Improved RT-DETR Based on MobileNetV4 for Vehicle Detection

Yang Zhang
Glasgow College
University of Electronic Science and Technology of China
Chengdu, China
2021190905018@std.uestc.edu.cn

***Abstract*—In this study, we present an enhanced version of the RT-DETR model integrated with MobileNetV4, designated as RT-DETR-MobileNetV4-Small/Medium, tailored for efficient vehicle detection. This model enhances computational efficiency while retaining high detection accuracy, making it ideal for deployment in resource-constrained environments. Key innovations include a streamlined backbone architecture using MobileNetV4 that significantly reduces parameter count and computational demands. Extensive evaluations on a COCO-format vehicle dataset demonstrate the model's superior performance. Specifically, our RT-DETR-MobileNetV4-Small achieves accuracy comparable to, and in some cases superior to, other RT-DETR models while utilizing only 11M parameters and 38 GFLOPs. This represents a reduction of 65% in parameters and 75% in computational complexity compared to RT-DETR-L, and a 45% reduction in parameters and 54% reduction in GFLOPs compared to RT-DETR R18. These results highlight the potential of RT-DETR-MobileNetV4 in applications requiring real-time processing on mobile and edge devices, offering a promising solution for advanced object detection tasks in dynamic environments.**

Keywords- RT-DETR, MobileNetV4, Vehicle Detection, Model lightweight

I. INTRODUCTION

In recent years, deep learning technology has profoundly transformed the field of computer vision, particularly in object detection. Traditional algorithms such as Faster R-CNN and the YOLO series have demonstrated outstanding performance across various standard datasets but typically require substantial computational resources [1][2]. These demands limit their application in resource-constrained environments, such as mobile devices.

Recently, the RT-DETR model, an optimization of the traditional DETR built on the Transformer architecture, has been developed to enhance real-time object detection capabilities [3]. RT-DETR, with its streamlined network structure and more efficient processing flow, reduces the need for high computational resources. This model retains the advantages of DETR's end-to-end design, eliminating the necessity for manually designed anchors and complex post-processing steps, while significantly reducing computational costs in real-time applications through the introduction of more efficient encoder and decoder structures [4][5][6].

Although the original RT-DETR has made progress in reducing computational resource consumption, its performance and efficiency in extremely resource-limited environments, such as on mobile devices, still need enhancement [4][9]. To address this issue, we propose an improved version of RT-DETR, which incorporates a further light-weight design based on MobileNetV4. This modification not only preserves the advantages of the original model but also enhances the model's operational speed and accuracy through innovative adjustments to the network architecture and an optimized loss function. Moreover, our enhancements include specific fine-tuning of the model to meet diverse real-time processing requirements and hardware configurations.

We validated the effectiveness of the improved RT-DETR model through experiments conducted on vehicle datasets in COCO format with 1000 pictures. Our RT-DETR-MobileNetV4-Small achieves mAP50 of 97.3% and mAP50-95 of 76.3% while utilizing only 11M parameters and 38 GFLOPs. This represents a reduction of 65% in parameters and 75% in computational complexity compared to RT-DETR-L, and a 45% reduction in parameters and 54% reduction in GFLOPs compared to RT-DETR R18.

II. RELATED WORK

A. RT-DETR (Real-time Detection Transformer)

RT-DETR (Real-time Detection Transformer) is a Transformer-based real-time object detection model designed to meet the demands of real-time processing, particularly on resource-constrained devices. This model addresses the high computational cost of the traditional DETR (Detection Transformer) and significantly enhances inference speed without sacrificing accuracy through an improved architectural design [4].

RT-DETR features an efficient hybrid encoder design that reduces computational complexity and increases processing speed by decoupling the interactions and fusion between different feature scales. Additionally, the model incorporates a minimal uncertainty query selection mechanism that optimizes uncertainty during the query initialization process, thus enhancing the accuracy and reliability of object detection [5][6]. This advancement provides RT-DETR with flexible speed adjustment capabilities across various real-time application scenarios, adapting to different user needs without retraining.

