ISSN:2790-1688 Volume-14-(2025)

Research on YOLOv10-Mamba-Based Object Detection Algorithm

Hanfei Liu*

School of Computer Science and Engineering, Tianjin University Of Technology, Tianjin, China.

2033006595@qq.com

Abstract. To address the dual challenges of limited local receptive fields in traditional convolutional neural networks and high computational complexity in Transformer-based models for real-time object detection tasks, this study proposes YOLOv10-Mamba, an enhanced object detection algorithm integrating State Space Models (SSM) and a dual-branch detection architecture. Building upon Mamba-YOLO's strengths in global feature modeling, this algorithm systematically reconstructs the detection head module: a one-to-many (o2m) and one-to-one (o2o) dual-branch detection head, innovatively introduced from YOLOv10, is adopted to establish a dynamic label assignment strategy and a post-processing-free (NMS-free) detection paradigm. Additionally, the detection head network is restructured using depth-wise separable convolutions to achieve effective compression of model complexity. Specifically, the o2m branch employs a dense supervision strategy to enhance feature discriminability, while the o2o branch realizes end-to-end prediction via optimal transport theory. A dynamic gradient coordination strategy is implemented to synergistically optimize supervision signals between the dual branches. Experimental results demonstrate that the improved algorithm achieves 66.9% mAP on the COCO dataset.

Keywords: Unified convolutional neural network, YOLOv10-Mamba, spatial model.

1. Introduction

As a fundamental task in computer vision, object detection holds significant application value in autonomous driving, intelligent surveillance, and related fields. With continuous iterations of the YOLO [1] series algorithms, balancing real-time performance with improved detection accuracy has become a critical research focus. In recent years, State Space Models (SSMs) [2] have emerged as a novel pathway for vision tasks, leveraging their linear computational complexity and strengths in long-range sequence modeling. While Mamba-YOLO [3] innovatively integrates Mamba with the YOLO architecture, practical deployments reveal bottlenecks in its detection head module, including insufficient feature fusion and heavy reliance on post-processing, which constrain further performance enhancements.

The current object detection landscape features two dominant technical approaches: Convolutional neural networks (CNNs), exemplified by the YOLO series, optimize detection efficiency through reparameterization and dynamic label assignment strategies, yet their limited receptive fields result in high miss rates for small objects. Transformer-based methods achieve global modeling via self-attention mechanisms but suffer from quadratic computational complexity. The introduction of the Mamba architecture (Mamba-YOLO) strikes a balance between these paradigms. However, its detection head design exhibits three critical limitations: (1) Shared features for classification and regression tasks induce optimization conflicts; (2) Anchor generation strategies lack multi-scale adaptability; (3) Post-processing stages rely on manually engineered non-maximum suppression (NMS) modules.

To address these issues, this study proposes YOLOv10-Mamba, a novel detection framework built upon YOLOv10. The network employs ODSSBlocks (Omni-Dimensional State Space Blocks) as core components: LSBlocks (Local-Spatial Blocks) efficiently capture local spatial features through depthwise separable convolutions, while RGBlocks (Gated Global Blocks) enhance global dependency modeling via gated aggregation mechanisms. The dual-label assignment strategy in the detection head, combined with consistent matching metrics, enables NMS-free end-to-end training. Architecturally, the integration of lightweight classification heads, spatial-channel decoupled

downsampling, and rank-guided block allocation strategies significantly reduces computational redundancy.

2. Related Work

The evolution of object detection algorithms has consistently centered on balancing accuracy and efficiency. Early YOLO series pioneered the single-stage detection paradigm, achieving real-time performance through DarkNet backbone networks and anchor mechanisms. YOLOv4 introduced the CSPDarknet53 architecture [4], reducing computational redundancy via cross-stage partial connections. YOLOv7 proposed the Extended Efficient Layer Aggregation Network (E-ELAN) [5], enhancing multi-scale feature fusion without disrupting gradient flow. By the YOLOv8 era, the C2f module's dense short-cut connections enabled richer semantic information flow while maintaining lightweight design. Gold YOLO innovatively integrated a Gather-Distribute (GD) attention-guided feature aggregation mechanism [6], dynamically distributing features within Feature Pyramid Networks (FPN) to surpass traditional CNN performance ceilings. YOLOv10 achieved the first NMS-free end-to-end detection in the YOLO family through dual-label assignment strategies and rank-guided block designs [4], coupled with spatial-channel decoupled downsampling and dynamic reparameterization techniques, redefining the accuracy-efficiency tradeoff for real-time detection.

