

Lightweight RT-DETR with Collaborative Attention and Multi-Scale Gated Fusion

Mingrui Han and Xu Zhao*

Abstract—To address the critical challenge of balancing accuracy and efficiency in real-time object detection, this paper proposes a lightweight-enhanced RT-DETR detector. Our method achieves coordinated optimization at both the backbone and encoder levels. First, a novel C2f-SCSPA module is introduced to replace the original ResNet backbone. It integrates a Synergistic Channel-Space-Pixel Attention (SCSPA) mechanism within an efficient C2f architecture, facilitating adaptive semantic refinement and multi-scale information aggregation during feature extraction. Second, a Multi-Scale Adaptive Gated Fusion (MSAGF) module is incorporated at the encoder level, enabling context-aware and spatially adaptive fusion for effective cross-scale feature interaction. Comprehensive experiments on the MS COCO dataset demonstrate that the proposed model achieves a significant improvement, with mAP50 rising to 68.29% and mAP@0.5-0.95 to 50.5%. Evaluated on the PASCAL VOC dataset, it attains a competitive mAP@0.5 of 82.4% while maintaining a lightweight profile of merely 9.9M parameters and 19.2 GFLOPs. Remarkably, the model delivers a high inference speed of 27 FPS on an NVIDIA Jetson Orin Nano edge device. These results validate its superior balance between accuracy and efficiency, showcasing strong potential for practical, resource-constrained applications.

I. INTRODUCTION

Object detection is a core computer vision task that requires balancing accuracy and speed for real-time applications like autonomous driving. Complex scenarios demand models with strong multi-scale perception and feature fusion capabilities. Current mainstream approaches are based on convolutional or Transformer architectures: the YOLO series maintains leading real-time performance through its grid prediction and multi-scale mechanisms, but its reliance on post-processing such as non-maximum suppression may introduce latency and suboptimal results. End-to-end detectors represented by RT-DETR [1] streamline the pipeline by employing Transformer encoder-decoder architecture for direct set prediction while preserving real-time efficiency.

Despite significant progress, RT-DETR still exhibits several limitations. First, RT-DETR employs a ResNet backbone. While offering strong feature extraction, it introduces high computational cost due to its complex architecture, making it less suitable for strictly resource-constrained environments. Second, during feature fusion, the model primarily relies on feature map splitting and convolutional operations, lacking explicit global context modeling and an adaptive fusion mechanism. This restricts the model's feature discriminability and generalization in complex scenes.

Mingrui Han and Xu Zhao are with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China. This work has been funded in part by the NSFC grants 62176156.

*Corresponding Author. Email: zhaoxu@sjtu.edu.cn

To address these challenges, we propose two novel modules: C2f-SCSPA for backbone enhancement and MSAGF for encoder fusion. The C2f-SCSPA module introduces a Synergistic Channel-Space-Pixel Attention (SCSPA) mechanism into the efficient C2f architecture, which allows the backbone network to adaptively focus on regions of varying detection difficulty, enhancing feature representation quality while maintaining low computational overhead. The Multi-Scale Adaptive Gated Fusion (MSAGF) module employs a context-aware and spatially adaptive fusion strategy to dynamically establish connections across multi-scale features in the encoder. By adaptively weighting and fusing features based on their semantic importance and spatial context, MSAGF enables more refined and flexible feature interaction.

Through the synergistic integration of C2f-SCSPA and MSAGF into the RT-DETR framework, we build a detection architecture that balances discriminative power and efficiency. The enhanced backbone provides richer and more adaptive multi-scale features, while the intelligent fusion mechanism ensures optimal integration of these features, leading to a holistic improvement in detection accuracy and computational efficiency.

The main contributions of this paper are summarized as follows: (i) For the backbone network, we propose the C2f-SCSPA module, which effectively enhances the model's discriminative power and representational capacity for multi-scale features; (ii) For the encoder design, we propose the MSAGF module, which employs a context-aware and spatially adaptive fusion strategy to improve the interaction and fusion efficiency across features at different scales; (iii) The proposed approach achieves comprehensive improvements in detection accuracy and robustness within the RT-DETR framework, delivering strong performance on standard benchmarks such as COCO [2] and VOC [3], while striking an excellent balance between efficiency and performance.

