Y. Li, J. Chen, Computer Vision–Based Counting Model for Dense Steel Pipe on Construction Sites, J. Constr. Eng. Manag. (2022). https://doi.org/10.1061/(asce)co.1943-7862.0002217.

[7] Y. Li, Y. Lu, J. Chen, A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector, Autom. Constr. (2021). https://doi.org/10.1016/j.autcon.2021.103602.

[8] Z. Fan, J. Lu, B. Qiu, T. Jiang, K. An, A.N. Josephraj, C. Wei, Automated Steel Bar Counting and Center Localization with Convolutional Neural Networks, (2019) 1–9. http://arxiv.org/abs/1906.00891.

[9] Y. Zhu, C. Tang, H. Liu, P. Huang, End-Face Localization and Segmentation of Steel Bar Based on Convolution Neural Network, IEEE Access. (2020). https://doi.org/10.1109/ACCESS.2020.2989300.

[10] Y. Shin, S. Heo, S. Han, J. Kim, S. Na, An Image-Based Steel Rebar Size Estimation and Counting Method Using a Convolutional Neural Network Combined with Homography, Buildings. 11 (2021). https://doi.org/10.3390/buildings11100463.

[11] A.C. Hernández-Ruiz, J.A. Martínez-Nieto, J.D. Buldain-Pérez, Steel bar counting from images with machine learning, Electron. (2021). https://doi.org/10.3390/electronics10040402.

[12] G. Zhu, B. Qiu, J. Liu, Z. Wang, Z. Fan, Steel Bars Counting and Center Localization Method Based on Convolutional Neural Network and Distance Clusting, in: 2023 42nd Chinese Control Conf., 2023: pp. 8775–8779. https://doi.org/10.23919/CCC58697.2023.10241126.

[13] S. Wang, Effectiveness of traditional augmentation methods for rebar counting using UAV imagery with Faster R-CNN and YOLOv10-based transformer architectures, Sci. Rep. 15 (2025) 33702. https://doi.org/10.1038/s41598-025-18964-1.

[14] S. Wang, Domain-adaptive faster R-CNN for non-PPE identification on construction sites from body-worn and general images, Sci. Rep. (2026). https://doi.org/10.1038/s41598-026-35148-7.

[15] C. Li, A. Zhou, A. Yao, Omni-dimensional dynamic convolution, ArXiv Prepr. ArXiv2209.07947. (2022).

[16] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, Z. Huang, Efficient multi-scale attention module with cross-spatial learning, in: ICASSP 2023-2023 IEEE Int. Conf. Acoust. Speech Signal Process., IEEE, 2023: pp. 1–5.

[17] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, W. Zuo, Enhancing geometric factors in model learning and inference for object detection and instance segmentation, IEEE Trans. Cybern. 52 (2021) 8574–8586.

[18] S. Ma, Y. Xu, Mpdiou: a loss for efficient and accurate bounding box regression, ArXiv Prepr. ArXiv2307.07662. (2023).