Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 1, 2026

Towards Efficient Retail Text Localization: A Lightweight Network with Saliency-Guided Sparse Attention

Zili Li *

Henan Polytechnic University, Henan 454003, China

Abstract: Retail product packaging text detection is crucial for intelligent label auditing and smart retail applications. While deep learning models, particularly the YOLO series, have demonstrated remarkable real-time inference capabilities, detecting text in complex retail scenarios presents formidable challenges due to extreme scale variations, dense distributions, and severe background interference. During the repeated downsampling processes in standard YOLO architectures, the fine-grained features of microscopic text are prone to degradation, and introducing conventional global attention mechanisms heavily inflates computational costs. To address these issues, this paper proposes an enhanced lightweight, real-time object detector optimized for retail packaging text. First, to prevent the loss of minute text details, we propose a Context Guide Downsample block that jointly aggregates local, surrounding, and global contextual descriptors during spatial resolution reduction. Second, to break the quadratic computational bottleneck of traditional self-attention, a Dynamic Saliency-guided Sparse Attention module is introduced into the encoder. DSSA adaptively filters out background redundancies and establishes efficient long-range dependencies across text regions. Finally, a Multi-scale Feature Mapping module replaces conventional feature concatenation, employing a scale-aware non-linear modulation strategy to align and fuse heterogeneous hierarchical semantics in the neck network. Extensive experiments on the Food-Product-Image dataset demonstrate that the proposed model achieves a superior trade-off between detection accuracy and inference speed, outperforming existing state-of-the-art lightweight detectors in complex text localization tasks.

Keywords: Retail Packaging Text Detection, Lightweight Object Detection, YOLO Architecture, Context-Guided Downsampling, Sparse Attention, Multi-Scale Feature Fusion.

1. Introduction

Retail product packaging serves as a fundamental medium for conveying critical consumer information, encompassing essential details such as product names, ingredient lists, nutrition facts, and expiration dates. The automated detection and localization of these text elements are crucial steps for intelligent food label auditing, supply chain traceability, and automated checkout systems in smart retail. Traditionally, extracting this information relies heavily on manual visual inspection or heuristic-based image processing pipelines, which are labor-intensive, time-consuming, and highly susceptible to complex environmental variations, making it difficult to meet the demands of modern high-speed retail scenarios.

In recent years, deep learning-based object detection algorithms, particularly the YOLO (You Only Look Once) [1] series, have dominated the field of industrial and commercial visual inspection due to their remarkable balance between precision and real-time inference speed. While baseline models like YOLOv8 and YOLOv11 [2] achieve satisfactory results on general public datasets, retail product text detection presents unique and formidable challenges. First, packaging text exhibits extreme scale variations, ranging from massive brand logos to microscopic expiration dates. During the repeated downsampling operations of standard Convolutional Neural Networks (CNNs) in YOLO architectures, the fine-grained texture information of these minute text strokes is prone to severe degradation or complete loss. Second, text instances on packaging often span extensively across the image, requiring models to capture long-range dependencies and global contexts. However, introducing standard self-attention mechanisms to YOLO architectures significantly increases the computational overhead, causing a bottleneck for real-time deployment. Finally, industrial retail environments introduce complex background interference, and the direct fusion of non-adjacent hierarchical features in traditional YOLO necks often leads to semantic distortion and spatial misalignment, restricting the overall detection accuracy across diverse scales.

To fundamentally address the aforementioned challenges, this paper proposes an enhanced lightweight real-time text detection model based on the YOLO architecture, tailored for complex retail product packaging. Specifically, to mitigate the feature loss of microscopic text during spatial resolution reduction, we introduce a Context Guide Downsample block in the early stages of the network. Inspired by the human visual system, the CG block [3] jointly extracts and aggregates local features, surrounding contexts, and global descriptors, effectively preserving discriminative details without inflating parameters. Furthermore, to capture the long-range geometric dependencies of text lines efficiently, we embed a Dynamic Saliency-guided Sparse Attention module. DSSA utilizes an adaptive sparsification strategy to explicitly filter out background redundancies, enabling robust global context modeling while maintaining a linear computational complexity. Finally, to resolve semantic gaps and spatial conflicts during cross-layer feature transmission, we used a Multi-scale Feature Mapping module [4] in the neck network. By applying a scale-aware non-linear modulation mechanism, MFM dynamically aligns and fuses hierarchical features, thereby boosting the model's robustness

to extreme scale variations.