Experiments on the COCO dataset demonstrate that RT-DETR maintains low latency while achieving accuracy comparable to or even superior to advanced YOLO detectors. Specifically, RT-DETR achieved over 54% average precision (AP) on the COCO validation set and processed over 100

979-8-3315-3508-7/25/$31.00 ©2025 IEEE

frames per second on the NVIDIA T4 GPU, significantly outperforming other real-time object detection models.

Despite its success in speed and accuracy, RT-DETR, like other Transformer-based detection models, still has room for improvement in handling small-sized objects. Future research may focus on further optimizing the model structure and reducing computational demands to expand its applicability in domains requiring high real-time performance, such as autonomous driving and video surveillance.

B. MobileNetV4

MobileNetV4, the latest generation model designed by Google for mobile devices, aims to provide efficient model designs for the mobile ecosystem. Core features include the introduction of the Universal Inverted Bottleneck (UIB) search block, a unified and flexible structure that merges the Inverted Bottleneck (IB), ConvNext, Feed Forward Network (FFN), and a novel Extra Depthwise (ExtraDW) variant [7]. Additionally, MobileNetV4 incorporates the Mobile MQA, an attention module optimized for mobile accelerators, delivering a significant 39% speedup in inference. To further enhance accuracy, MobileNetV4 also employs a novel distillation technique [8].

The UIB block in MobileNetV4, by integrating two optional depthwise convolutions, improves upon the traditional inverted bottleneck structure [8]. This design not only simplifies the network architecture but also enhances the flexibility in spatial and channel mixing, expands the receptive field, and boosts computational efficiency. The Mobile MQA block is optimized for mobile accelerators by simplifying the multi-head attention mechanism, which reduces memory access requirements, thus enhancing operational intensity and inference speed.

MobileNetV4 utilizes an optimized Neural Architecture Search (NAS) strategy that combines coarse-grained and fine-grained searches, significantly enhancing search efficiency. This approach not only improves the model's universality across various hardware platforms but also ensures its efficiency on mobile CPUs, DSPs, GPUs, and specialized accelerators like the Apple Neural Engine and Google Pixel EdgeTPU [8].

The MobileNetV4 model has been extensively tested across multiple standard datasets, particularly achieving up to 87% accuracy on ImageNet-1K, while operating at a mere 3.8 milliseconds on the Pixel 8 EdgeTPU [8]. These results demonstrate MobileNetV4's exceptional capability in balancing high accuracy with high efficiency.

III. Model Overview

The RT-DETR-MobileNetV4-small/medium network architecture incorporates several innovative modules to achieve efficient and precise object detection. This architecture utilizes a specific version of MobileNetV4 as the backbone network along with a highly optimized detection head, designed to provide a lightweight yet powerful model suitable for a range of applications from server-grade hardware to edge devices.

