ISSN: 2664-9640

A Survey on Pruning Methods and Attention Mechanisms for YOLOv8-Based Models

Zhexu Zhang[a]

College of Electronic Information and Communications, Beijing University of Technology, Beijing 100124, China

[a]Email: 1320231055@emails.bjut.edu.cn

Abstract

The C2f module in YOLOv8 enhances multi-scale feature fusion, yet parameter redundancy still hinders its deployment on edge devices. Current model compression techniques often overlook two critical issues: (i) topological mismatch among branched network paths, and (ii) lack of synergy between pruning criteria and distillation loss functions. This survey systematically reviews recent YOLOv8 lightweight adaptation efforts, critically analyzing limitations of existing structured pruning, attention integration, and knowledge distillation methods from two perspectives: topological coupling of network architecture and coordination of multi-objective optimization. Building upon the internal mechanics of the C2f module, we formally formulate a path-level channel dependency graph and propose a joint evaluation metric that integrates channel importance with distillation sensitivity. This provides a viable technical pathway toward lossless, high-ratio compression of YOLOv8.

Keywords

YOLOv8; Structured Pruning; Attention Mechanism; Knowledge Distillation; C2f Module.

1. Introduction

Table 1. Performance comparison of mainstream YOLOv8 lightweight adaptation methods (validated on public datasets)

Method	Param. Compression	△mAP50	△FLOPs	Dataset	Key Deficiency
LAMP[4]	65.3%	↑2.2%	↓29.6%	KITTI	Ignores C2f path dependency; magnitude ranking sensitive to BN statistics
CWD+Pruning[3]	73.5%	↓2.7%	↓73.2%	VisDrone	Distillation temperature requires manual tuning; single-dataset validation
SimYOLO[5]	0%	↑1.1%	Negligible	NEU-DET	SimAM hyperparameter λ fixed; compression potential unexplored; accuracy gain dataset-dependent
BN pruning + INT8 [1]	64.2%	↓4.2%	—	Custom	Large accuracy loss;no distillation compensation

Note: ↑ denotes improvement over baseline; ↓ denotes degradation or resource reduction. SimYOLO [5] is an accuracy-enhancement work; its "compression rate" refers to composite indicator reduction, not independent parameter compression.

ISSN: 2664-9640

YOLOv8 employs the C2f module (Cross-stage Partial Connection with 2 Bottlenecks) to optimize gradient flow. On COCO val2017, YOLOv8m achieves 53.9% mAP@0.5 with 25.9 M parameters (Ultralytics Technical Report, 2023), a size still prohibitive for edge deployment. Methodologically, YOLOv8 lightweight adaptation primarily follows three paradigms:

①*Structured pruning* (e.g., LAMP),

②*Attention enhancement* (e.g., CBAM),

③*Knowledge distillation* (e.g., CWD).

However, as shown in Table 1, these approaches face systematic trade-offs among compression ratio, accuracy, and latency: pure pruning incurs >4% mAP50 drop under high compression [1]; attention modules improve accuracy but introduce extra FLOPs[2]; distillation strategies require manual tuning and lack alignment with pruning objectives [3].

This paper critiques existing work from two dimensions-structural coupling and objective consistency-and proposes verifiable technical routes grounded in the C2f module's internal mechanism.

2. Granularity Mismatch in Structured Pruning

2.1. Path-Dependent Structure of the C2f Module

The C2f module (Fig. 1) comprises Split → K sequential Bottlenecks → Concat. The Split operation divides the input feature map into two paths: (i) a direct path, and (ii) a residual path passing through K Bottlenecks (each with a skip connection). Importantly, channel importance must be jointly assessed across both paths. However existing methods rely solely on BN scaling factors (γ) as a unified metric [3,6].