The main contributions of this paper are summarized as follows:

We propose an enhanced lightweight YOLO-based detector specifically optimized for retail product text detection, achieving a superior trade-off between detection accuracy and real-time inference speed on edge devices.

We introduce the Context Guide (CG) Downsample block to prevent the degradation of minute text features. By simultaneously leveraging local, surrounding, and global contextual information, it significantly enriches the feature representation during the downsampling process.

We design the Dynamic Saliency-guided Sparse Attention (DSSA) module, which efficiently establishes long-range spatial dependencies. Through a saliency-driven sparse sampling strategy, it filters out background noise and breaks the quadratic computational bottleneck of standard self-attention.

We apply a novel Multi-scale Feature Mapping (MFM) module that replaces conventional feature concatenation. It adaptively modulates and aligns heterogeneous features across different hierarchical levels, resolving spatial misalignments and semantic conflicts in the neck network.

Extensive comparative and ablation experiments on the retail packaging dataset (Food-Product-Image) validate that the proposed model significantly outperforms existing state-of-the-art lightweight methods in complex text detection tasks.

2. Proposed Method

2.1. Overview Method

To address the formidable challenges of multi-scale text distribution and severe background interference in retail product packaging, this paper proposes an end-to-end lightweight detection network based on the YOLO architecture. As illustrated in Figure X, the overall framework primarily consists of a feature extraction backbone and a meticulously designed synergistic Neck network. In the initial stage, the backbone extracts a hierarchical feature pyramid across three distinct scales, denoted as P1, P2, and P3. To break the computational bottleneck of self-attention while maintaining a global receptive field, we introduce a Dynamic Saliency-guided Sparse Attention module specifically at the deepest semantic level (P3), replacing the original computationally expensive C2PSA module in the YOLO11 baseline. By leveraging an adaptive sparse strategy to explicitly filter out background noise, DSSA not only significantly enhances the lightweight characteristics of the model but also effectively captures the long-range dependencies of complex packaging text. Concurrently, the shallow and medium-level features (P1 and P2) undergo standard convolutional processing for channel alignment, generating the refined representations P1', P2', and P3' as a solid foundation for hierarchical fusion, as depicted in Figure 1.

During the feature transmission and aggregation stage, the model achieves a deep transformation of cross-scale semantics through a top-down and bottom-up routing process. To eliminate the semantic gaps and spatial misalignments caused by conventional direct concatenation, the Multi-scale Feature Mapping (MFM) module is deployed as the primary splicing unit to smoothly align and merge the high-resolution shallow inputs with deep semantic features. Immediately following each MFM fusion node, a Context Guide Downsample (CG) block is embedded to perform comprehensive feature aggregation. By jointly excavating local details, surrounding contexts, and global representations, the CG block ensures that the fine-grained geometric properties of dense text lines are preserved and emphasized during multi-stage propagation. Finally, the fully aggregated multi-scale feature streams are fed into the decoupled detection head, significantly enhancing the localization robustness and classification accuracy for microscopic text instances on irregular packaging surfaces.

