Journal of Computers Vol. 37 No. 4, August 2026, pp. 57-73
https://doi.org/10.63367/19911599202608370400 5

LiteOCR: A Lightweight Text Detection Framework Integrating Cross-Modal Semantic Distillation and Dynamic Resolution Decision-Making

Dian-Hui Mao[1], Ming-Dong Xie[1], Zhi-Hua Zhao[2],
and Fang Fang[3*]

[1] School of Computer and Artificial Intelligence, Beijing Technology and Business University,
Beijing 100048, China
maodh@th.btbu.edu.cn, 2671661857@qq.com

[2] School of Law, China University of Political Science and Law,
Beijing 102249, China
zhaozhihua@cupl.edu.cn

[3] Commodity Barcode Department, GS1 China,
Beijing 100011, China
fangf@ancc.org.cn

Received 21 July 2025; Revised 25 January 2026; Accepted 28 January 2026

Abstract. To address the urgent demand for efficient and robust optical character recognition (OCR) technology in the e-commerce sector, this paper introduces LiteOCR—a lightweight OCR framework designed for edge intelligence deployment. This framework integrates three core technologies: cross-modal knowledge distillation, semi-supervised learning mechanisms, and dynamic computation gating strategies. The framework aims to effectively overcome the significant limitations of traditional OCR systems, which suffer from insufficient robustness in complex environments and the difficulty of running resource-intensive large models efficiently on computationally constrained devices. Its core design principle focuses on achieving an optimal balance between adaptive capabilities for complex and variable visual conditions in e-commerce scenarios and real-time processing performance on edge devices, providing an effective solution for efficient text information extraction at the edge. Experiments demonstrate that on a custom-built e-commerce dataset, LiteOCR achieves an F1 score of 83.3%. Despite having only 15% of the model parameters compared to its teacher model (GOT-OCR 2.0), the performance gap narrows to just 1.8 percentage points. This comprehensively validates the effectiveness of cross-modal attention distillation in preserving semantic accuracy, dynamic resolution gating in enhancing computational efficiency, and semi-supervised feature alignment in reducing data dependency, providing a practical lightweight solution for high-precision OCR deployment at the edge.

Keywords: e-commerce domain, OCR, cross-modal attention distillation, semi-supervised learning, dynamic gating

1 Introduction

In the e-commerce sector, optical character recognition (OCR) technology has emerged as a critical enabler for core business processes such as multimodal product information management, structured processing of massive electronic invoices, and end-to-end logistics tracking. Confronted with daily demands for entering tens of thousands of product records, traditional manual data entry—hampered by high labor costs and inherent limitations in ensuring systematic and real-time data management—can no longer meet the rapid development needs of modern e-commerce. The introduction of OCR technology, through automated data extraction and structured processing, has effectively reduced human resource investment while significantly enhancing operational efficiency and data reliability.

However, existing mainstream OCR systems (such as PP-OCRv3 [1] and DB-Net [2]) demonstrate outstanding performance in general image text detection tasks. Yet when processing complex images specific to

* Corresponding Author

e-commerce scenarios, their fixed computational strategies and reliance on a single modality reveal significant limitations. Specifically, these limitations manifest in two key aspects: First, scene complexity undermines model generalization capabilities. Challenges such as reflective transparent packaging, densely packed 8pt micro-font text, multilingual text mixing, and low-resolution images severely test the recognition accuracy of traditional models. Second, industrial deployment challenges persist. While integrating large language models enhances semantic understanding, their high edge-side inference latency, stringent hardware requirements for multi-billion-parameter models, and potential data privacy compliance risks make them ill-suited for meeting the real-time demands of e-commerce scenarios.

In recent years, researchers have attempted to optimize OCR model efficiency through techniques such as knowledge distillation [3] and dynamic networks [4]. MobileNetV3 [5] employs channel pruning and 8-bit quantization to compress the model to 10MB, yet suffers over 9% accuracy loss. CLIP-OCR [6] enhances semantic understanding via visual-language pre-training, but its computational complexity (FLOPs > 150G) prevents deployment on edge devices.