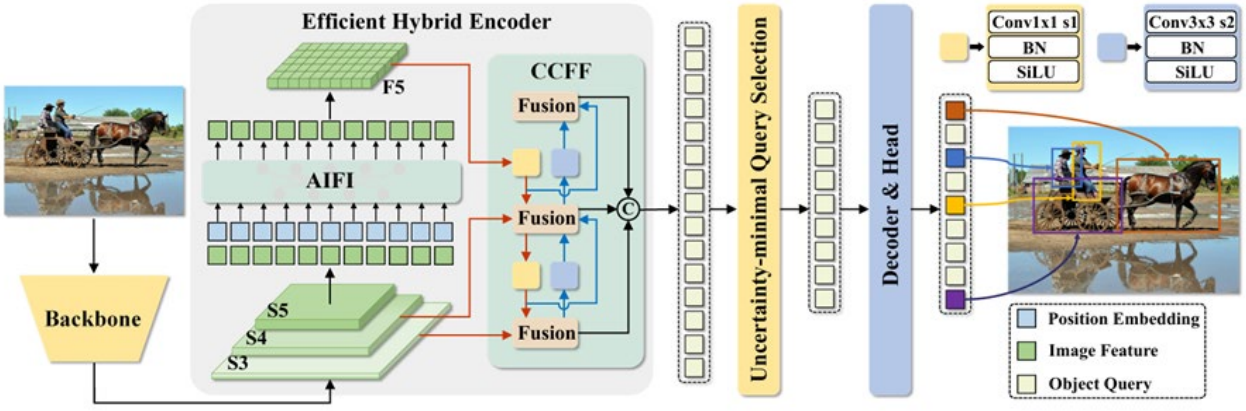

Figure 1. The original backbone of RT-DETR is replaced by MobileNetV4-small/medium [2]

The backbone network employs the MobileNetV4ConvSmall/Medium version, specifically designed for lightweight devices. This choice reflects dual optimization for efficiency and performance, allowing the model to maintain low parameter count and computational demands while providing adequate feature extraction capabilities. The initial layer, starting with MobileNetV4ConvSmall, is tailored for small devices to reduce computational load while ensuring the performance needed for real-time object detection.

According to figure 1.[2] The design of the detection head focuses on combining convolutional layers and up-sampling layers to achieve feature fusion and object detection across different scales. This segment initially projects feature through standard convolutional layers, which are then further processed by the AIFI module and another convolutional layer. Up-sampling layers enlarge the feature map, and a Concat operation fuses features from different layers, aiding the model in capturing multi-scale information from fine to coarse granularity. Additionally, the RepC3 modules, used repeatedly at various stages, refine features and prepare for the detection task. These modules enhance the non-linear expression of features, increasing the accuracy and reliability of the detection.

In the final stages of the detection process, the specially designed RT-DETR Decoder is employed. This decoder is responsible for processing fused features and outputting the final detection results. It integrates information from different feature layers and adjusts the position and size of each detection box to optimize target recognition accuracy.

In summary, the RT-DETR-MobileNetV4-small/medium model represents a critical iteration over the RT-DETR-L by replacing the original backbone network with MobileNetV4ConvSmall. This modification significantly enhances model efficiency and performance, not only optimizing processing speed and accuracy but also enabling the model to meet broader application demands, particularly for real-time processing on resource-constrained devices.

IV. Experiment

A. Datasets and Implementation Details

In our experiments, we employed a COCO-format vehicle dataset consisting of 1,000 images, with 800 designated for the training set and 200 for the validation set. Each image is annotated with vehicle categories, encompassing a variety of vehicle models and complex background scenarios. Our experimental setup included the use of the PyTorch framework for model implementation and training, supported by a 3060 laptop GPU for computational tasks. We utilized the AdamW optimizer for training the model. The resolution of all images was standardized to 640×640. The training of the model was conducted from scratch epochs.