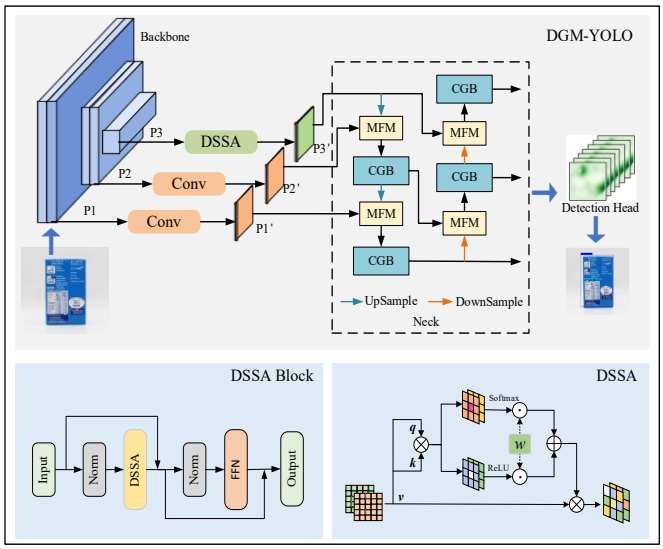

Figure 1. DGM-YOLO Network Structure

2.2. Dynamic Saliency-guided Sparse Attention

While Attention-based detectors excel in global modeling, the standard self-attention mechanism suffers from quadratic time and space complexity, which becomes a significant bottleneck when processing high-resolution feature sequences. To address this, we propose the Dynamic Saliency-guided Sparse Attention (DSSA) in the connect stage. DSSA operates on an adaptive sparsification strategy, which explicitly filters out background redundancies while maintaining the global receptive field.

Given an input feature token sequence $X \in \mathbb{R}^{L\times C}$(where L is the sequence length and C is the channel dimension), DSSA first evaluates the spatial importance of each token through a saliency prediction branch. This branch consists of a lightweight linear projection $W_s \in \mathbb{R}^{c\times 1}$ and a Sigmoid activation, generating a saliency score vector S:

$$S = \sigma(XW_s) \in \mathbb{R}^{L\times 1}$$

Each scalar value s_i in S reflects the semantic contribution of the corresponding spatial pixel. To maximize the retention of critical information within a sparse pathway, the sequence is sorted in descending order based on S, and the indices of the top-k highest-response candidates are dynamically extracted to form an index set $\mathcal{J}_{topk}$.

During the similarity measurement phase, the Query (Q) and Key (K) vectors undergo a dot-product operation with a relative position bias B, generating the raw correlation matrix M:

$$M = \left(\frac{QK^T}{\sqrt{d}} + B\right)$$

Where d represents the feature dimension of the attention head, acting as a scaling factor to prevent the mapping

function from entering the saturation region and to ensure gradient stability.

Instead of executing a dense fully-connected mapping among Q, K, and Value (V), DSSA performs heterogeneous attention aggregation. The interaction is strictly established between the complete Query vector Q and the sparse key-value pairs extracted by the saliency index set. The computational process is formulated as:

$$\mathrm{DSSA}(Q,K,V) = \mathrm{Softmax}\left(\frac{Q(\mathrm{Gather}(K,\mathcal{I}_{topk}))^T}{\sqrt{d_k}}\right)\mathrm{Gather}(V,\mathcal{I}_{topk})$$

By leveraging this non-uniform sampling strategy, the computational complexity is significantly optimized to O(L·K). More importantly, it forces the attention weights to focus strictly on the geometric centers, edges, and semantic regions of the targets, effectively masking low-response background areas. During the convergence process, DSSA autonomously transitions from a coarse-grained global scan to fine-grained local saliency focusing. This mechanism provides a higher signal-to-noise ratio for downstream bipartite matching.

2.3. Context Guide Downsample Block

In standard convolutional networks, repeated downsampling operations often lead to severe degradation of minute object features. Inspired by the human visual system, which relies heavily on contextual information for scene understanding, we introduce the Context Guide Downsample Block(CG) to aggregate comprehensive spatial representations during the downsampling process, as depicted in Figure 2.