II. RELATED WORK

Research on real-time object detectors has consistently pursued an optimal trade-off between accuracy and inference speed. The technological evolution in this field can be broadly categorized into two distinct eras: the conventional Convolutional Neural Network (CNN) era and the subsequent Transformer-integrated era.

During the CNN-dominant period, the research focus centered on optimizing convolutional architectures and detection head designs. YOLOv1 [4] is the first CNN-based one-stage object detector to achieve true real-time object detection.

And SSD [5] established an efficient foundation for single-stage detectors by enabling direct predictions on multi-scale feature maps. YOLOv5 [6] further systematized this approach by introducing components like the C3 module and adaptive anchor computation, achieving a robust accuracy-speed balance in practical deployment. A common characteristic of models from this era was their heavy reliance on hand-crafted components, such as predefined anchor boxes, and post-processing steps like Non-Maximum Suppression (NMS).

The introduction of DETR [7] in 2020 marked a paradigm shift towards end-to-end object detection based on set prediction, offering a simplified pipeline free of manual design heuristics. However, its initial formulation suffered from slow convergence and high computational complexity. Subsequent research has primarily focused on mitigating these bottlenecks, leading to a clear trajectory of improvements.

Deformable DETR [8] addressed the computational issue by proposing a deformable attention mechanism, which restricts each query's attention to a sparse set of key sampling points, significantly reducing complexity and accelerating convergence. Concurrently, Conditional DETR [9] facilitated training by decoupling content and spatial attention, enabling more precise spatial localization. To stabilize the challenging bipartite matching process during training, DN-DETR [10] introduced a denoising training strategy that guides the model by reconstructing corrupted ground-truth annotations. Another line of work sought to ground the abstract object queries. DAB-DETR [11] reformulated queries as dynamic anchor boxes, providing explicit spatial priors from the onset of training. Building upon these advances, DINO [12] comprehensively integrated techniques including contrastive denoising training and dynamic anchor initialization, and employed a mixed query selection mechanism to achieve state-of-the-art performance.

The drive towards real-time application spurred further architectural innovations. The pursuit of real-time performance has driven further architectural innovations aimed at reducing computational overhead. Efficient DETR [13] and Sparse DETR [14] decrease computational cost by strategically reducing the number of encoder/decoder layers or selectively updating a sparse set of object queries, respectively. Following this direction, Lite DETR [15] alleviates redundancy through an alternating update scheme for multi-scale encoder features. Similarly, Focus DETR [16] reduces computation by filtering out background regions, allowing the model to concentrate attention on more promising foreground areas.. This progression culminated in RT-DETR [1], which realized high-performance real-time inference for an end-to-end Transformer detector via an efficient hybrid encoder and an IoU-aware query selection mechanism, representing a significant milestone.

In parallel, the YOLO series has sustained its evolution by strategically integrating advancements from the end-to-end paradigm. YOLOv8 [17] enhanced its backbone architecture with the C2f module and refined its training objective through a Task-aligned Assigner, achieving a more robust

balance between accuracy and inference speed. Building upon this, YOLOv9 [18] introduced innovative concepts like Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) to strengthen gradient flow and feature representation capabilities. Following this trajectory, YOLOv10 [19] made a significant architectural shift by adopting a fully end-to-end design, eliminating NMS through dual label assignment and incorporating a Partial Self-Attention (PSA) module for enhanced global context modeling. The subsequent development, YOLOv11 [20], has focused extensively on practical deployment efficiency, leveraging techniques such as Neural Architecture Search (NAS) and streamlined model scaling to optimize performance across diverse hardware constraints.

III. METHOD

A. Model Overview

The improved model consists of a C2f-SCSPA backbone, a hybrid encoder, and an RT-DETR decoder with several auxiliary prediction heads. The overview of the model is illustrated in Fig.1.

To pursue a lightweight design, we adopt a backbone constructed with efficient C2f modules to replace the original ResNet. This module is further enhanced by integrating a synergistic channel-space-pixel attention mechanism, leading to the proposed C2f-SCSPA module(*cf.* Sec. III-B).

