Improving YOLOv5 for the KITTI Dataset with Ghost Convolutions and Coordinate Attention

Vyom Sagar Shikhar Kapoor
University of Alabama at Birmingham
{vsagar@uab.edu,skapoor2@uab.edu}

Abstract

Object detection is a core computer vision task in which an algorithm identifies and localizes objects within an image or video. In applications such as autonomous vehicles, surveillance systems, and medical imaging, detection models must be both accurate and computationally efficient. This work investigates whether the YOLOv5s architecture can be improved by integrating lightweight convolution modules and attention mechanisms. Specifically, we study the effects of replacing standard convolutions with Ghost Modules and injecting Coordinate Attention into the backbone.

We evaluate four model variants (baseline YOLOv5s, Ghost-only, CA-only, and a combined Ghost+CA model) on the KITTI dataset. Our analysis focuses on performance trade-offs, transfer-learning compatibility, parameter reductions, and accuracy impacts. Ghost Modules substantially reduce parameters and GFLOPs with minimal accuracy degradation, while Coordinate Attention enhances spatial localization but complicates pretrained weight transfer. The hybrid Ghost+CA model achieves strong computational efficiency while retaining competitive detection accuracy. These findings highlight promising directions for resource-efficient object detection in real-world autonomous systems.

1. Introduction

Object detection plays a critical role in perception systems for autonomous vehicles. For real-time decision-making, detectors must be both fast and accurate. YOLO-based models have become a standard for balancing these needs, but further improvements are required for deployment on constrained hardware such as embedded systems and automotive-grade processors.

The KITTI dataset captures real-world driving scenes with diverse lighting, occlusions, and object scales making it a strong benchmark for testing detection robustness. YOLOv5s performs well on KITTI, yet its convolution-heavy backbone and neck limit efficiency.

Our work explores whether two modern architectural enhancements Ghost Modules and Coordinate Attention can reduce computational cost while preserving detection accuracy. Our study is guided by the following key questions:

- **Architectural novelty:** Can lightweight and attention-based modules meaningfully improve YOLOv5s on KITTI?
- **Empirical novelty:** How do these modifications impact transfer learning, convergence behavior, parameter efficiency, and accuracy trade-offs?

To investigate these questions, we implement and evaluate four model variants: the baseline YOLOv5s architecture, YOLOv5s with Ghost Modules, YOLOv5s with Coordinate Attention blocks, and a hybrid model incorporating both enhancements.

We analyze quantitative and qualitative results, emphasizing how architectural changes affect pretrained weight compatibility, model efficiency, and overall detection performance.

2. Related Work

2.1. YOLO Architectures

YOLO-based detectors have evolved to balance speed and accuracy in real-time object detection [12, 13]. YOLOv3–YOLOv7 advanced real-time detection using CSP connections and PANet-style feature fusion [9, 14]. YOLOv5 introduced the SPPF module, enhanced training strategies, and flexible model scaling [3, 7].

2.2. Lightweight Convolutions

Lightweight convolutional modules aim to reduce the computational burden of deep neural networks while maintaining representational capacity. MobileNets [5] popularized depthwise separable convolutions for efficient inference on mobile devices. GhostNet [2] introduced Ghost Modules, which generate intrinsic feature maps using standard convolutions and then expand them using inexpensive linear

operations.

2.3. Attention Mechanisms

Attention mechanisms enhance feature representation by selectively emphasizing informative channels or spatial locations. Squeeze-and-Excitation (SE) networks [6] introduced channel-wise recalibration, while CBAM [15] extended this idea with joint channel and spatial attention. Coordinate Attention (CA) [4] embeds positional information into channel attention, improving long-range dependencies and localization, which is highly relevant for driving scenes with small objects like pedestrians and cyclists.

2.4. KITTI Object Detection

The KITTI vision benchmark suite [1] provides real-world urban driving scenes with challenging illumination, occlusions, and object scales. Research on KITTI typically emphasizes robustness to scale variation and environmental conditions.