To address challenges in OCR tasks within e-commerce scenarios, this paper introduces LiteOCR—a product label recognition framework integrating lightweight cross-modal distillation with dynamic resolution decision-making. By employing cross-modal semantic distillation, the architecture reduces model parameters to 13.4M. Combined with locally texture-aware dynamic computation optimization, it achieves end-to-end latency below 100ms on NVIDIA Jetson Nano edge devices, achieving an F1-score exceeding 83.3% for key field recognition. The core innovations and contributions of this research are as follows:

(1) Cross-Modal Attention Distillation Architecture: A hierarchical attention transfer method based on the GOT-OCR 2.0 model. This approach achieves joint channel-spatial dynamic feature enhancement through bidirectional knowledge transfer pathways, overcoming the limitations of rigid feature representations and coarse cross-modal alignment in traditional unimodal distillation. It significantly enhances the model's discriminative robustness in challenging scenarios involving noise, occlusion, and data scarcity.

(2) Designing a Semi-Labeled Data-Driven Training Scheme Based on the Mean-Teacher Framework: By combining strong-weak data augmentation with feature consistency loss, this approach maintains detection accuracy and enhances model generalization under reduced labeled data, addressing the issue of scarce annotated data.

(3) Fine-grained Dynamic Resolution Gating Strategy: A text-density-driven mechanism for dynamic convolution kernel selection. A lightweight density prediction head estimates local text density in real-time, driving dynamic activation patterns for kernels across stages. During inference, this strategy adaptively adjusts computational paths based on image content—significantly reducing overall computational load by eliminating redundant computations in sparse text regions while preserving full-precision processing in dense, complex areas. This achieves substantial computational savings without compromising recognition accuracy in critical regions.

The final structure of the paper is as follows: Section 2 discusses recent research trends in OCR technology. Section 3 details the implementation of our method. Section 4 presents experiments comparing model performance through cross-validation and ablation studies. Section 5 concludes the paper.

2 Related Works

Traditional OCR systems face performance bottlenecks when processing complex text images, particularly in e-commerce scenarios characterized by high background noise, low resolution, and multilingual text mixing. Most conventional OCR systems rely on large models that excel in standard tasks but leave significant room for improvement in computational efficiency and inference latency. To address these issues, academia has proposed various methods in recent years to optimize OCR efficiency and accuracy, encompassing techniques like model compression, cross-modal learning, and dynamic networks. Han, Li [7-8] and others effectively reduced model size using pruning techniques, decreasing parameter count and computational load by removing unimportant connections or neurons. Chitty-Venkata and Song et al. [9-10] employed quantization techniques to convert floating-point model parameters into low-bit fixed-point representations, thereby shrinking model size and accelerating inference. For instance, PP-OCRv3 [1] adopted cross-modal joint learning to reduce its model size to 15MB while demonstrating outstanding performance across general scenarios. However, such methods suffer significant performance degradation in complex text scenarios, primarily because unimodal compression

strategies disrupt the visual-semantic feature coupling. Although Yang et al. [11] attempted to mitigate performance loss through knowledge distillation from recognition layers, their approach failed to preserve fine-grained cross-modal associative features due to limitations in shallow-layer output distribution transfer. Additionally, Liu et al. [12] explored how structured pruning techniques could better preserve model architecture to mitigate performance impacts.

Multimodal OCR research focuses on constructing collaborative image-text representation spaces. Lv et al. [13] achieves feature alignment through contrastive learning, while Florencio et al. [14] employs a Transformer architecture to fuse multimodal sequences, achieving high accuracy in natural scene text recognition and demonstrating the Transformer's robust capability in processing sequential data. However, the backbone networks of both models exhibit high computational load and complexity, rendering them unsuitable for industrial scenarios demanding real-time performance and lightweight solutions. While Wan et al.'s [15] dynamic modality selection method activates specific branches based on input complexity, thereby partially optimizing computational resource usage, inference latency on edge devices remains excessively high, failing to meet practical industrial demands.