Specifically, features extracted from the fourth, sixth and ninth modules of the backbone are fed into the encoder, denoted as $\{P4, P6, P9\}$ respectively. The hybrid encoder transforms multi-scale features into a sequence of image features through intra-scale interaction and cross-scale fusion, incorporating our proposed Multi-Scale Adaptive Gated Fusion (MSAGF) module for effective feature integration(*cf.* Sec. III-C).

Finally, a fixed number of encoder features are selected as initial object queries for the decoder and iteratively optimized by the Transformer decoder to generate categories and bounding boxes.

B. Improved Backbone with C2f-SCSPA

To reduce the computational cost of the model, we replace the ResNet backbone in RT-DETR with the lightweight C2f module. To address its limited ability in adaptive feature weighting, we further propose the C2f-SCSPA module, which incorporates a Synergistic Channel-Space-Pixel Attention (SCSPA) mechanism.

The mathematical formulation of the triple attention synergy is as follows.

The channel attention mechanism aims to model the importance relationships between different channels in a feature map. Its core computation can be summarized in three key steps. First, the input features X undergo both global average pooling and global max pooling to aggregate spatial information. Next, the resulting vectors from both pooling operations are fed into a shared Multilayer Perceptron (MLP) for non-linear mapping and dimensionality reduction. Finally, the

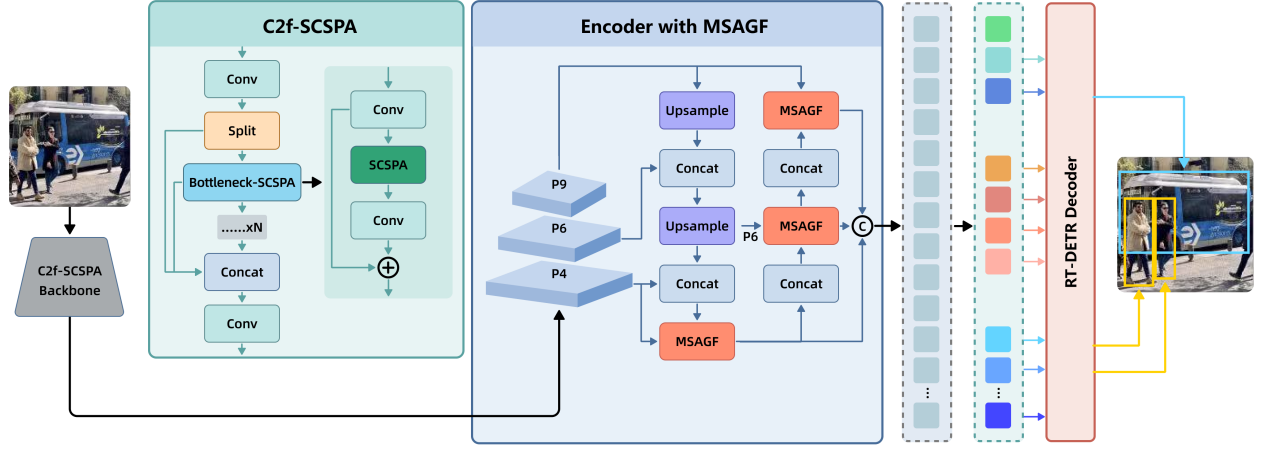


Fig. 1: Model overview. The model consists of three components: the efficient C2f-SCSPA backbone, an encoder incorporating the MSAGF module, and the RT-DETR decoder, illustrating the end-to-end pipeline from input image to output detections.