3. Methodology

We examine four model variants built upon the YOLOv5s architecture: the baseline model, a Ghost-enhanced model, a Coordinate Attention (CA) model, and a hybrid model combining both modifications.

3.1. Baseline YOLOv5s

The baseline follows the standard YOLOv5s configuration [7], consisting of:

- a CSPDarknet backbone for efficient feature extraction [14],
- a PANet-style neck for multi-scale feature fusion at $P3$, $P4$, and $P5$ [9],
- an SPPF block providing multi-scale receptive fields before the detection head [3].

This model contains approximately 7.03M parameters and requires 15.8 GFLOPs for a 640×640 input.

3.2. Ghost Module Integration

Ghost Modules provide a lightweight alternative to standard convolutions by exploiting redundancy in feature representations [2, 5].

This design reduces both parameter count and computational cost while maintaining representational capacity. In our modified architecture, GhostConv layers replace selected standard convolutions in the backbone.

Benefits:

- Eliminates redundant feature generation,
- Reduces computational load and FLOPs,
- Maintains expressive capacity through intrinsic and generated feature maps.

3.3. Coordinate Attention Integration

Coordinate Attention (CA) enhances spatial feature encoding by embedding positional information directly into channel attention [4].

Benefits:

- Encodes explicit positional information,
- Enhances object localization in cluttered scenes,
- Improves detection of thin, elongated, or small objects.

3.4. Hybrid Ghost + CA Model

The hybrid model integrates both Ghost Modules and Coordinate Attention to combine lightweight feature generation with enhanced spatial focus. While this configuration aims to reduce parameters and improve localization, it also increases architectural deviation from the YOLOv5s baseline, potentially reducing pretrained weight compatibility and affecting convergence.

4. Dataset

4.1. KITTI Overview

The KITTI Object Detection dataset [1] contains real-world street scenes from urban environments. It is widely used for evaluating detection performance under challenging conditions due to:

- **7,481** annotated images,
- **8** object classes: Car, Van, Truck, Pedestrian, Cyclist, Tram, Misc, Person Sitting
- frequent **occlusion** and **truncation**,
- diverse **lighting** and **weather** conditions.

These characteristics make KITTI a strong benchmark for assessing robustness in real-world autonomous driving scenarios.

4.2. Preprocessing and Splits

All images are resized to 640×640, and standard YOLO augmentations (Mosaic, scaling, flipping, color jitter) are applied. The dataset is split into:

- approximately **5,000** training images,
- approximately **1,500** validation images.

Annotations are converted from the KITTI label format into YOLO-style `.txt` files containing class ID and normalized bounding boxes.

4.3. Dataset Statistics

Table 1 summarises the dataset properties used in our experiments.

5. Results and Evaluation

We evaluate four YOLOv5s variants baseline, Ghost-only, CA-only, and the hybrid Ghost+CA model under identical training conditions. Our analysis examines accuracy,

Table 1. Summary of KITTI dataset statistics used in our experiments.

Property	Value
Total images	7,481
Training images	$\approx$ 5,000
Validation images	$\approx$ 1,500
Number of classes	8
Avg. objects per image	$\approx$ 7

Table 2. Comparison of YOLOv5s variants on KITTI. All models are trained for 100 epochs with 640×640 input resolution.

Model	Params	GFLOPs	P	R	mAP_{50}
YOLOv5s (Baseline)	7.03	15.8	0.915	0.863	0.910
Ghost only	6.07	13.9	0.900	0.850	0.908
CA only	7.35	17.1	0.913	0.856	0.908
Ghost + CA (Hybrid)	6.39	15.2	0.906	0.823	0.892

computational efficiency, parameter reduction, pretrained-weight compatibility, and qualitative performance.

5.1. Training Setup

All models are trained for 100 epochs using stochastic gradient descent with warm restarts [10], momentum 0.937, weight decay 5×10^{-4}, and an initial learning rate of 0.01 following a cosine schedule. We use the SiLU activation [11] as in the YOLOv5 implementation [7].