Beyond these approaches, recent research has explored multimodal pretraining. Parikh et al. [16] applied this to OCR by pretraining on large-scale text-image datasets, enabling models to learn richer cross-modal representations. For instance, some studies jointly train text and image encoders to enhance understanding of textual content within images.

Dynamic network design aims to adaptively adjust model structure or computational paths based on input data characteristics or computational resource constraints, thereby enhancing computational efficiency while maintaining performance. This approach has achieved progress in domains such as image classification (Wang et al.) [17] and object detection (Ghiasi et al.) [18]. Techniques like separable convolutions (Chollet et al.) [19], attention mechanisms (Vaswani et al.) [20], and dynamic routing have provided novel insights for implementing dynamic networks.

However, its application in OCR still faces challenges. ABCNetv2 [21] employs deformable convolutions to adapt to curved text geometries, enhancing the model's recognition capability for irregular text. Yet, its fixed computational path leads to redundant memory consumption. DynamicViT [22] dynamically prunes visual features based on token importance scores, reducing computational load but disrupting textual sequence continuity, thereby increasing recognition error rates. This indicates that applying dynamic network designs to OCR tasks requires more nuanced consideration of textual sequence properties and cross-modal correlations.

A review of existing research reveals three key shortcomings in current approaches: 1) Inefficient cross-modal knowledge transfer: Traditional compression methods disrupt cross-modal feature correlations, leading to performance degradation in complex scenarios. Existing cross-modal OCR methods rely on coarse-grained image-text matching, neglecting hierarchical associations between text regions and semantic concepts, resulting in inefficient knowledge transfer; 2) Coarse-grained dynamic computation: Existing dynamic strategies rely on global complexity assessments, failing to precisely adapt to local text characteristics; 3) Heavy reliance on annotated data: Semi-supervised OCR research (Yang et al. [23]) depends on strong annotation constraints, limiting its applicability in low-resource scenarios. This paper addresses these issues through systematic innovations: hierarchical attention distillation, pixel-level gated decision-making, and multi-view feature alignment.

3 Methodology

This study addresses the dual challenges posed by the multimodal characteristics of product images (visual appearance, textual descriptions, multilingual labels) and complex background noise (promotional watermarks, irregular layouts, multiscale text) in e-commerce scenarios. It simultaneously provides a feasible solution for low-computational-power environments and complex, unlabeled scenarios. The overall model architecture is illustrated in Fig. 1.

The left side of the figure shows the pre-training phase of the model, the upper half is the student model channel, and the lower half is the teacher model channel. The leftmost part is an example of a semi-supervised feature on its training method; the middle part is a cross-modal attention distillation method, only part of the module is given in the example, and the details will be introduced in the following section; and the right half is an example of a dynamic gating module.