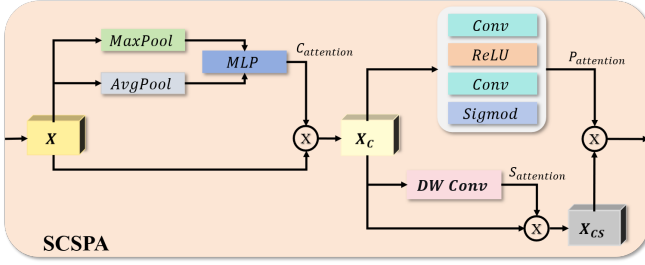


Fig. 2: Structure of the Synergistic Channel- Spatial-Pixel Attention (SCSPA) module

outputs from the two MLP pathways are summed element-wise and processed through a Sigmoid activation function to generate the final channel attention weights $C_{\text{attention}}$.

$$C_{\text{attention}} = \sigma(\text{MLP}(\text{GAP}(X)) + \text{MLP}(\text{GMP}(X))) \quad (1)$$

These weights $C_{\text{attention}} \in \mathbb{R}^{C \times 1 \times 1}$ are subsequently used to recalibrate the original input features X along the channel dimension.

Then the channel-weighted features are given by:

$$X_c = X \otimes C_{\text{attention}} \quad (2)$$

where $\otimes$ denotes element-wise multiplication along the channel dimension.

The spatial attention mechanism captures multi-scale spatial context and emphasizes informative regions within feature maps. This is achieved through multi-scale depthwise separable convolutions that aggregate contextual information across different receptive fields, enabling the model to focus on the most salient spatial locations.

$$S_{\text{attention}} = \text{Conv}_{1 \times 1} \left(\sum_{k \in \{7, 11, 21\}} D_k(X_c) \right) \quad (3)$$

$$D_k(X) = \text{Conv}_{(k, 1)} (\text{Conv}_{(1, k)}(X)) \quad (4)$$

The resulting spatially-weighted features are then obtained by applying these attention weights to the input features through element-wise multiplication.

$$X_{cs} = X_c \otimes S_{\text{attention}} \quad (5)$$

The pixel attention mechanism performs fine-grained pixel-wise calibration, enabling the model to adaptively emphasize or suppress individual pixel positions based on their local context. This provides more refined feature optimization than channel and spatial attention, thereby enhancing the representation of detailed local patterns. The pixel attention weights are computed as

$$P_{\text{attention}} = \sigma(\text{Conv}_{1 \times 1}(\delta(\text{Conv}_{1 \times 1}(X_c)))) \quad (6)$$

The final output features are obtained by applying these pixel-wise attention weights to the input features.

$$X_{\text{output}} = X_{cs} \otimes P_{\text{attention}} \quad (7)$$

The SCSPA module achieves multi-scale feature calibration from global context to local details through the synergy of channel, spatial, and pixel attention, significantly enhancing the model's focus on critical features. And the enhanced C2f-SCSPA module improves multi-scale representation in complex scenarios while maintaining high computational efficiency.

C. Hybrid Encoder with MSAGF

To overcome the limitations of self-attention and adaptive fusion in the original RT-DETR encoder, this paper introduces a Multi-Scale Adaptive Gated Fusion (MSAGF)

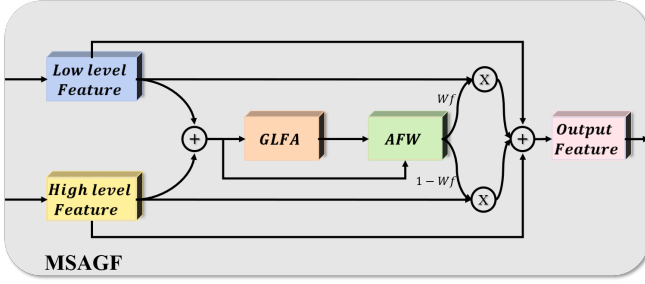


Fig. 3: Schematic diagram of the Multi-scale Adaptive Gated Fusion (MSAGF) module

module, forming an enhanced hybrid encoder. The module employs a collaborative dual-path design that aggregates multi-scale context through global-local feature extraction and guides fusion via an adaptive weighting mechanism.