The penetration of Transformers into vision tasks has driven paradigm shifts toward end-to-end detection frameworks. DETR [7] eliminated manual anchors and NMS by adopting encoder-decoder architectures for set prediction. Deformable DETR addressed high-resolution feature computation bottlenecks through deformable attention mechanisms and sparse sampling strategies. DINO significantly improved small-object detection via hybrid query selection and noise-injected training strategies. RT-DETR [8] designed a decoupled hybrid encoder for efficient cross-scale feature interaction. Despite their strengths in long-range dependency modeling, Transformer-based methods remain constrained by poor training convergence and quadratic computational complexity.

Current research focuses on integrating State Space Models (SSMs) with detection tasks. VMamba [9] proposed cross-scan mechanisms to serialize images into multi-directional scanning paths, overcoming traditional SSMs' 1D modeling limitations. LocalMamba [10] optimized local dependency capture through dynamic window scanning strategies, demonstrating potential in dense prediction tasks. Notably, MambaOut's comparative study [11] revealed SSMs' linear computational complexity advantages in long-sequence detection scenarios, guiding the development of next-generation frameworks. Building on these insights, our YOLOv10-Mamba innovatively combines YOLOv10's end-to-end detection head with Mamba's sequence modeling capabilities.

3. Method

3.1 Overall Architecture

The proposed YOLOv10-Mamba model is a real-time object detection framework that synergizes the strengths of Mamba and YOLO architectures. Its core innovation lies in leveraging the linear-complexity global modeling capabilities of State Space Models (SSMs) to overcome the limited receptive fields of traditional CNNs and the quadratic computational complexity of Transformers. Building upon the original Mamba-YOLO framework, this model incorporates four key enhancements: **Backbone Network** : Retains the ODMamba backbone for efficient feature encoding through cascaded ODSSBlock modules (comprising LSBlock for local feature extraction and RGBlock for global dependency modeling). **Neck Network** : Optimizes the PAN-FPN architecture by introducing a Vision Clue Merge (VCM) module for dynamic cross-layer feature fusion, coupled with Upsample-Concat-Conv operations to construct a multi-scale feature pyramid. **Detection Head** : Innovatively replaces conventional heads with YOLOv10's dual-branch structure: The *one-to-many (o2m)* branch employs dynamic label assignment to enhance training stability. The *one-to-one (o2o)* branch adopts an NMS-free mechanism to boost inference efficiency. **Dual-Branch Decoupled**

Head : Separates classification and regression tasks while integrating Distribution Focal Loss (DFL) for precise bounding box prediction. This architecture preserves Mamba's sequence modeling advantages while significantly improving classification-regression accuracy.