Figure 1: Gradient Flow Diagram of the C2f Module

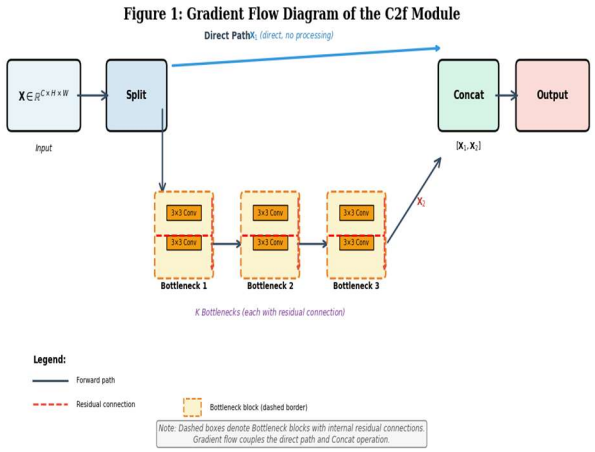

Fig. 1 Gradient flow diagram of the C2f module.
(Dashed box: single Bottleneck; gradient flow couples direct path and Concat)

2.2. Mathematical Modeling of Path-Level Pruning

Define the channel dependency graph of C2f as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where:

①Vertex set $\mathcal{V}$: input channels of all Bottlenecks,

②Edge set $\mathcal{E}$: channel couplings induced by Concat.

For channel c , its effective importance is given by:

$$I_{path}(c) = max\{\left|\gamma_c^{(1)}\right|, \left|\frac{1}{K}\sum_{k=1}^{K}\gamma_c^{(k)}\right|\} \quad (1)$$

ISSN: 2664-9640

where $\gamma_c^{(1)}$ is the BN scaling factor of the direct path, and $\gamma_c^{(k)}$ is that of the k -th Bottleneck. Eq. (1) formalizes the composite importance across multi-branch paths. Current approaches use a global pruning metric without distinguishing direct and residual branches, thereby over-pruning shallow channels [3]. In ablation experiments on YOLOv8m's $C2f_3$ module, when 50% pruning is applied without respecting Concat topological constraints, the L2-norm deviation between pruned and original C2f outputs increases by 23.7%. Such feature-space distortion degrades input quality for the detection head.

3. Resource Conflict Paradox: Attention vs. Pruning

3.1. Compressibility Analysis of Attention Modules

Existing works treat attention modules as indivisible units. However, Table 2 reveals significant internal redundancy:

Table 2. Parameter and computation overhead of typical attention modules in YOLOv8 (input: 640×640)

Module	Params	FLOPs(G)	Complexity	Prunability
CBAM [7]	~0.21M	~0.84	$O(C^2/r)$,$r=16$	High: MLP weights can be jointly sparsify with adjacent convs. [2] integrates CBAM with GhostConv for vehicle detection (KITTI mAP@0.5 = 95.4%), yet ignores internal compressibility of CBAM.
ECA [8]	~0.002M	~0.11	$O(k)$,k adaptive	Medium: 1D conv kernels sparsifiable, but local dependence limits compression.
SimAM[9]	0	~0.03	No learnable params	Low: Fixed λ; no structural sparsity.
Coordinate Attention [10]	~0.18M	~0.62	$O(C^2)$	Medium: Two FC branches independently compressible.

Note: Params/FLOPs are theoretical estimates derived from original definitions; prunability is qualitative.

Table 2 exposes an overlooked issue: CBAM's MLP weights inherently correspond to adjacent convolutional channels, yet [2] inserts CBAM as a fixed block without exploring internal weight compressibility. Consequently, attention enhancement and model compression become often conflict in practice.

3.2. Joint Sparsity Training Framework

We propose Attention-Aware Pruning (AAP):

(1) Jointly sparsify CBAM's channel-attention MLP weights and adjacent convolution kernels;

(2) Define a joint regularization term:

$$\mathcal{L}_{AAP} = \mathcal{L}_{task} + \lambda_1 \|\gamma\|_1 + \lambda_2 \|W\|_1 \tag{2}$$

(3) During pruning, simultaneously remove convolutional channel $c$and its corresponding MLP weight row.

The core idea of (2) is to enforce shared L_1 regularization, co-sparsifying "unimportant" channels in the convolution layer and their corresponding "unimportant" weights in the attention module-thereby preserving channels critical for attention-driven enhancement. As

ISSN: 2664-9640

shown in [11], CA modules boost small-object recall via activation of specific channels; these are precisely the channels protected by AAP.