The GLFA module takes two aligned features as input: the high-level semantic features F_{high} and the low-level detailed features F_{low} . Their element-wise sum yields an initial fused feature F_0 .

$$F_0 = F_{high} + F_{low} \quad (8)$$

Global and Local Feature Aggregation: The GLFA module processes the aggregated feature F_0 to simultaneously extract local details and global contextual information, enhancing representation through an adaptive dual-path design. The core operations are as follows:

The GLFA module takes the initial feature F_0 as input and extracts both local details and global context simultaneously through a parallel dual-path structure. First, the local feature f_{local} is derived via a multi-branch convolution designed to capture fine-grained features under different receptive fields.

$$f_{local} = \text{Conv}_{\text{fuse}}\left(\mathcal{S}(\text{Concat}(C_1, C_2, C_3, C_4))\right) \quad (9)$$

where $C_i = \text{Conv}_i(F_0)$ denotes the output of the i -th convolutional branch, and $\mathcal{S}(\cdot)$ represents the channel shuffle operation. Meanwhile, the global path captures long-range dependencies via a self-attention mechanism.

$$\begin{aligned} Q, K, V &= \text{QKV_Block}(F_0), \\ A &= \text{Softmax}\left(\frac{QK^\top}{\sqrt{C}}\right), \\ f_{\text{global}} &= \text{Linear}(AV) + x. \end{aligned} \quad (10)$$

Finally, the local and global features are fused through residual addition to produce the refined output:

$$F_{GLFA} = f_{local} + f_{global} \quad (11)$$

This design enables the module to effectively integrate multi-scale information, thereby enhancing the representational capacity of features in complex scenes.

Adaptive Fusion Weighting: The AFW module generates spatially adaptive weights by combining the initially

aggregated feature F_0 and the contextually enriched feature F_{GLFA} . The two features are first concatenated along the channel dimension.

$$F_{\text{concat}} = \text{Concat}(F_0, F_{GLFA}) \quad (12)$$

A grouped convolution (group number = C) is then applied, followed by a sigmoid activation, yielding the soft weight map:

$$W_f = \sigma(\text{Conv}_{\text{groups}=C}(F_{\text{concat}})) \quad (13)$$

where σ denotes the sigmoid function, which normalizes each weight to $[0, 1]$ to enable soft feature selection.

Finally, the MSAGF module fuses the multi-level features using the adaptive weight map W_f . The output is computed as:

$$F_{\text{out}} = F_0 + W_f \otimes F_{high} + (1 - W_f) \otimes F_{low} \quad (14)$$

where $\otimes$ represents element-wise multiplication.

MSAGF module effectively overcomes the limitations of plain feature fusion in the original encoder. It enhances the model's ability to accurately extract features in complex scenes, especially for small and occluded objects, while maintaining a lightweight design.

IV. EXPERIMENT

A. Experimental Settings

Dataset To comprehensively evaluate the proposed model, we conducted experiments on MS COCO and PASCAL VOC datasets. The model was first trained on the MS COCO 2017 train2017 split (118,287 images) and validated on val2017 (5,000 images) to examine its detection capability under complex scenes, diverse object scales, and challenging contextual variations. Subsequently, to assess generalization, the model was further trained on the PASCAL VOC 2012 train set and evaluated on its official val set for final comparative analysis.

Evaluation metric Object detection performance is assessed using precision, recall, and mean Average Precision (mAP). Precision reflects the proportion of correct positive predictions, while recall measures the model's ability to identify all relevant objects. mAP@0.5 averages precision at an IoU threshold of 0.50, whereas mAP@0.5:0.95 evaluates performance across IoU thresholds from 0.50 to 0.95, offering a stricter measure of localization accuracy. Model efficiency is evaluated via parameter count (Params) and computational complexity (GFlops). Params indicate model size, and GFlops estimate inference speed and energy consumption. Together, these metrics provide a comprehensive assessment for applications requiring both high accuracy and practical efficiency.