B. Computational Results

TABLE I. Relevant parameters and data

Model	Backbone	#Epochs	#Params(M)	GFLOPs	mAP50	mAP50-95
YOLOv5-L	/	300	46	109	90.7	69.8
YOLOv5-X	/	300	86	205	92.1	72.5
YOLOv8-L	/	/	43	165	94.3	73.9
YOLOv8-X	/	/	68	257	95.9	76.1
DETR-DC5	R50	500	41	187	87.2	67.6
DETR-DC5	R101	500	60	253	88.8	69.1
Anchor-DETR-DC5	R50	50	39	172	88.3	68.9
Conditional-DETR-DC5	R50	108	44	195	89.3	70.1
Conditional-DETR-DC5	R101	108	53	262	90.2	69.8
Efficient-DETR-DC5	R50	36	35	210	90.4	70.3
Efficient-DETR-DC5	R101	36	54	289	91.8	71.7
SMCA-DETR	R50	108	40	152	92.6	73.1
SMCA-DETR	R101	108	58	218	93.4	73.8
Deformable-DETR	R50	50	40	173	95.1	74.8
DINO-Deformable-DETR	R50	36	47	279	96.4	75.9
RT-DETR-L	HGNetv2	72	32	110	95.5	75.5
RT-DETR-X	HGNetv2	72	67	234	97.2	76.2
RT-DETR R18	R18	72	20	59	94.7	75.2
RT-DETR R34	R34	72	30	89	95.5	75.6
RT-DETR R50	R50	72	42	136	96.2	76.1
RT-DETR R101	R101	72	76	259	97.9	77.1
RT-DETR MNv4-Medium	MobileNetV4-Medium	**96**	**17**	**38**	**98.4**	**78.7**
RT-DETR MNv4-Small	MobileNetV4-Small	**96**	**11**	**27**	**97.3**	**76.3**

According to Table 1. The research results of the RT-DETR MobileNetV4 Small/Medium versions underscore their unique position in the modern object detection domain, particularly excelling in model size, computational efficiency, and detection accuracy.

1) Model Size

The RT-DETR MobileNetV4 is designed to be both efficient and lightweight. Taking the RT-DETR MNv4-Small as an example, it has only 11M parameters, approximately 84% less than other high-performance models such as the YOLOv8-X, which has 68M parameters. This significant reduction in parameter size brings notable advantages in terms of storage occupation, computational resource consumption, and energy efficiency, making it particularly suitable for resource-constrained environments. For instance, in mobile or embedded systems, a smaller model means faster download speeds, lower memory requirements, and quicker startup times, which are

crucial for applications that require rapid deployment and frequent updates. Moreover, although the RT-DETR MNv4-Medium has a relatively larger parameter count of 17M, it still represents a reduction of about 78% compared to the RT-DETR R101's 76M parameters. This demonstrates that even while aiming for higher detection accuracy, the model design maintains a high degree of parameter efficiency.

2) Computational Efficiency

In terms of computational efficiency, the RT-DETR MobileNetV4 series exhibits excellent performance optimization. For example, the RT-DETR MNv4-Medium has 38 GFLOPs, substantially lower than the 279 GFLOPs of the DINO-Deformable-DETR, reducing computational complexity by nearly 87%. This significant reduction not only makes the model more suitable for real-time applications such as video surveillance and real-time traffic management systems but also greatly lowers the demands on processors, thus reducing energy consumption and extending device operational times. Lower GFLOPs mean that the model can operate on a wider range of devices, including those with less powerful CPUs, without sacrificing performance. The RT-DETR MNv4-Small further reduces GFLOPs to 27, the lowest among all models compared, making it an ideal choice for operation on edge devices.

3) Detection Accuracy

Accuracy is a core metric for assessing the performance of object detection models. In the mAP50 index, the RT-DETR MNv4-Medium performs exceptionally well with a score of 98.4%, surpassing all other models including the high-performance YOLOv8-X, which scores 95.9%. This indicates a clear advantage of the RT-DETR MNv4-Medium in detection accuracy, particularly in applications that require precise object localization, such as pedestrian and vehicle detection in autonomous vehicles. Additionally, its mAP50-95 score of 78.7% is also the best-performing, reflecting the model's robustness across various IoU thresholds. The performance of the RT-DETR MNv4-Medium in this range confirms its superior detection consistency and reliability. Although the RT-DETR MNv4-Small has smaller parameters and computational requirements, with mAP50 and mAP50-95 scores of 97.3% and 76.3%, respectively, it still demonstrates exceptional detection capabilities, making it highly suitable for applications that demand rapid and accurate feedback.