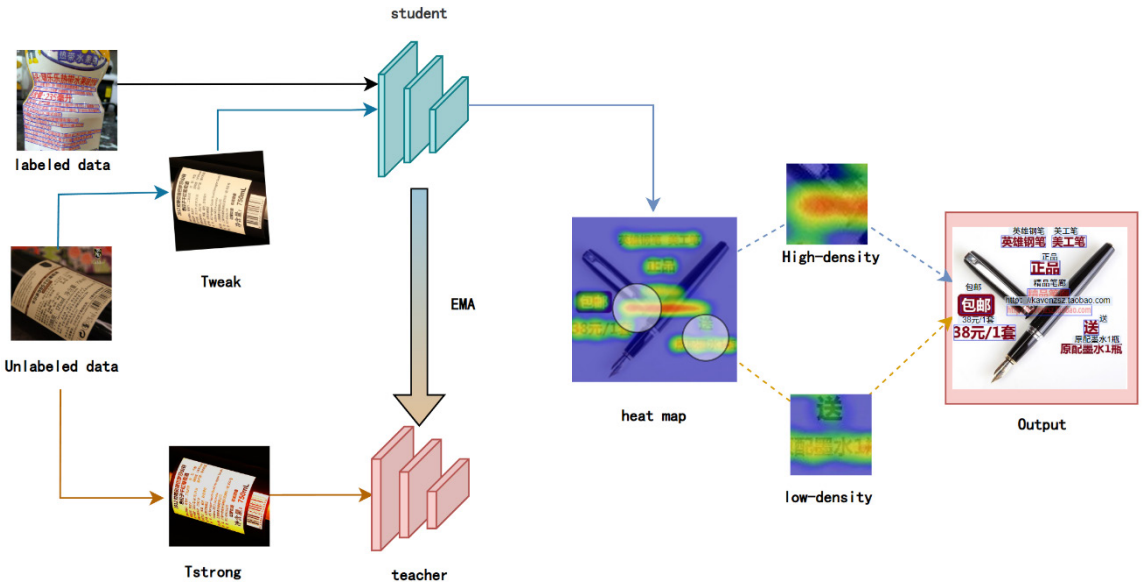


Fig. 1. Logic diagram of the overall architecture of the model

3.1 Cross-modal Attention Distillation Architecture

Existing OCR models' lightweight process is often due to cross-modal semantic correlation breaks leading to key information (e.g., price, specification parameters) recognition accuracy significantly decreased, this phenomenon is particularly prominent in e-commerce high-noise scenarios, and the existing OCR recognition models based on large model correlation (Omniparser [15], GOT-OCR [24], etc.) are large is not easy to be applied to low hardware performance scenarios. While traditional unimodal distillation methods only focus on feature alignment in a single modality of text or image, they lack targeted optimization of local key regions and cannot take advantage of complementary multimodal information. Therefore, we use a cross-modal attention distillation architecture, which uses a hierarchical attention migration mechanism (decoupling the joint visual-textual representation space) that can transform the cross-modal attention patterns of the teacher model into migratable semantic priors and establishes deep inter-modal correlations through dynamic projection matrices, thus realizing a fundamental shift of the knowledge transfer paradigm in the model compression process.

In this study, GOT-OCR2.0 is selected as the teacher model for cross-modal attention distillation, mainly based on its significant advantages in multimodal fusion capability, complex scene robustness, and end-to-end architectural design, which are highly compatible with the OCR requirements in e-commerce scenarios, and MobileNet-V3 can inherit its parsing capability for complex graphic layouts as the student model. At the same time, it meets the lightweight demand of mobile deployment and provides technical support for real-time product information extraction, multilingual label management, and other applications of e-commerce platforms.

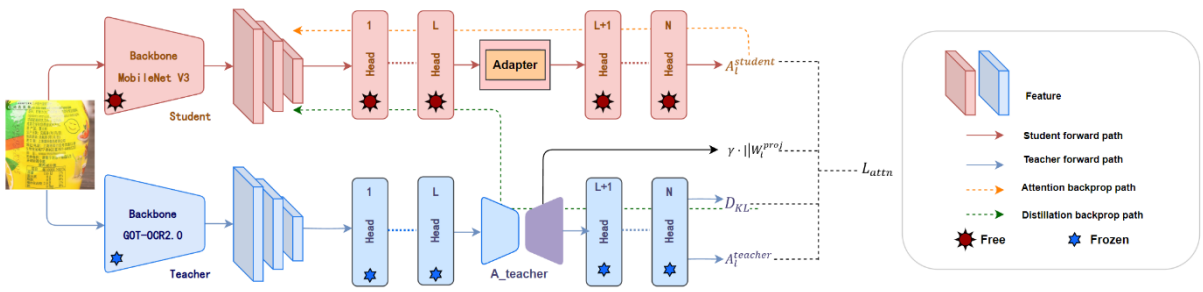


Fig. 2. Cross-modal attention distillation architecture

The overall architecture of cross-modal attention distillation is shown in Fig. 2. It includes a bidirectional knowledge transfer pathway: given a teacher-student model pair, the teacher model generates language prior features through a visual encoder and a text encoder, which are mapped to the visual feature space via a cross-modal projection matrix Wproj; the student model extracts features through a lightweight visual encoder. The visual encoder needs to retain text structural information while filtering out irrelevant background noise, and by decoupling the cross-modal semantic association features of the GOT-OCR2.0 model, a hierarchical knowledge transfer loss function is constructed:

$$L_{attn} = \Sigma_{l=1}^{L} \lambda_l \cdot [D_{KL}(A_l^{teacher} \mid A_l^{student}) + \gamma \cdot \left| W_l^{proj} \right|_F^2] \tag{1}$$