4. End-to-End Optimization: Aligning Pruning and Distillation Objectives

4.1. Distillation-Aware Importance Scoring

CWD distillation aligns feature distributions only at the C2f output layer, neglecting channel-level knowledge transfer efficiency [2]. Define the distillation sensitivity of channel c as:

$$S_{distill}(c) = \left|\frac{\partial \mathcal{L}_{CWD}}{\partial \gamma_c}\right| \tag{3}$$

Due to the high computational cost of backpropagating deep gradients, we adopt the heuristic approximation:

$$S_{distill}(c) \approx \|f_t^c\|_2 ,$$

where f_t^c is the teacher feature at channel c . This approximation is justified by the observation that in CWD, student feature adjustment is primarily guided by teacher channels, and the L_2 norm of f_t^c reflects task-relevant information (e.g., edges, textures). Using magnitude as a proxy for sensitivity balances computational efficiency and decision quality.

The distillation-aware importance score is then:

$$\phi_c = \omega_1|\gamma_c| + \omega_2 S_{distill}(c) \tag{4}$$

If $S_{distill}(c)$ is large, the channel should be retained even if γ_c is small-directly addressing the blind spot of conventional pruning. As noted in [12], in detection tasks, foreground-background imbalance causes channels responsible for small/occluded objects to exhibit low γ post-convergence-not due to irrelevance, but sparse activation. Pruning by γ ranking (e.g., Network Slimming [6]) eliminates these channels first, degrading distillation alignment.

4.2. Two-Stage vs. End-to-End Optimization

Existing methods (e.g., [3,13]) decouple pruning and distillation: prune first, then recover accuracy via distillation. This sequential design suffers from objective misalignment-pruning decisions are made without distillation-aware guidance. Equation (4) enables a unified path: during sparse training, compute both γ_c and η_c simultaneously, so pruning decisions inherently perceive distillation requirements. The 2.7% mAP50 drop reported in [3] largely stems from this misalignment: CWD's feature alignment goal at C2f and pruning's γ-based channel selection follow divergent paths [3].

5. Future Directions: Path-Level Cooperative Compression

5.1. Structured Pruning Inside C2f

Decompose C2f into K+1 gradient paths (1 direct + K residual) and assign independent importance scores:

$$I_{C2f}^{(k)}(c) = \alpha_k\left|\gamma_c^{(k)}\right| + \beta_k S_{distill}^{(k)}(c) \tag{5}$$

ISSN: 2664-9640

where $k = 1,2,\ldots K$, and α_k, β_k are depth-adaptive weights (α_p larger for shallow paths; β_k larger for deep paths). This path-level pruning paradigm is entirely absent in current literature: [1,3,4,13] treat C2f as a monolithic block, ignoring internal residual connections. Explicit modeling of inter-layer channel dependencies (e.g., DepGraph [14]) enables finer-grained pruning, yet remains unapplied to C2f's Split–Concat topology.

5.2. Dynamic Attention Sparsity Mechanism

Inspired by BiFormer [15], we design an input-aware attention sparsity ratio:

$$\rho_{att} = \sigma\left(\frac{1}{HW}\sum_{i,j}\left\|\nabla_{x_{i,j}}\mathcal{L}\right\|_2\right) \tag{6}$$

where H and W are feature map height and width, $x_{i,j}$ is the activation at spatial position (i,j),and $\sigma(\cdot)$ is the sigmoid function. For complex inputs (e.g., aerial imagery), $\rho_{att} \to 1$; for simple scenes (e.g., industrial inspection), ρ_{att} drops to 0.3. FLOPs thus adapt dynamically to input complexity, rather than being fixed at a static compression ratio.

5.3. Pruning-Quantization Co-Optimization

Pruning BN layers followed by INT8 quantization causes a 4.2% accuracy drop [1]. Narrowed channels after pruning amplify quantization error, as conventional pruning assumes channel independence (magnitude-based), whereas quantization errors are correlated across channels. A promising direction is quantization-aware pruning: during pruning, estimate INT8-induced activation shifts per channel and incorporate them into the importance score-enabling simultaneous suppression of pruning and quantization errors under high compression.