C. Comparison between F1-confidence curve

According to figure 2. F1-Confidence Curve analysis of the RT-DETR MobileNetV4 Small model demonstrates that at a confidence threshold of 0.330, it achieves its highest F1 score of 0.97, indicating an excellent balance of precision and recall at lower confidence levels. This highlights the model's high sensitivity and accuracy in recognizing target categories. The model maintains high F1 values up to a confidence threshold of 0.8, with a sharp decline occurring only as it approaches 1.0. This shows the model's good stability across a wide range of confidence levels. In contrast, the RT-DETR-L reaches its highest F1 score of 0.96 at a confidence threshold of 0.493, requiring higher confidence to achieve the optimal balance of precision and recall, and its F1 score begins to decline earlier, indicating a higher sensitivity to confidence level adjustments.

In terms of performance comparison and superiority, the RT-DETR MobileNetV4 Small exhibits high flexibility in confidence threshold settings, achieving high F1 scores even at lower confidence levels. This indicates better adaptability to data fluctuations and variations in real applications, which is particularly crucial in real-time or dynamically changing environments such as traffic monitoring or mobile surveillance devices. From the maximum F1 scores, the RT-DETR MobileNetV4 Small slightly leads RT-DETR-L, providing users with a more reliable detection tool for the "Car" category.