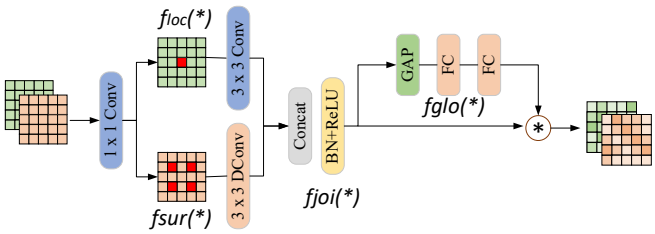

Figure 2. CGBlock Structure

The fundamental intuition behind the CG block is to fully exploit three levels of spatial information: the local feature, the surrounding context, and the global context. As an integrated unit, the CG block consists of four cascaded extractors: a local feature extractor floc(·), a surrounding context extractor fsur(·), a joint feature extractor fjoi(·), and a global context extractor fglo(·).

Specifically, in the first step, floc(·) is instantiated as a standard 3×3 convolutional layer to capture the localized texture details of the target region. Simultaneously, fsur(·)is implemented using a 3×3 dilated convolution, which provides an enlarged receptive field to learn the surrounding contextual dependencies without introducing extra parameters. The outputs of these two extractors are then aggregated by fjoi(·), which is designed as a concatenation operation followed by Batch Normalization and Parametric ReLU activations, yielding the joint feature representation.

In the second step, to further refine the joint feature, fglo(·) is employed to extract the global context of the entire scene. This is achieved by utilizing a global average pooling layer to squeeze the spatial dimensions, followed by a multi-layer perceptron (MLP) to generate a channel-wise descriptor. Finally, a scaling operation is applied to adaptively re-weight the joint feature using the extracted global descriptor, emphasizing discriminative text features while suppressing irrelevant background noise. To facilitate gradient back-propagation and enhance feature complexity, Global Residual Learning is adopted, bridging the initial input directly to the output of the global feature extractor.

2.4. Mutil-scale Feature Mapping Block

In the feature post-processing stage, semantic distortion and spatial misalignment frequently occur when multi-scale feature sequences are transmitted across hierarchical levels. To resolve this, the Multi-scale Feature Mapping (MFM) module is designed as the critical bridge connecting the encoder and the decoder, as depicted in Figure 3. The core function of MFM is to achieve deep alignment between high-level semantics and low-level high-resolution details through a non-linear modulation mechanism.

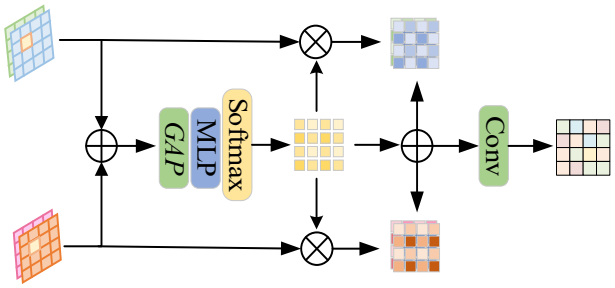

Figure 3. MFM Structure

Unlike conventional methods that rely on simple addition or concatenation, MFM adopts a scale-aware weighted modulation strategy. The input consists of a set of feature maps {F1, F2, … , Fn} extracted from different stages of the encoder, varying in spatial resolutions and channel dimensions. Initially, a set of lightweight projection operators aligns these heterogeneous features into a unified latent space. Subsequently, a global context modeling unit generates a set of dynamic modulation factors $\Gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ where each γ_i represents the adaptive contribution weight of a specific scale to the final detection task.

The core fusion logic of the MFM module is expressed as the following mapping process:

$$F_{fused} = \sum_{i=1}^{n} \gamma_i \odot \Phi_i(\mathrm{Resize}(F_i))$$

Where Resize(·) ensures spatial alignment of features from different scales, and $\Phi_i(\cdot)$ denotes the non-linear transformation function designed to capture cross-dimensional interactive features. Through this formulation, MFM adaptively adjusts the fusion ratio between shallow localization cues and deep classification semantics based on the size distribution of the targets.