Implementation Details All experiments are conducted on an NVIDIA RTX 5090 GPU for model development and training. We employ the RT-DETR architecture with a decoder configuration using the RT-DETRDecoder module,

TABLE I: Ablation study results

Model Configuration	Epoch	Precision	Recall	mAP@0.5	mAP@0.5-0.95	Params(M)	GFLOPs
Baseline	300	64.2	48.5	54.3	39.5	6.2	12.1
Baseline+SCSPA	300	66.5	53.8	59.5	42.9	6.2	12.39
Baseline+MSAGF	300	67.8	52.7	61.6	41.4	9.8	18.82
Baseline+SCSPA+MSAGF	30.0	68.6	61.6	68.3	50.5	9.9	19.17

which processes multi-scale feature maps from three levels. The model is primarily trained and evaluated on the COCO 2017 dataset for 300 epochs with a batch size of 128, using the AdamW optimizer. Input images are resized to 640×640 pixels. Subsequently, the model is further evaluated on the PASCAL VOC dataset. For deployment-oriented evaluation, the inference speed is measured on an NVIDIA Jetson Orin Nano developer kit under TensorRT FP16.

B. Ablation Study

To evaluate the contribution of each proposed component—C2f-SCSPA backbone and MSAGF fusion module—we conduct an ablation study on the baseline model. The results are summarized in TABLE I.

Baseline+SCSPA: Incorporating the SCSPA module further improves model performance (precision 67.8%, mAP@0.5 61.6%, mAP@0.5:0.95 41.4%), while increasing parameters to 9.8M and computation to 18.82G FLOPs. This validates the effectiveness of MSAGF in facilitating context-aware, cross-scale feature fusion, with the performance gains coming at an acceptable additional computational cost.

Baseline+MSAGF: Integrating the MSAGF module into the baseline leads to further improvements, yielding a precision of 67.8%, mAP@0.5 of 61.6%, and mAP@0.5:0.95 of 41.4%. This advancement comes with an increased model capacity, raising parameters to 9.8M and FLOPs to 18.82G. This performance improvement confirms that the module effectively achieves adaptive cross-scale feature fusion with contextual awareness.

Baseline+SCSPA+MSAGF: The integration of both C2f-SCSPA and MSAGF modules yields the best overall performance. The full model achieves 68.6% precision, 61.6% recall, 68.3% mAP@0.5, and 50.5% mAP@0.5:0.95 with only 9.9M parameters and 19.17G FLOPs. This demonstrates clear synergy: C2f-SCSPA provides more discriminative multi-scale features, which are then optimally fused by MSAGF, leading to a remarkable 25.8% relative improvement in mAP@0.5 over the baseline.

The ablation study confirms the core value of both modules: the C2f-SCSPA backbone significantly enhances feature extraction at minimal cost, while the MSAGF module enables powerful adaptive fusion with a controlled complexity increase. Their co-design successfully pushes the accuracy-efficiency frontier, validating the joint optimization approach for lightweight, high-performance detection.

C. Main Results and Comparison

To comprehensively evaluate the performance of the proposed method, we conducted extensive comparative experi-

ments on the COCO and VOC datasets, comparing it with current mainstream lightweight detectors. The results are shown in Table II and Table III, respectively.

As shown in Table II, our full model achieves 68.3% mAP@0.5 and 50.5% mAP@0.5:0.95 on the COCO validation set, surpassing all compared lightweight models. With only 9.9M parameters and 19.17G FLOPs, it achieves an excellent balance between computational efficiency and detection accuracy. Compared to the YOLO series, our method demonstrates significant advantages. It leads YOLOv8-S and YOLOv9-S by 6.5% in mAP@0.5, respectively. Even compared to the latest YOLOv11-S, it achieves a 4.4% improvement in mAP@0.5 with only a 0.9M parameter increase.

Against Transformer-based detectors, our model exhibits overwhelming efficiency advantages. Compared to RT-DETR-R18, our approach reduces parameters by over 50% and computation to about one-third, while maintaining superior accuracy. Compared to the current best-performing DEIM-D-FINE-S (10M parameters, 25G FLOPs), our model achieves higher mAP@0.5 and mAP@0.5:0.95 with lower computational cost (19.17G FLOPs), further validating its dual advantages in lightweight design and accuracy.

To further validate the model’s generalization capability and practical deployment efficiency, we conducted comparative experiments on the PASCAL VOC dataset and the NVIDIA Jetson Orin Nano edge device, as shown in Table III. On the VOC dataset, our model achieved 82.4% mAP@0.5, outperforming YOLOv10-S (82.1%) and RT-DETR-R18 (81.3%), while performing comparably to D-FINE-S (82.7%). With fewer parameters and significantly reduced computational complexity, this result reaffirms the high efficiency of our approach. In edge deployment speed tests, our model attained an inference speed of 27 FPS on Jetson Orin Nano, significantly surpassing other compared models.