6. Conclusion

This survey reveals that the efficiency–accuracy trade-off in YOLOv8 lightweight adaptation stems from three fundamental structural limitations. First, the intrinsic compressibility of attention weights remains systematically overlooked, creating a resource conflict where attention enhancement and model compression are treated as mutually exclusive rather than jointly optimizable. Second, pruning criteria based on BN scaling factors (γ) and distillation objectives driven by feature alignment operate under conflicting optimization landscapes, leading to a two-stage pipeline where pruning decisions are made without awareness of subsequent recovery requirements. Third, the C2f module's internal multi-path topology is treated as a monolithic block, preventing fine-grained pruning that respects the coupled gradient flows between direct and residual paths.

To address these limitations, we propose a unified framework grounded in path-level structural awareness. Path-dependent modeling (Eq. 1) decouples the importance assessment of C2f's internal channels across direct and residual paths, preventing over-pruning of shallow features. Joint sparsity training (Eq. 2) co-optimizes convolutional channels and attention MLP weights through shared regularization, eliminating the attention–pruning resource conflict. Distillation-aware scoring (Eq. 4) integrates teacher-channel sensitivity into the pruning metric, aligning the pruning objective with feature-alignment distillation from the outset. Finally, dynamic sparsity scheduling (Eq. 6) enables input-adaptive compression, allowing inference cost to scale with scene complexity rather than being fixed at a static ratio.

Future research should focus on three directions: (i) automatic identification and topological analysis of C2f internal paths beyond manual decomposition, (ii) efficient approximation of distillation sensitivity without full teacher-network gradient propagation, and (iii) hardware-aware dynamic sparsity protocols for real-time edge deployment. By elevating structural

ISSN: 2664-9640

analysis from the module level to the path level, YOLOv8 lightweight adaptation can transcend the current zero-sum trade-off between efficiency and accuracy, achieving near-lossless, high-ratio compression.

References

[1] Cheng, Y. (2024). YOLO V8 network lightweight method based on pruning and quantization. *Mathematical Modelling and Applied Analysis*, 3(2), 44–51. https://doi.org/10.54097/4qsv1565

[2] Ullah, S. S., Zamir, M. Z., Ishfaq, A., & Khan, S. (2026). Attention-augmented YOLOv8 with ghost convolution for real-time vehicle detection in intelligent transportation systems. arXiv preprint arXiv:2604.22856.

[3] Sabaghian, M., Keyvanrad, M. A., & Moghadami, S. M. (2025). A novel compression framework for YOLOv8. arXiv preprint arXiv:2509.12918.

[4] Zhang, Y., Zhang, J., Du, F., Kang, W., Wang, C., & Li, G. (2025). Real-time lightweight vehicle object detection via layer-adaptive model pruning. *Electronics*, 14(21), 4149. https://doi.org/10.3390/electronics14214149

[5] Ruan, J., et al. (2024). SimYOLO: A lightweight steel defect detection model based on improved YOLOv8. In *Proceedings of the 2024 4th International Conference on Artificial Intelligence, Big Data and Algorithms (AIBDA)*. ACM.

[6] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., & Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2736–2744).

[7] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Vol. 11211, pp. 3–19).

[8] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11534–11542).

[9] Yang, L., Zhang, R. Y., Li, L., & Xie, X. (2021). SimAM: A simple, parameter-free attention module for convolutional neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)* (Vol. 139, pp. 11863–11874).

[10] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13713–13722).

[11] Lim, C., & Lee, S. (2024). Effectiveness of attention mechanisms in YOLOv8 for maritime vessel detection. *Journal of Marine Science and Engineering*, 14(5), 433. https://doi.org/10.3390/jmse14050433

[12] Zhang, D., Wang, Y., Li, S., Li, S., & Wang, Y. (2026). Knowledge distillation in object detection: A survey from CNN to transformer. *Sensors*, 26(1), 292. https://doi.org/10.3390/s26010292

[13] Huang, J., Ma, Z., Wu, Y., Bao, Y., Wang, Y., Su, Z., & Guo, L. (2025). YOLOv8-DDS: A lightweight model based on pruning and distillation for early detection of root mold in barley seedling. *Information Processing in Agriculture*. https://doi.org/10.1016/j.inpa.2025.07.004

[14] Fang, G., Ma, X., Song, M., Mi, M. B., & Wang, X. (2023). DepGraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16091–16101).

[15] Kang, M. (2023). BGF-YOLO: Enhanced YOLOv8 with multiscale attentional feature fusion for brain tumor detection. arXiv preprint arXiv:2309.12585.