

Feature-Enhanced RT-DETR for Small Object Detection in UAV Imagery

Miju Won
School of Electronic Engineering
Soongsil University
Seoul, South Korea
mjwon@soongsil.ac.kr

Chushi Yu
School of Electronic Engineering
Soongsil University
Seoul, South Korea
csyu@soongsil.ac.kr

Yoan Shin*
School of Electronic Engineering
Soongsil University
Seoul, South Korea
yashin@ssu.ac.kr

Abstract—Unmanned aerial vehicle (UAV)-based imagery often contains small objects with weak boundary and localization cues due to large viewing distances, complex backgrounds, and dense object distributions. In this paper, we propose a feature-enhanced real-time detection transformer (RT-DETR)-r18 model for UAV small object detection, while preserving the NMS-free end-to-end detection pipeline. The proposed model integrates the shallow feature enhancement module (SFEM) at the P2 stage, the content-aware reassembly of features (CARAFE)-Upsampling module in the neck, and C3k2 blocks after feature fusion to refine multi-scale feature representations. Experiments on the VisDrone2019 dataset show that the proposed model improves both mAP_{50} and $mAP_{50:95}$ compared with the baseline RT-DETR-r18. In particular, the improvement in $mAP_{50:95}$ indicates enhanced bounding box localization under stricter intersection over union (IoU) thresholds. These results demonstrate that the proposed model improves UAV small object detection accuracy, while maintaining practical end-to-end processing speed.

Index Terms—UAV imagery, small object detection, RT-DETR-r18, shallow feature enhancement, CARAFE-Upsampling.

I. INTRODUCTION

Unmanned aerial vehicle (UAV)-based object detection is widely used in disaster monitoring, traffic surveillance, and remote reconnaissance. However, UAV images are often captured from high altitudes and wide fields of view, causing target objects to appear small.

Small objects in UAV imagery are prone to detection and localization errors due to a limited number of pixels, complex backgrounds, and dense object distributions. Transformer-based detectors such as detection transformer (DETR) perform object detection in an end-to-end manner and reduce the dependence on hand-crafted post-processing such as non-maximum suppression (NMS) [1]. While real-time DETR (RT-DETR) is advantageous due to its NMS-free end-to-end detection structure [2], lightweight models such as RT-DETR-r18 may suffer from weakened spatial details and boundary cues of small objects during the iterative downsampling process.

In this paper, we propose a feature-enhanced RT-DETR-r18 model for UAV small object detection. The proposed model enhances spatial details and boundary cues by integrating

the shallow feature enhancement module (SFEM) at the P2 stage, the content-aware reassembly of features (CARAFE)-Upsampling in the neck, and C3k2 blocks after feature fusion. We analyze the effect of these modules on detection accuracy, bounding box localization quality, and the accuracy-speed trade-off using end-to-end frames per second (FPS).

II. PROPOSED METHOD

A. Overall Architecture

In this paper, the enhanced architecture is designed to improve small object features, while preserving the NMS-free end-to-end detection pipeline of RT-DETR-r18. As shown in Fig. 1, the baseline model consists of a ResNet-18 backbone, an efficient hybrid encoder, intersection over union (IoU)-aware query selection, and an RT-DETR decoder, forming an NMS-free end-to-end detection pipeline.

Compared with the baseline, the proposed model introduces three modifications as illustrated in Fig. 2. Firstly, SFEM is inserted at the P2 stage of the backbone to enhance shallow spatial and boundary information before further downsampling. The enhanced P2 feature is propagated to the subsequent backbone stages to generate P3, P4, and P5 feature maps. Secondly, nearest-neighbor upsampling in the top-down neck path is replaced with CARAFE-Upsampling for content-aware feature reassembly. Thirdly, C3k2 blocks are applied after feature fusion to refine the fused multi-scale features before the RT-DETR decoder.

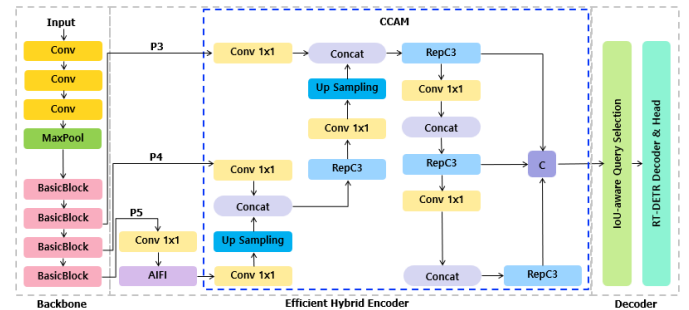


Fig. 1. Architecture of the baseline RT-DETR-r18 model.