3. Experiments and Results

3.1. Dataset and Implementation Details

To comprehensively evaluate the applicability and generalization of the proposed method in structured text detection, we conducted experiments in Food-Product-Image dataset [5].

This dataset characterizes critical information extraction from product packaging, developed by Brosch et al. at the

Software Systems Research Group, Trier University of Applied Sciences, Germany. Image acquisition and annotation strictly adhere to the GS1 Global Product Classification standard. All images were captured under controlled illumination using a Canon EOS 2000D camera at 4000×4000 resolution. A 30-category label system is defined to annotate key retail product regions, supporting holistic annotation and end-to-end attribute-value extraction, covering product name, brand, barcode, nutrition, net content, alcohol content, organic, vegan, and recyclable symbols. Comprising 250 products and 1,034 images (4.14 per product), it features an average label density of 8.14 per image and 33.65 per product. Universal labels (name, brand) cover ~80% of images; domain-specific labels are less frequent. Images are dominated by front, back, and non-front perspectives. Packaging is primarily carton, box, and bag, including 97 plastic and 69 paper items.

All experiments were implemented in the PyTorch 2.4.0 deep learning framework with CUDA 12.2 for GPU acceleration. The hardware platform consists of a dedicated server equipped with a 14-core Intel Xeon(R) Gold 6330 CPU, 62 GB of system RAM, and an NVIDIA GeForce RTX 3090 GPU with 24 GB of video memory. Model training was performed from scratch without any pre-trained weights, and training was terminated upon convergence of the loss function. According to the convergence behavior observed during training, the model was trained for 350 epochs with input image sizes of 1280×1280 pixels. The training process employed stochastic gradient descent (SGD) as the optimizer with a learning rate of 0.001, weight decay of 0.0005, momentum of 0.937, and batch sizes of 16 in experimental settings.

3.2. Evaluation Metrics

Common evaluation metrics for scene text detection include Precision (P), Recall (R), and mean Average Precision (mAP). Model lightweightness is assessed using metrics such as the number of parameters and GFLOPS.

Precision measures the proportion of correctly predicted positive samples among all predicted positive samples, defined as:

$$Precision = \frac{TP}{TP + FP}$$

Where TP denotes true positives (correctly detected target information) and FP denotes false positives (incorrectly predicted positive samples).

Recall represents the ratio of correctly detected positives to all ground-truth positives, computed as:

$$Recall = \frac{TP}{TP + FN}$$

Where FN denotes false negatives (positives missed by the model).

Average Precision (AP) is the area under the precision–recall curve, expressed as:

$$AP = \int_0^1 P\,recision(Recall)\,d(Recall)$$

Mean AP (mAP) is the average AP across all classes:

$$mAP = \frac{1}{N}\sum_{i=1}^{N} A\,P_i$$

Where APi is the AP of the i-th class and N is the total number of classes.

mAP@0.5 refers to mAP computed at an IoU threshold of 0.5, while mAP@0.5:0.95 denotes the average mAP over IoU thresholds from 0.5 to 0.95 at 0.05 intervals.

3.3. Comparisons of Prior Arts

To comprehensively evaluate the effectiveness and superiority of the proposed DGM-YOLO in complex retail packaging text detection, we conducted extensive comparative experiments against several representative state-of-the-art (SOTA) object detectors. The selected baselines encompass classic two-stage detectors (Faster R-CNN[6]), efficient multi-scale networks (EfficientDet), and the mainstream real-time YOLO series (ranging from YOLOv5n to the latest YOLOv12n). The quantitative evaluation results, measured by mAP@0.5, mAP@0.5:0.95, Recall, and inference speed (FPS), are detailed in Table 1.