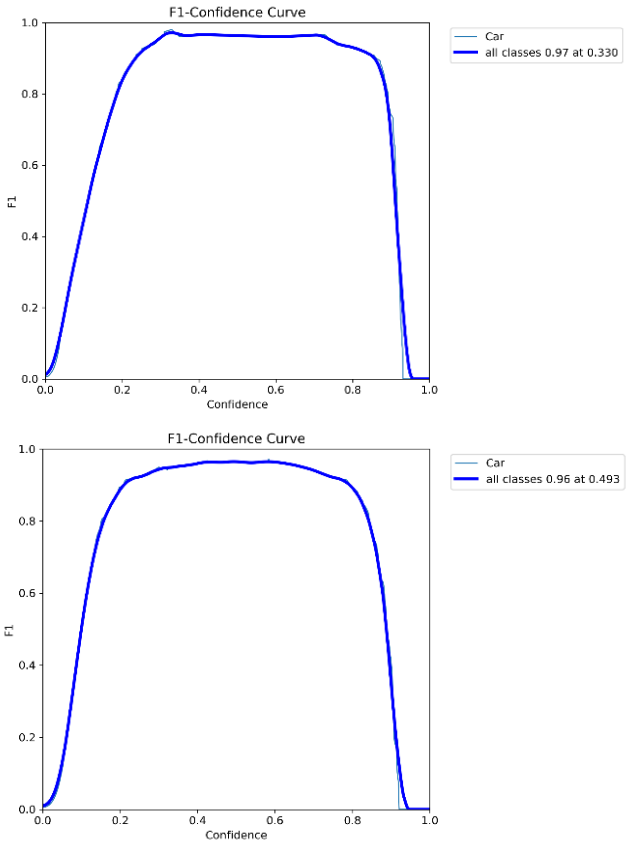

Figure 2. F1-confidence curve of RT-DETR MobileNetV4 and RT-DETR-L

D. Comparison between Metrics

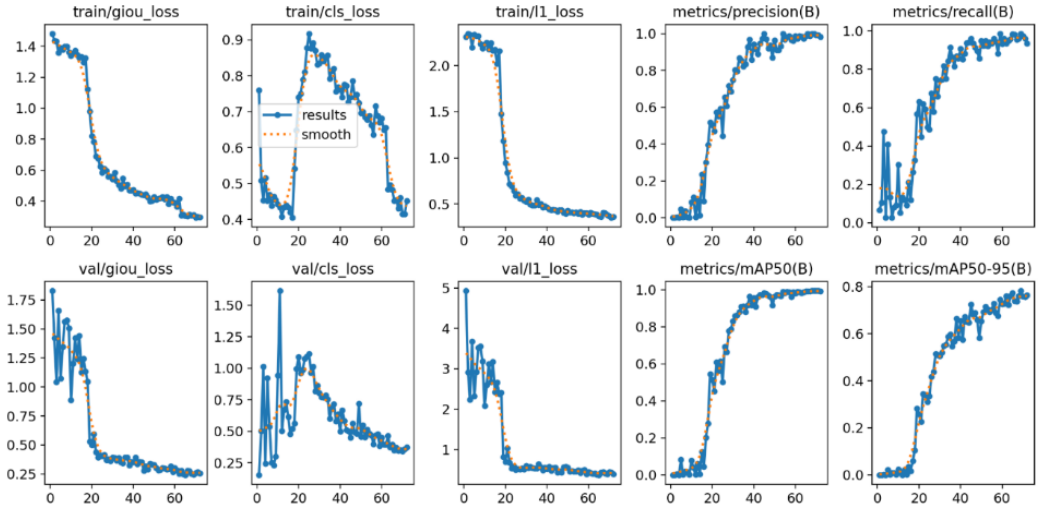

Figure 3. Metrics of RT-DETR MobileNetV4-Small

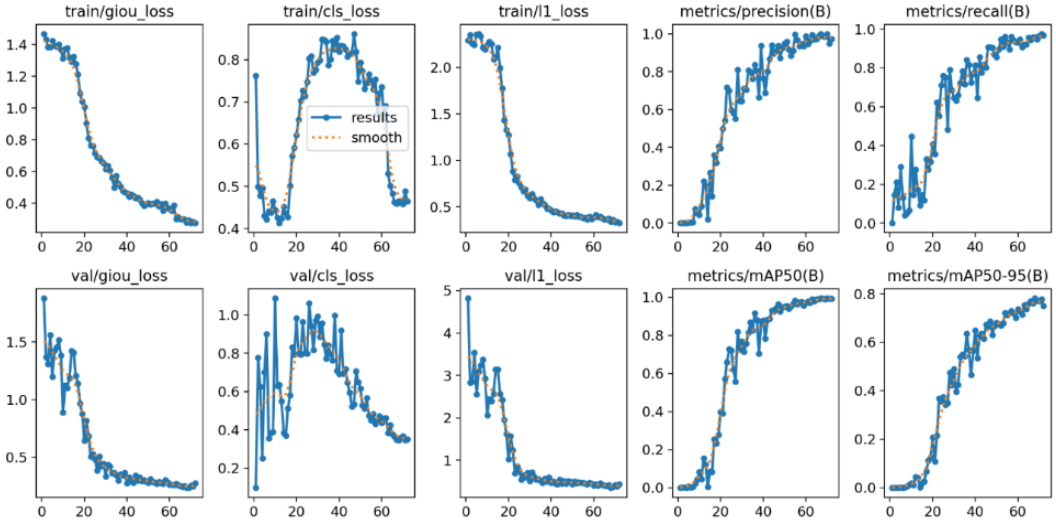

Figure 4. Metrics of RT-DETR-L

1) Training and Validation Loss Comparison.

According to figure 3 and figure 4 RT-DETR-MobileNetV4-Small model exhibits faster convergence and lower steady-state losses in both training and validation GIoU loss metrics compared to the RT-DETR-L. This suggests that the MobileNetV4-small architecture is more precise in geometric matching, enabling better alignment between predicted bounding boxes and ground-truth, which is critical for the accuracy of the model.

In terms of classification loss, RT-DETR-MobileNetV4-Small demonstrates a notably more stable and lower loss curve. This reflects higher efficiency and accuracy in classification tasks within the object detection process. Lower classification loss directly correlates with the model's ability to accurately determine object categories, which is essential for reliable detection outcomes.

The RT-DETR-MobileNetV4-Small model also shows faster convergence and lower losses in L1 loss, which pertains to localization. This further indicates its superior capability in precise object localization, ensuring that detected objects are accurately positioned within the image.

2) Performance Metrics (Precision and Recall).