*Corresponding author.

This work was supported in part by the NRF of Korea grant funded by the MSIT under Grant RS-2025-02214082, and by the MSIT under the ITRC support program (IITP-2026-RS-2023-00258639) supervised by the IITP.

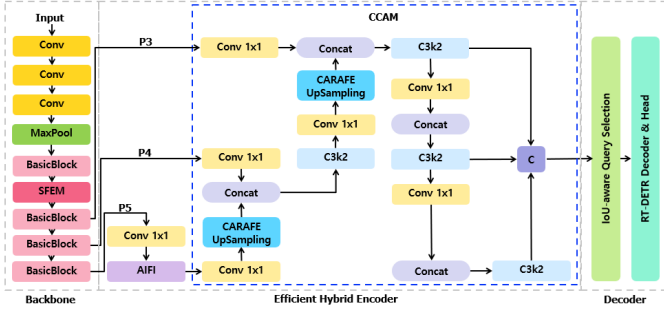


Fig. 2. Architecture of the proposed feature-enhanced RT-DETR-r18 model.

B. Feature Enhancement Modules

1) *SFEM*: The SFEM is applied to the P2 feature map to enhance fine spatial and boundary information for small-object detection. As shown in Fig. 3, it consists of bottleneck layers and a FastResidual refinement block. The input feature is projected and split into preserved and refined branches, which are subsequently concatenated to generate the enhanced P2 feature. Following Drone-DETR [3], the SFEM operation can be described as

$$[U_0, U_1] = \text{Split}(\text{Conv}_{1 \times 1}(X_{P2})), \quad (1)$$

$$U_{i+1} = \mathcal{F}_{FR}(U_i), \quad i = 1, \dots, n, \quad (2)$$

$$X_{P2}^{enh} = \text{Conv}_{1 \times 1}(\text{Concat}(U_0, U_1, \dots, U_{n+1})), \quad (3)$$

where X_{P2} and X_{P2}^{enh} denote the input and enhanced P2 feature maps, respectively. The initial 1×1 convolution projects the input feature and divides it into two branch features, U_0 and U_1 . The branch U_0 preserves the original shallow spatial information, while U_1 is progressively refined by the FastResidual operation $\mathcal{F}_{FR}(\cdot)$. The preserved and refined features are concatenated and projected through the final 1×1 convolution to produce a compact enhanced representation with the channel dimensions required by the subsequent backbone stage. The enhanced P2 feature is then propagated through the backbone to generate P3, P4, and P5 feature maps with strengthened small-object information.

2) *CARAFE-Upsampling*: The CARAFE-Upsampling module replaces nearest-neighbor upsampling in the top-down neck path. As shown in Fig. 4, the input feature is used in two paths: kernel prediction and content-aware reassembly. In the kernel prediction path, the input feature is compressed, encoded, and normalized to generate location-wise reassembly kernels. In the reassembly path, local features from the same input feature are reassembled using the predicted kernels to produce the upsampled feature. Based on CARAFE [4], the operation can be described as

$$W_{l'} = \psi(\mathcal{N}(X_l, k_{enc})), \quad (4)$$

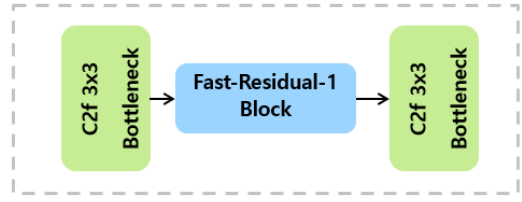


Fig. 3. Structure of the SFEM.

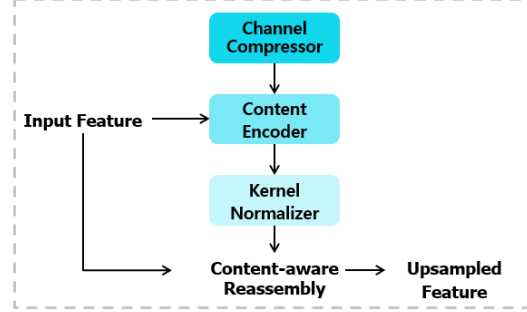


Fig. 4. Structure of the CARAFE-Upsampling module.

$$X'_{l'} = \sum_{q \in \mathcal{N}(l, k_{up})} W_{l',q} \cdot X_q, \quad (5)$$

where X_l denotes the input feature at location l , and $X'_{l'}$ denotes the upsampled feature at location l' . $W_{l'}$ is the location-wise reassembly kernel predicted from the local neighborhood $\mathcal{N}(X_l, k_{enc})$, and $W_{l',q}$ denotes the weight assigned to the feature value X_q in the reassembly region $\mathcal{N}(l, k_{up})$. Through this weighted aggregation, CARAFE reconstructs the upsampled feature according to local feature content rather than fixed interpolation.