Table 1. Comparison with SOTA models

Model	mAP@0.5	mAP@0.50:0.95	Recall	FPS
Faster-RCNN	41.9	27.2	46.9	23.7
Efficient-Det	35.3	20.6	40.7	30.6
YOLOv5n	42.9	28.7	50.9	63.7
YOLOv8n	48.5	34.5	59.5	72.5
YOLOv10n	45.1	32.5	55.5	68.8
YOLOv11n	47.4	33.6	58.6	75.3
YOLOv12n	47.9	34.0	61.7	80.5
DGM-YOLO	50.2	35.3	60.5	85.5

As observed from the experimental results, the proposed DGM-YOLO achieves the most outstanding comprehensive performance, establishing a new SOTA baseline for lightweight text detection. Specifically, DGM-YOLO secures an impressive 50.2% in mAP@0.5 and 35.3% in the more stringent mAP@0.5:0.95 metric, simultaneously maintaining a highly competitive Recall of 60.5%.

When compared to the classic two-stage architecture Faster R-CNN, DGM-YOLO demonstrates overwhelming advantages across all primary metrics. Faster R-CNN suffers from a low mAP@0.5 of 41.9% and a sluggish inference speed of 23.7 FPS, rendering it unsuitable for real-time deployment on edge devices. Similarly, EfficientDet struggles with the dense and multi-scale text distribution inherent in our dataset, yielding a sub-optimal mAP@0.5 of only 35.3% and a Recall of 40.7%.

Furthermore, in direct comparison with the highly optimized YOLO family, the superiority of DGM-YOLO remains evident in both detection accuracy and localization robustness. Compared to the widely adopted YOLOv8n baseline, our model achieves a 1.7% absolute improvement in mAP@0.5 and a 0.8% boost in mAP@0.5:0.95. More importantly, when evaluated against the latest and most advanced YOLOv12n baseline, DGM-YOLO still manages to secure a significant performance leap, outperforming it by 2.3% in mAP@0.5 and 1.3% in mAP@0.5:0.95. Although YOLOv12n exhibits a marginally higher Recall (61.7% vs. 60.5%), DGM-YOLO's substantial lead in mean Average Precision confirms its superior ability to accurately classify and tightly bound complex text instances.

Beyond detection accuracy, inference efficiency is a critical

criterion for practical industrial applications. Remarkably, DGM-YOLO achieves the highest inference speed among all evaluated models, reaching an exceptional 85.5 FPS. This is notably faster than the baseline YOLOv11n (75.3 FPS) and YOLOv12n (80.5 FPS). The unprecedented real-time processing capability, coupled with the highest overall precision, unequivocally proves that our proposed components (CGB, DSSA, and MFM) successfully mitigate background noise and redundant computations, achieving an optimal trade-off between detection accuracy and efficiency.

3.4. Ablation Studies

To systematically validate the independent contributions and synergistic effects of the proposed components—namely, the Dynamic Saliency-guided Sparse Attention (DSSA), the Context Guide Block (CGB), and the Multi-scale Feature Mapping (MFM)—we conducted comprehensive ablation experiments based on the YOLOv11n architecture. The performance of each variant, evaluated in terms of detection accuracy (mAP@0.5) and model complexity (Parameters in Millions), is summarized in Table 2.

Table 2. Ablation study results

Model	DSSA	CGB	MFM	mAP@0.5	Parameters
Baseline				47.4	2.588
Baseline 1	√			49.6	2.406
Baseline 2		√		48.2	2.498
Baseline 3			√	47.7	2.511
DGM-YOLO	√	√	√	50.2	2.389

Effectiveness of DSSA: We first evaluate the impact of replacing the standard attention module in the baseline with our proposed DSSA. As shown in Table 2, Baselinel (incorporating only DSSA) achieves a substantial performance leap, boosting the mAP@0.5 from 47.4% to 49.6% (an absolute improvement of 2.2%). More importantly, this significant accuracy gain is accompanied by a remarkable reduction in model size, decreasing the parameter count from 2.588M to 2.406M. This clearly demonstrates that the saliency-driven sparse sampling strategy in DSSA effectively filters out background redundancies and focuses computational resources on critical text regions, thereby successfully breaking the quadratic complexity bottleneck of global attention mechanisms while enhancing representation capability.