In equation (1), $A_l^{teacher} \in R^{H \times W \times C}$ and $A_l^{student} \in R^{H \times W \times C}$ denote the cross-modal attention matrices of layer l of the teacher-student model, and $Wproj$ is the learnable eigenprojection matrix, respectively. The KL scatter term $DKL(\cdot)$ constrains the similarity of the attention distributions, and the Frobenius paradigm term regularizes the projections by regularizing the matrix's singular value distribution to alleviate the modal divide problem. When the semantic feature mapping is introduced into the channel attention module of the student model, the lower bound of its visual feature discriminability can be expressed as:

$$D(F_s) \geq D(F_t) - \in \cdot \sqrt{\dim(H)} \tag{2}$$

In equation (2), $D(\cdot)$ is the feature discriminative measure, $\in$ is the semantic embedding error, and H is the Hilbert feature space.

3.2 Semi-supervised Dynamic Feature Consistency Constraints

To enhance the generalization of the model during the fine-tuning training phase, especially in dealing with massive product images, user behavior data, etc., the scarcity of labeled data is a prevalent bottleneck, and how to effectively utilize a large number of unlabeled product images or user interaction information is crucial to improve the performance of OCR recognition in downstream tasks such as product classification, similar product recommendation, and visual search. Although the Mean-Teacher framework [25] has proven successful by enforcing consistency of predictions between different perturbed views of the same input, this consistency is usually measured at the output layer. In order to motivate the model to adapt to e-commerce scenarios with high product diversity and fine-grained categories, we design a semi-supervised dynamic feature consistency constraint mechanism, which extends consistency regularization directly to the feature space.

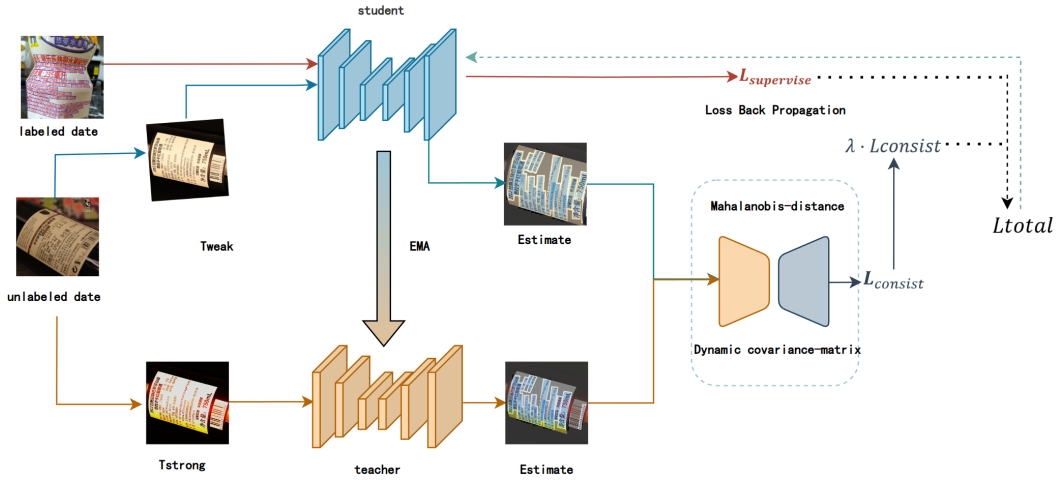


Fig. 3. Hierarchical modular diagram of the training process of the semi-supervised feature alignment architecture

The cross-entropy loss (supervised loss) is computed after the labeled data is input to the student model, the unlabeled data is input to the student and teacher models via weak-strong dual-view enhancement, respectively, and the feature layer outputs are aligned by the dynamic covariance matrix-estimated Mahalanobis distance. The teacher model parameters are updated by EMA (η=0.996), the feature space alignment module is weighted and fused with the supervised loss in the gradient backpropagation stage (λ=3.0), and the feature distribution alignment error converges to the theoretical optimal interval (Δ<0.1) when the amount of unlabeled data reaches the critical threshold.