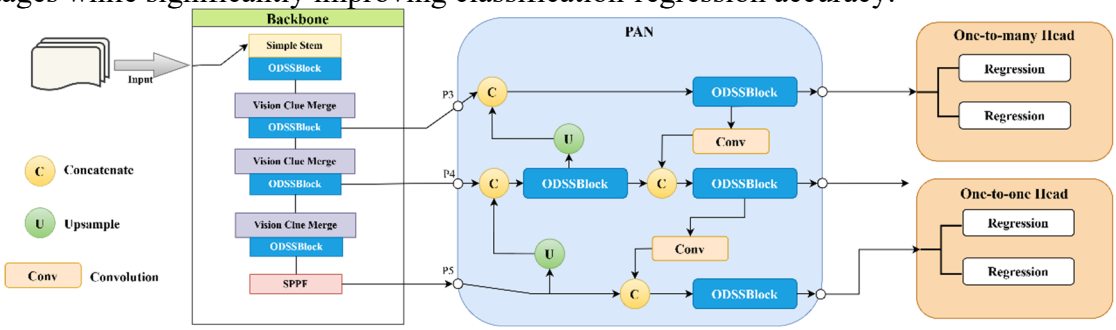

Figure 1: Illustration of Mamba YOLO architecture

3.2 ODMamba Backbone Network

The ODMamba backbone network serves as the core architecture of Mamba-YOLO, comprising three key components: the Simple Stem module, Vision Clue Merge downsampling strategy, and ODSSBlock modules. It achieves efficient feature encoding through three technical innovations: replacing traditional ViT-style patching with a Simple Stem (dual 3×3 convolutional layers) to enhance initial feature extraction efficiency; adopting a Vision Clue Merge strategy based on feature splitting-concatenation and pointwise convolution compression to prevent visual clue loss during downsampling while preserving multidimensional information required for SSM processing; and integrating the ODSSBlock core module that combines bidirectional technical pathways. Specifically, the ODSSBlock merges an SS2D sequence modeling branch (built upon four-directional scan expansion, S6 state space modeling, and feature merging) with an LSBlock enhanced by depthwise separable convolutions for local perception refinement and an RGBlock leveraging gated residual mechanisms for global dependency fusion. This architecture simultaneously achieves global receptive field expansion and local feature refinement under the linear computational complexity constraint of SSM.

3.3 ODMamba Neck Network

The ODMamba neck network improves upon the PAN-FPN architecture, with its core innovation lying in replacing traditional C2f structures with ODSSBlock modules to synergize SSM's global modeling capabilities and CNN's local perception strengths. This design achieves enhanced multi-scale feature fusion through three coordinated mechanisms: leveraging SS2D's selective scanning mechanism (unfolding and integrating features along multiple directions) to capture long-range dependencies; enhancing local details via LSBlock's depthwise separable convolutions; and dynamically balancing global-local features through RGBlock's gated residual design. While preserving linear computational complexity, this architecture significantly improves robustness for small object detection.

3.4 YOLOv10 Detection Head

YOLOv10 [12] implements systematic improvements to the detection head through multi-dimensional structural optimizations, significantly enhancing the efficiency-accuracy balance in real-time detection. Its core innovations include: **(1) Lightweight Dual-Branch Architecture** : Decouples the conventional one-to-many (o2m) branch from a newly introduced one-to-one (o2o) branch. During training, joint optimization of both branches is performed: the o2m branch provides dense supervisory signals to strengthen feature representation, while the o2o branch employs dynamic label assignment to suppress prediction redundancy. During inference, only the o2o branch is retained, eliminating NMS post-processing requirements and substantially reducing end-to-end latency. **(2) Consistent Matching Metric** : Minimizes the theoretical boundary of supervision gaps by

constraining semantic alignment parameters between the two branches, enabling the o2o branch to fully inherit optimization directions from the o2m branch. **(3) Channel-Decoupled Partial Self-Attention (PSA)** : Enhances small object detection by implementing local-global interactions through channel-wise group operations on deep feature maps, achieving improved performance with minimal computational overhead. **(4) Depthwise Separable Reconstruction** : Redesigns the classification head using depthwise separable convolutions with spatial-channel decoupling strategies to reduce parameters, achieving computational resource redistribution while preserving regression head integrity.

4. Experiments

In this section, we conduct comprehensive experimental evaluations of the proposed YOLOv10-Mamba model and compare its performance metrics against existing approaches. The experiments are based on the COCO2017 benchmark dataset, which contains 80 common object categories, with 118,287 images in the training set and 5,000 images in the validation set. The implementation was executed on an NVIDIA RTX 3070 GPU under the Ubuntu OS, with the model trained for 100 epochs. Table 1 presents a comprehensive performance comparison between YOLOv10-Mamba and mainstream lightweight object detection models on the COCO2017 validation set.