Effectiveness of CGB: Next, we assess the contribution of the Context Guide Block during the downsampling stage. Comparing Baseline2 with the original baseline, the integration of CGB yields a 0.8% increase in mAP@0.5 (reaching 48.2%) while simultaneously reducing the parameters to 2.498M. This validates our hypothesis that explicitly modeling local, surrounding, and global contexts during spatial resolution reduction prevents the loss of fine-grained textures, proving highly effective for detecting minute and irregularly shaped retail packaging text.

Effectiveness of MFM: When the Multi-scale Feature Mapping module is deployed independently in the neck network (Baseline3), the model attains an mAP@0.5 of 47.7% alongside a parameter reduction to 2.511M. Although the individual accuracy gain (0.3%) is relatively modest compared to DSSA and CGB, the consistent reduction in parameters confirms that the scale-aware non-linear modulation strategy is computationally more efficient than traditional dense concatenation, effectively mitigating semantic misalignments across hierarchical levels.

Synergistic Effects in DGM-YOLO: Finally, the integration of all three proposed modules yields the complete DGM-YOLO framework. This unified architecture achieves an unprecedented mAP@0.5 of 50.2%—a striking 2.8% absolute improvement over the baseline. Crucially, this peak accuracy is attained with the lowest model complexity among all variants, utilizing only 2.389M parameters. This exceptional outcome indicates that DSSA, CGB, and MFM are not merely additive components but function synergistically. The CGB preserves vital low-level details early in the network; DSSA efficiently models long-range dependencies in deep semantic layers; and MFM seamlessly aligns these multi-scale representations in the neck. Consequently, DGM-YOLO establishes a superior architecture that maximizes detection precision while minimizing computational overhead for complex retail text inspection.

3.5. Qualitative Analysis of Detection Results

To intuitively demonstrate the superiority of the proposed DGM-YOLO in real-world retail scenarios, we conducted a comprehensive qualitative evaluation. We first compare our model with mainstream baselines to highlight its precision in dense layouts, and then showcase its robustness across diverse packaging modalities.

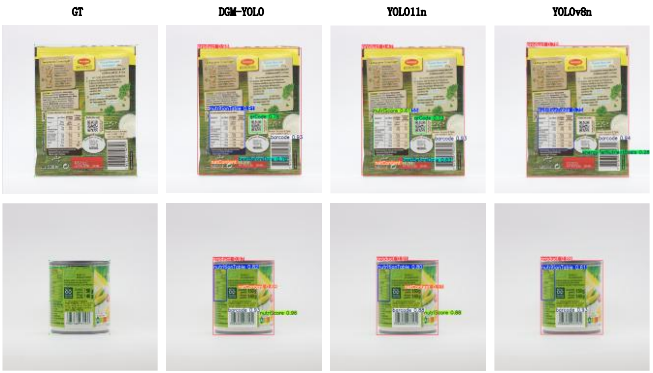

Figure 4. Detection results of the Comparative method

Figure 5. Overview Detection results

Figure 4 illustrates the qualitative comparison between the proposed DGM-YOLO and baseline models (YOLOv8n, YOLOv11n) under challenging conditions involving dense text distributions and curved surface distortions. In the highly clustered pouch packaging scenario (top row), baseline models exhibit significant feature degradation; YOLOv8n completely misses microscopic targets like qrCode and bestBeforeDate while generating false positives, and YOLOv11n suffers from bounding box overlapping and misclassifications. DGM-YOLO successfully localizes all fine-grained elements with high confidence, producing tight, independent bounding boxes highly consistent with the Ground Truth. Similarly, on the cylindrical can (bottom row) where text undergoes severe geometric distortion, both baselines fail to detect the netContent and the marginal nutriScore icon due to feature dilution. DGM-YOLO accurately captures these distorted elements. This comparative analysis demonstrates that the proposed Context Guide (CG) and DSSA modules effectively prevent the loss of microscopic details during feature extraction, empowering the architecture to robustly resolve spatial and semantic conflicts in densely deformed text regions.