3) *C3k2 for Post-fusion Feature Refinement*: C3k2 is used after feature fusion to refine the multi-scale feature representation. After top-down fusion, shallow spatial cues and high-level semantic context are aggregated into the same feature stream. C3k2 is introduced as a post-fusion refinement block to improve local feature mixing before the fused features are passed to the RT-DETR decoder.

III. EXPERIMENTAL SETUP

A. Dataset

We use the VisDrone2019-DET [5] dataset to evaluate UAV object detection performance. VisDrone2019-DET is a widely used UAV-based object detection benchmark that contains images captured from diverse urban scenes under different capture altitudes, viewpoints, illumination conditions, and object densities. The dataset includes ten object categories: pedestrian, people, bicycle, car, van, truck, tricycle, awning-tricycle, bus, and motor. With many small and densely distributed objects, VisDrone2019-DET is suitable for evaluating small-object detection performance in UAV imagery. The dataset consists of 6,471 training images, 548 validation images, and 1,610 test-dev images. In this study, the training set is used

TABLE I
ABLATION STUDY OF THE PROPOSED MODULES ON THE VisDrone2019-DET VALIDATION SET.

Model	Module			P (%)	R (%)	mAP ₅₀ (%)	mAP _{50:95} (%)	Params (M)	GFLOPs	FPS
	C3k2	SFEM	CARAFE							
RT-DETR-r18	–	–	–	59.70	43.76	45.54	27.50	20.11	57.9	63.17
	–	✓	–	59.74	43.92	45.93	28.25	20.12	57.9	61.32
	–	–	✓	61.23	44.77	46.51	28.54	20.24	57.3	59.24
	✓	✓	–	61.09	44.98	46.50	28.48	20.81	57.3	49.36
	✓	–	✓	61.32	45.05	46.64	28.69	20.96	61.9	34.63
	–	✓	✓	59.63	45.14	46.39	28.75	20.05	58.8	58.27
Ours	✓	✓	✓	60.50	44.82	46.71	28.82	20.51	58.4	44.84

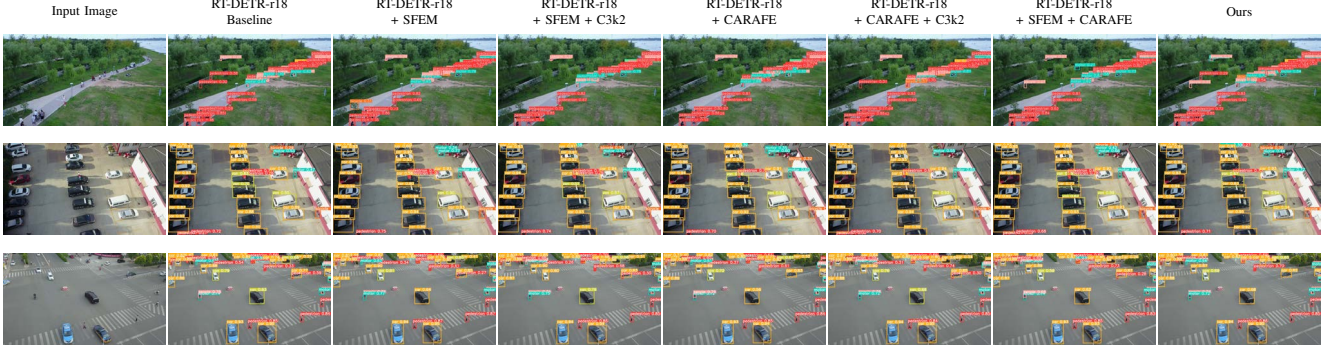


Fig. 5. Qualitative comparison on three representative VisDrone2019-DET validation images. The rows correspond to 0000023_00000_d_0000008, 0000026_01000_d_0000026, and 0000330_02201_d_0000811, respectively, while the columns show different model configurations.

to train each model, and the final detection performance is evaluated on the 1,610-image test-dev set.

B. Implementation Details

All experiments are performed under identical settings on Ubuntu 20.04 using Python 3.9 and the PyTorch 2.8.0 framework with an NVIDIA RTX 3090 GPU. RT-DETR-r18 is used as the baseline model. The input image size is set to 640×640 , the batch size is set to 4, and all models are trained for 150 epochs using the AdamW optimizer. The initial learning rate is set to 0.0001, the momentum is set to 0.9, and the weight decay is set to 0.0001. These settings are applied consistently to evaluate the effect of each module on object detection accuracy and end-to-end processing speed.