COCO-pretrained weights are loaded for all matching layers [8]. Layers introduced by Ghost or Coordinate Attention modules are randomly initialized. Evaluation is performed using a held-out validation set of 1,500 KITTI images.

5.2. Quantitative Comparison

Table 2 compares parameter count, computational cost, and detection accuracy across the four model variants.

Ghost-only achieves nearly identical mAP_{50} to the baseline while reducing parameters by **14%** and GFLOPs by **12%**. CA-only maintains similar accuracy but adds computational overhead. The hybrid Ghost+CA model is the most parameter-efficient, but its recall and mAP decrease due to reduced pretrained-weight compatibility.

5.3. Accuracy Across Training Epochs

Figure 1 shows the mAP@0.5 curves for all variants over 100 epochs.

Ghost-only converges nearly as fast as the baseline, while CA-only and Ghost+CA models converge more slowly because fewer pretrained layers are reused.

5.4. Transfer Learning Compatibility

Table 3 quantifies model compatibility with COCO-pretrained weights. This trend aligns with findings from transfer-learning research [16].

Table 3. Impact of pretrained-weight compatibility on convergence and accuracy.

Model	Loaded / Total Layers	mAP@0.5
Baseline	342 / 349 (98%)	0.910
Ghost only	294 / 397 (74%)	0.908
CA only	246 / 458 (54%)	0.908
Ghost + CA	198 / 506 (39%)	0.892

Models with lower pretrained-weight reuse exhibit slower convergence and lower final accuracy. The effect is most pronounced for the hybrid architecture.

5.5. Qualitative Evaluation

Qualitative results (Figs. 2–3) demonstrate detection behavior on challenging KITTI scenes.

5.6. Key Insights

The experiments reveal several consistent trends:

- Ghost Modules provide the strongest accuracy–efficiency trade-off.
- Coordinate Attention sharpens localization but lowers pretrained-weight compatibility.
- Ghost+CA yields the highest efficiency but the lowest transferability and accuracy.

This relationship between architectural modification, pretrained-weight reuse, and final accuracy forms a central empirical finding of this work.

6. Conclusion

This work investigated architectural and empirical modifications to the YOLOv5s detector on the KITTI dataset, focusing on lightweight convolutional modules and attention mechanisms. Our findings highlight several key insights:

Ghost Modules offer the strongest balance between accuracy and computational efficiency. They substantially reduce parameter count and GFLOPs while maintaining mAP performance comparable to the baseline. This makes Ghost-enhanced YOLOv5s a promising option for deployment on resource-constrained hardware.

Coordinate Attention improves spatial localization by embedding positional cues into the feature representation. However, its integration disrupts pretrained-weight compatibility, leading to slower convergence and a slight reduction in final accuracy. These results suggest that CA may require additional tuning or deeper architectural integration to realize its full potential.