To validate the generalization capability of DGM-YOLO, Figure 5 visualizes detection results across a diverse array of complex packaging modalities, materials, and extreme scale variations. The model exhibits exceptional robustness against severe physical and environmental interferences. It flawlessly detects massive global targets, such as the entire product bounding box, simultaneously with extremely minute text, exemplified by the netContent on the yogurt lid and the bestBeforeDate on the snack bag. This precise concurrent localization strongly validates the efficacy of the Multi-scale Feature Mapping (MFM) module in aligning heterogeneous semantics across disparate spatial resolutions. On flexible packaging prone to severe geometric deformations and wrinkles, including plastic-wrapped vegetables and distorted snack bags, DGM-YOLO accurately identifies the barcode and nutritionTable without being compromised by structural irregularities. When confronted with strong specular reflections on metallic soda cans and glass jars, or highly cluttered backgrounds typical of yogurt packaging, the model consistently maintains high-confidence predictions for critical attributes. This visual evidence underscores the remarkable adaptability of DGM-YOLO for real-world unstructured retail environments.

4. Conclusion

In this paper, we address the formidable challenges of extreme scale variations, dense distributions, and complex background interference inherent in retail product packaging text detection. To overcome the feature degradation and computational bottlenecks of conventional object detectors, we propose DGM-YOLO, a highly efficient and robust lightweight network. The proposed architecture is fundamentally strengthened by three novel components. First, the Context Guide (CG) Downsample block is introduced in the early stages to jointly aggregate local, surrounding, and global contextual descriptors, effectively preventing the loss of fine-grained text features during spatial resolution reduction. Second, the Dynamic Saliency-guided Sparse Attention (DSSA) module is embedded to perform efficient token interaction. By adaptively filtering background redundancies, DSSA breaks the quadratic computational complexity of standard self-attention while establishing robust long-range dependencies across text regions. Finally, the Multi-scale Feature Mapping (MFM) module replaces conventional feature concatenation, employing a scale-aware non-linear modulation strategy to align heterogeneous hierarchical semantics and resolve spatial misalignments.

Extensive experiments on the Food-Product-Image datasets demonstrate the superiority of our approach. DGM-YOLO achieves an outstanding mAP@0.5 of 50.2% and a real-time inference speed of 85.5 FPS with only 2.389M parameters. This exceptional trade-off between detection accuracy and computational efficiency establishes DGM-YOLO as a highly practical solution for deployment on resource-constrained edge devices in intelligent retail and automated auditing systems. Future work will focus on integrating Multi-modal large language models (LLMs) to further resolve semantic ambiguities in multi-lingual and irregular packaging layouts.

Acknowledgements

We sincerely thank the researchers who constructed and open-sourced the Food-Product-Image dataset, which provided an essential foundation for the training and comprehensive evaluation of the models in this study. We also extend our gratitude to the developers of the YOLO series and the PyTorch open-source communities for providing invaluable tools and foundational frameworks that facilitated our research.

References

[1] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[2] Rahima K, Muhammad H. YOLOv11: An Overview of the Key Architectural Enhancements [J]. arXiv, 2024.

[3] Wu T, Tang S, Zhang R, et al. CGNet: A Light-weight Context Guided Network for Semantic Segmentation [J]. arXiv, 2019.

[4] Zhang Y, Zhou S, Li H. Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing [C]. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).2024:2846~2855.

[5] Brosch C, Bouwens A, Bast S, et al. Creation and Evaluation of a Food Product Image Dataset for Product Property Extraction [J]. arXiv, 2025.

[6] Chen J R, Kao S H, He H, et al. Run, don't walk: chasing higher FLOPS for faster neural networks [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12021-12031.