Throughout the training phases, RT-DETR-MobileNetV4-small achieves higher precision, particularly in the later stages, indicating more reliable positive predictions. High precision means that a greater proportion of positive identifications made by the model are correct, which is vital for applications where accuracy is critical.

RT-DETR-MobileNetV4-small also exhibits higher recall rates, indicating its ability to detect a larger number of actual positives. High recall is crucial in applications to minimize the risk of missing true positive detections, especially in security-sensitive scenarios.

The model consistently maintains a lead in mAP and mAP50-95 metrics, showcasing its balanced performance across different IoU thresholds. This balanced performance

ensures that RT-DETR-MobileNetV4-small is effective and accurate in handling objects of various sizes and difficulty levels.

E. Comparison between Detection

Figure 5. RT-DETR-MobileNetV4-Small

Figure 6. RT-DETR R18

Figure 5 displays the results from RT-DETR-MobileNetV4-Small, where a single detection bounding box is depicted around a vehicle with a high confidence score of 0.9. This indicates that RT-DETR-MobileNetV4-Small is capable of precisely recognizing vehicles with high confidence, without generating any superfluous detection boxes. This aspect is particularly crucial in object detection systems as it reduces the possibility of false positives, thereby enhancing the system's overall efficiency and accuracy.

In contrast, Figure 6 shows the results from RT-DETR R18, where multiple detection errors occurred. The presence of multiple detection boxes with low confidence not only reflects uncertainty in detection but may also lead to confusion and inefficiency in subsequent processing stages. This is especially problematic in applications that require rapid and accurate responses, such as the environmental perception systems in autonomous vehicles.

V. Conclusion

In this study, we have introduced an enhanced RT-DETR model, termed RT-DETR-MobileNetV4-Small/Medium, which is specifically optimized for lightweight and efficient operation, making it particularly suited for deployment in resource-constrained environments. This model integrates a streamlined backbone architecture utilizing MobileNetV4, significantly reducing the parameter count and computational demands.

Our model's key innovation lies in its lightweight design, which significantly lowers computational requirements while maintaining competitive detection accuracy, thus addressing the pressing need for efficient processing in devices with limited computational resources. The RT-DETR-MobileNetV4-Small model, in particular, has demonstrated its capability to achieve comparable or superior accuracy to other advanced models while using substantially fewer parameters and GFLOPs. This represents a reduction of 65% in parameters and 75% in computational complexity compared to RT-DETR-L, showcasing its efficiency.

However, the study encounters limitations due to hardware constraints that restrict the usage of larger and potentially more informative datasets, possibly impacting the generalizability and robustness of the model. Future work will focus on overcoming these hardware limitations to enhance model training and evaluation, potentially incorporating larger datasets to validate the model's effectiveness across more diverse scenarios. The ultimate aim is to refine our lightweight model to deliver not only high accuracy and efficiency but also robustness across various operational contexts, which is critical for real-world applications.

References

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.

[2] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458.

[3] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

[4] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco[gnition (pp. 16965-16974).

[5] Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., & Liu, Y. (2024). RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer. arXiv preprint arXiv:2407.17140.

[6] Wang, S., Xia, C., Lv, F., & Shi, Y. (2024). RT-DETRv3: Real-time End-to-End Object Detection with Hierarchical Dense Positive Supervision. arXiv preprint arXiv:2409.08475.

[7] Sinha, D., & ElSharkawy, M. (2019, October). Thin mobilenet: An enhanced mobilenet architecture. In 2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON) (pp. 0280-0285). IEEE.

[8] Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., ... & Howard, A. (2024). MobileNetV4-Universal Models for the Mobile Ecosystem. arXiv preprint arXiv:2404.10518.

[9] Wang, Z., Zhan, J., Duan, C., Guan, X., Lu, P., & Yang, K. (2022). A review of vehicle detection techniques for intelligent vehicles. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3811-3831.