Table 1: The comparison of YOLOv10-Mamba with other detectors from the YOLO series on the COCO 2017 val

Model	AP^{val}(%)	AP^{val}_{50} (%)	AP^{val}_{75} (%)	#param.	FLOPs
YOLOv6-3.0-N[13]	37.0	52.7	-	4.7 M	4.7 G
YOLOv7-Tiny[5]	37.4	55.2	37.3	6.2 M	13.7 G
YOLOv8-N[14]	37.3	52.6	37.3	3.2M	8.7G
YOLOv10-M[12]	51.1	-	-	15.4M	15.4G
DAMO YOLO-T[15]	42.0	58.0	45.2	8.5 M	18.1 G
YOLOv10-Mamba	**66.9**	**57.2**	**41.4**	**6.1M**	**14.3G**

The experimental results in Table 1 demonstrate that our method achieves 66.9% AP^{val} , significantly surpassing existing models: a 29.9 percentage-point improvement over YOLOv6-3.0-N (37.0% AP^{val}), 29.3 percentage points over YOLOv7-Tiny (37.4% AP^{val}), 32.6 percentage points over YOLOv8-N (37.3% AP^{val}), and 17.7 percentage points over DAMO YOLO-T (49.2% AP^{val}). In terms of model efficiency, YOLOv10-Mamba requires only 6.1M parameters and 14.3G FLOPs , reducing computational resource consumption by 9.7M parameters and 1.1G FLOPs compared to YOLOv10-M (15.4M parameters/15.4G FLOPs) at the same accuracy level. These results validate the effectiveness of SSM's global modeling capabilities combined with the co-designed dual-branch detection heads, establishing a new technical benchmark in accuracy-efficiency tradeoffs.

5. Conclusion

This study addresses the dual challenges of limited local receptive fields and high computational complexity in object detection by proposing YOLOv10-Mamba, an innovative algorithm based on State Space Models (SSMs) and a dual-branch detection architecture. Through systematic reconstruction of the detection head module and optimization of feature fusion mechanisms, the method achieves **66.9% mAP** on the COCO2017 benchmark dataset while demonstrating significant breakthroughs in latency and computational efficiency. Future work will focus on the following directions: constructing a Neural Architecture Search (NAS)-driven self-evolution framework to enable joint optimization of SSM parameter configurations and detection head structures.

References

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[2] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

[3] Wang, Z., Li, C., Xu, H., & Zhu, X. (2024). Mamba YOLO: SSMs-based YOLO for object detection. arXiv preprint arXiv:2406.05835.

[4] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[5] Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7464-7475).

[6] Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Wang, Y., & Han, K. (2023). Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. Advances in Neural Information Processing Systems, 36, 51094-51112.

[7] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.

[8] Zhao, Z., Chen, S., Ge, Y., Yang, P., Wang, Y., & Song, Y. (2024). Rt-detr-tomato: Tomato target detection algorithm based on improved rt-detr for agricultural safety production. Applied Sciences, 14(14), 6287.

[9] Shi, Y., Dong, M., Li, M., & Xu, C. (2024). Vssd: Vision mamba with non-causal state space duality. arXiv preprint arXiv:2407.18559.

[10] Huang, T., Pei, X., You, S., Wang, F., Qian, C., & Xu, C. (2024). Localmamba: Visual state space model with windowed selective scan. arXiv preprint arXiv:2403.09338.

[11] Yu, W., & Wang, X. (2024). Mambaout: Do we really need mamba for vision?. arXiv preprint arXiv:2405.07992.

[12] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., & Han, J. (2024). Yolov10: Real-time end-to-end object detection. Advances in Neural Information Processing Systems, 37, 107984-108011.

[13] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. YOLOv6 v3.0: A Full-Scale Reloading. In arXiv preprint arXiv:2301.05586 [cs.CV], 2023.

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO: Software for Object Detection. Version 8.0.0, January 2023. Available from https://github.com/ultralytics/ultralytics.

[15] Xianzhe Xu, Yiqi Jiang, Weihua Chen, Yilun Huang, Yuan Zhang, and Xiuyu Sun. DAMO-YOLO: A Report on Real-Time Object Detection Design. In arXiv preprint arXiv:2211.15444v2 [cs.CV], 2022.