C. Evaluation Metrics

Following the COCO-style evaluation protocol [6], we evaluate each model in terms of detection accuracy, localization quality, model complexity, and real-time performance. precision (P), recall (R), mean average precision (mAP), and average precision (AP)-based metrics are reported, including mAP₅₀, mAP_{50:95}, AP₇₅, AP_s, AP_m, and AP_l. AP is computed from the precision-recall curve, and mAP_{50:95} averages AP over IoU thresholds from 0.5 to 0.95 with a step size of 0.05. In addition, AP₂₅ is used as a relaxed IoU-based metric to analyze coarse detection quality in UAV imagery. AP_s, AP_m, and AP_l are used for scale-specific analysis.

Parameters and giga floating-point operations (GFLOPs) are reported to compare model complexity. For real-time performance, we use end-to-end FPS, which includes pre-processing, model inference, and postprocessing. This metric reflects the practical processing speed of UAV-based object detection systems more directly than inference-only speed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Ablation Study

As shown in Table I, the proposed modules generally improve detection accuracy compared with the baseline RT-DETR-r18. SFEM improves mAP₅₀ and mAP_{50:95} by enhancing shallow spatial information at the P2 stage, while CARAFE-Upsampling provides further gains by performing content-aware feature reassembly in the top-down fusion path.

The combination of SFEM, CARAFE-Upsampling, and C3k2 achieves the highest mAP₅₀ and mAP_{50:95} among all evaluated models. Compared with the baseline, the proposed model improves mAP₅₀ from 45.54 to 46.71 and mAP_{50:95} from 27.50% to 28.82%, with a modest increase in parameter count from 20.11M to 20.51M. Although the end-to-end FPS decreases from 63.17 to 44.84, the proposed model maintains practical processing speed while improving detection and localization performance.

Fig. 5 provides a qualitative comparison of the module combinations on three representative images from the VisDrone2019-DET validation set. The proposed model shows more consistent detection of small and densely distributed

objects than the baseline and the individual module configurations. These qualitative results support the effectiveness of the combined modules.

B. Detailed AP Analysis

Table II compares detailed AP-based metrics between the baseline and the proposed model on the VisDrone2019-DET test set. The proposed model improves mAP_{50} , $mAP_{50:95}$, AP_{25} , and AP_{75} compared with RT-DETR-r18, indicating improved detection accuracy and localization quality under both relaxed and stricter IoU thresholds. For scale-specific AP, AP_s and AP_l increase slightly, while AP_m shows a small decrease. These results suggest that the proposed modules mainly improve localization-related AP metrics, while maintaining comparable scale-specific detection performance.

TABLE II
DETAILED AP COMPARISON ON THE VISDRONE2019-DET TEST SET.

Model	mAP_{50}	$mAP_{50:95}$	AP_{25}	AP_{75}	AP_s	AP_m	AP_l
RT-DETR-r18	46.5	28.3	49.4	28.8	9.69	22.0	13.5
Ours	46.7	28.6	49.8	29.1	9.84	21.8	14.0

V. CONCLUSION

In this paper, we proposed a feature-enhanced RT-DETR-r18 model for UAV small-object detection by combining SFEM, CARAFE-Upsampling, and C3k2 modules. The proposed structure preserves the NMS-free end-to-end detection pipeline while enhancing shallow P2 features, performing content-aware upsampling, and refining fused multi-scale features. Experiments on VisDrone2019-DET demonstrate improvements in mAP_{50} and $mAP_{50:95}$ over the baseline. In particular, the improvement in $mAP_{50:95}$ indicates enhanced localization quality under stricter IoU thresholds, while the ablation study confirms the effectiveness of the combined modules.

REFERENCES

- [1] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *Proc. ECCV 2020*, pp. 213–229, Online, Aug. 2020.
- [2] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," *Proc. IEEE/CVF CVPR 2024*, pp. 16965–16974, Seattle, USA, June 2024.
- [3] Y. Kong, X. Shang, and S. Jia, "Drone-DETR: Efficient small object detection for remote sensing image using enhanced RT-DETR model," *Sensors*, vol. 24, no. 17, Art. no. 5496, Aug. 2024.
- [4] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "CARAFE: Content-aware reassembly of features," *Proc. IEEE/CVF ICCV 2019*, pp. 3007–3016, Seoul, Korea, Oct. 2019.
- [5] D. Du, P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," *Proc. IEEE/CVF ICCVW 2019*, pp. 213–226, Seoul, Korea, Oct. 2019.
- [6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common objects in context," *Proc. ECCV 2014*, pp. 740–755, Zurich, Switzerland, Sept. 2014.