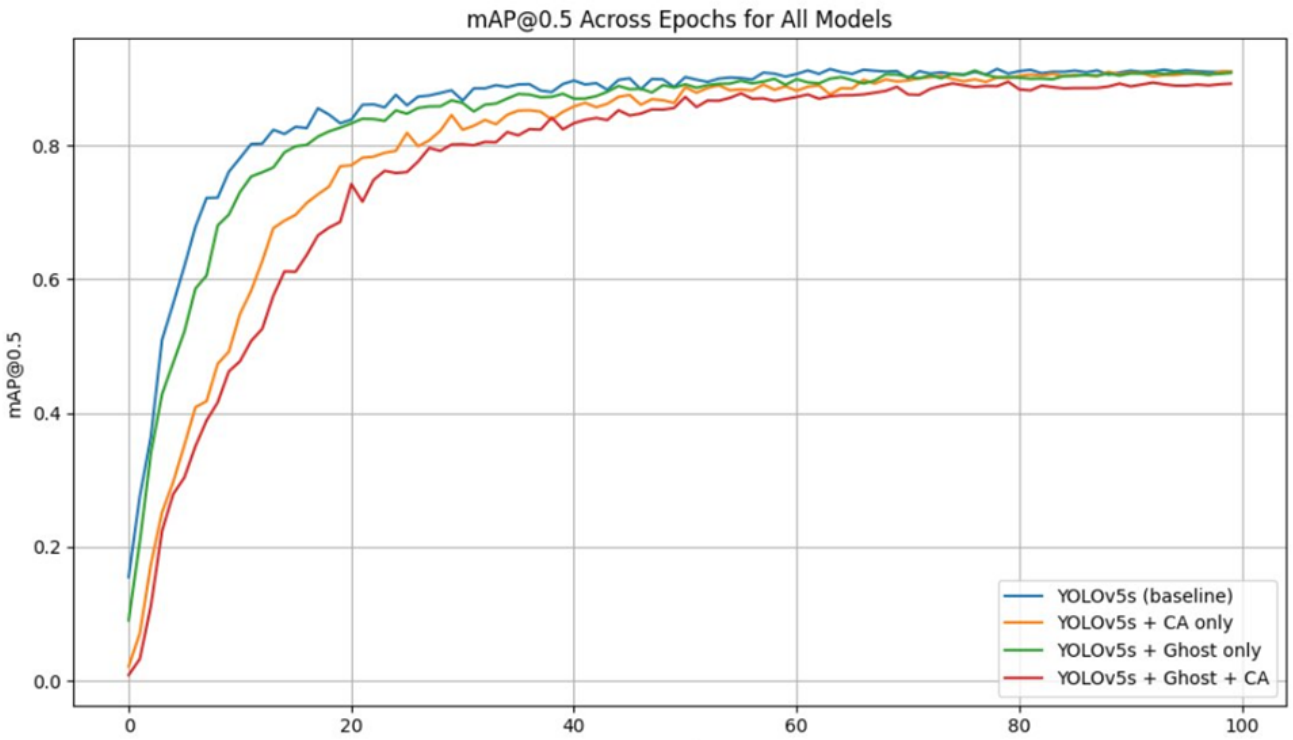

Figure 1. mAP@0.5 across epochs for all model variants. CA-based models converge more slowly due to reduced pretrained-weight alignment.

Figure 2. Qualitative example. Ghost-only closely matches baseline performance; CA-only improves localization on pedestrians and cyclists; Ghost+CA slightly reduces long-range detection confidence.

Figure 3. Additional qualitative comparison illustrating object-scale sensitivity and small-object challenges across models.

Hybrid Ghost + CA Model achieves meaningful parameter reductions and improved efficiency, yet suffers from the lowest pretrained-weight alignment among the variants. This negatively impacts convergence stability and overall detection accuracy, though the model shows promise with improved training strategies or alternative initialization methods.

Future Work

Future directions include:

- applying knowledge distillation to mitigate accuracy loss in CA-based models,
- evaluating real-world deployment on embedded devices such as Jetson Nano or Xavier,
- exploring additional attention mechanisms (e.g., CBAM, ECA, SimAM),
- investigating deeper integration of CA within PANet or SPPF modules.

Real-World Implications

Lightweight detectors particularly the Ghost-only YOLOv5 variant enable practical real-time perception on embedded and mobile platforms, including autonomous vehicles, UAVs, and robotic systems. These results demonstrate that carefully designed lightweighting strategies can deliver competitive accuracy while meeting stringent computational constraints.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pages 3354–3361, 2012. 2

[2] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chang Xu, and Enhua Xu. Ghostnet: More features from cheap operations. In *CVPR*, pages 1580–1589, 2020. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI*, 37(9):1904–1916, 2015. 1, 2

[4] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *CVPR*, pages 13713–13722, 2021. 2

[5] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 1, 2

[6] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 2

[7] Glenn Jocher et al. YOLOv5 by ultralytics. https://github.com/ultralytics/yolov5, 2020. Accessed: 2025-02-01. 1, 2, 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 3

[9] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768, 2018. 1, 2

[10] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3

[11] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 3

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1

[14] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hsuan Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *CVPRW*, pages 390–391, 2020. 1, 2

[15] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 2

[16] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, pages 3320–3328, 2014. 3