V. CONCLUSIONS

This paper advances the performance of the RT-DETR detector through two synergistic modules: C2f-SCSPA, which enhances multi-scale feature representation with a triple-synergistic attention mechanism, and MSAGF, which enables effective cross-scale integration through context-aware and spatially adaptive feature fusion. Together, they form an enhanced hybrid detection architecture that significantly improves both accuracy and robustness while maintaining high efficiency.

TABLE II: Comparison results with SOTA on COCO dataset

Model	Epoch	mAP@0.5	mAP@0.5:0.95	Params(M)	GFLOPs
YOLOv8-S [17]	500	61.8	44.9	11	29
YOLOv9-S [18]	500	61.8	46.8	7	26
YOLOv10-S [19]	500	63.0	46.3	7	22
YOLOv11-S [20]	500	63.9	47.0	9	22
RT-DETR(r18)*	300	49.6	30.6	20	57.3
RT-DETR(r18) [1]	/	63.8	46.5	20	61
D-FINE-S [21]	/	65.6	48.5	10	25
DEIM-D-FINE-S [22]	/	65.9	49.0	10	25
Ours	300	68.3	50.5	9.9	19.17

Note: * indicates our reproduced results. For methods without epoch specification (""/), the results are reported as provided in the original papers.

TABLE III: Comparison results on VOC and Edge Deployment Speed

Model	mAP@0.5	Params(M)	GFLOPs	FPS(Orin Nano)
YOLOv8-S	81.5	11	28.6	21
YOLOv10-S	82.1	7	21.6	24
RT-DETR(r18)	81.3	20	57.3	25
D-FINE-S	82.7	10	25.0	20
Ours	82.4	9.9	19.2	27

In summary, this work provides an efficient solution for high-precision real-time object detection, achieving an optimal balance between performance and computational cost. The proposed method sets a new benchmark for RT-DETR-based detectors and demonstrates strong potential for deployment in resource-constrained applications.

REFERENCES

- [1] 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," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 16 965–16 974.
- [2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [3] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [6] G. Jocher, A. Stoken, J. Borovec, L. Changyu, A. Hogan, L. Diaconu, J. Poznanski, L. Yu, P. Rai, R. Ferriday *et al.*, "ultralytics/yolov5: v3.0," *Zenodo*, 2020.
- [7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
- [8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.
- [9] D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, and J. Wang, "Conditional detr for fast training convergence," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3651–3660.
- [10] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.
- [11] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.
- [12] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.
- [13] Z. Yao, J. Ai, B. Li, and C. Zhang, "Efficient detr: improving end-to-end object detector with dense prior," *arXiv preprint arXiv:2104.01318*, 2021.
- [14] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse detr: Efficient end-to-end object detection with learnable sparsity," *arXiv preprint arXiv:2111.14330*, 2021.
- [15] F. Li, A. Zeng, S. Liu, H. Zhang, H. Li, L. Zhang, and L. M. Ni, "Lite detr: An interleaved multi-scale encoder for efficient detr," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 18 558–18 567.
- [16] D. Zheng, W. Dong, H. Hu, X. Chen, and Y. Wang, "Less is more: Focus attention for efficient detr," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 6674–6683.
- [17] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*. IEEE, 2024, pp. 1–6.
- [18] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.
- [19] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han *et al.*, "Yolov10: Real-time end-to-end object detection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 107 984–108 011, 2024.
- [20] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.
- [21] Y. Peng, H. Li, P. Wu, Y. Zhang, X. Sun, and F. Wu, "D-fine: Redefine regression task in detrs as fine-grained distribution refinement," *arXiv preprint arXiv:2410.13842*, 2024.
- [22] S. Huang, Z. Lu, X. Cun, Y. Yu, X. Zhou, and X. Shen, "Deim: Detr with improved matching for fast convergence," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 15 162–15 171.