Our approach builds on Mean-Teacher's student-teacher architecture, preserving the exponential moving average (EMA) updating rule for the teacher model parameters ($\theta' \leftarrow \eta\theta' + (1-\eta)\theta$) to ensure the stability of the objective as shown in Fig. 3. The core innovation lies in the way to handle unlabeled commodity data x_u (e.g., commodity images). We adopt a strong-weak data enhancement strategy to generate a weakly enhanced view *Tweak* (random cropping (retention ≥ 0.8), ±15° rotation), and a strongly enhanced view *Tstrong* (motion blur (kernel size 15×15), HSV color dithering (Δ = 0.2)). The student model fθ handles *Tweak*, while the teacher model fθ' handles *Tstrong*. Also, to measure the consistency between the commodity feature vectors, we utilize the Mahalanobis distance. Compared to metrics such as Euclidean distance, the Mahalanobis distance provides a more principled measure of feature spatial similarity by taking into account the correlation between feature dimensions and their respective variances, which helps to capture the subtle differences between commodity features. The feature consistency loss, *Lconsist*, is defined as the square of the Mahalanobis distance:

$$Lconsist = Ex_u (f\theta\left(Tweak\left(x_u\right)\right) - f\theta'\left(Tstrong\left(x_u\right)\right))^T \ \Sigma^{-1}\left(f\theta - f\theta'\right)L \tag{3}$$

In equation (3), A key aspect of the method is that the feature covariance matrix Σ is dynamically estimated during training, typically using running statistics computed on small batches of data. This allows the consistency metric to adapt to the changing distribution of commodity features during model learning, and the teacher model parameters are updated via EMA:

$$\theta' \leftarrow \eta\theta' + \left(1-\eta\right)\theta, \eta = 0.996 \tag{4}$$

The overall training objective for the student model combines the standard supervised loss Lsupervised on labeled data (x,y) (e.g., cross-entropy loss for a product classification task) and the feature consistency loss after a weighted λ has been applied to unlabeled product data:

$$Ltotal = Lsupervised + \lambda \cdot Lconsist \tag{5}$$

Finally, this loss is back-propagated back to the student model to achieve dynamic parameter tuning.

3.3 Dynamic Resolution Gating Module

To solve the efficiency-accuracy tradeoff problem under the edge device arithmetic constraint, we incorporate a dynamic resolution decision module based on local complexity perception. Given an input image $I \in R^{H\times W\times C}$, we first traverse the image with a sliding window (32×32 with a step size of 16), compute the text density $D(P_i)$ of each region Pi with the multiscale edge response $\| \nabla I \| k$, and use kernel estimation to generate the initial heat distribution map as shown in Fig. 4. Meanwhile, the multi-scale edge response $\| \nabla I \|_k$ is normalized to generate the edge activation map, and the initial heat map and the edge activation map are resampled to the original map resolution by bilinear interpolation, and the feature fusion is carried out by using the channel-attention mechanism to form a comprehensive heat map. The local complexity function is defined as:

$$C\left(Pi\right) = \alpha D\left(Pi\right) + \beta \sum_{k=1}^{3} \gamma_k \| \nabla I \|_k \tag{6}$$

In equation (6), where α is the text density term coefficient, β is the edge response supplement term coefficient, and γ_k is the multiscale weighting coefficient. To ensure the system's resource allocation for critical information in dense text regions while maintaining model accuracy and preventing failure in sparse peripheral areas, we defined these two coefficients as $\alpha = 0.6$ and $\beta = 0.4$ through experimental parameter tuning.

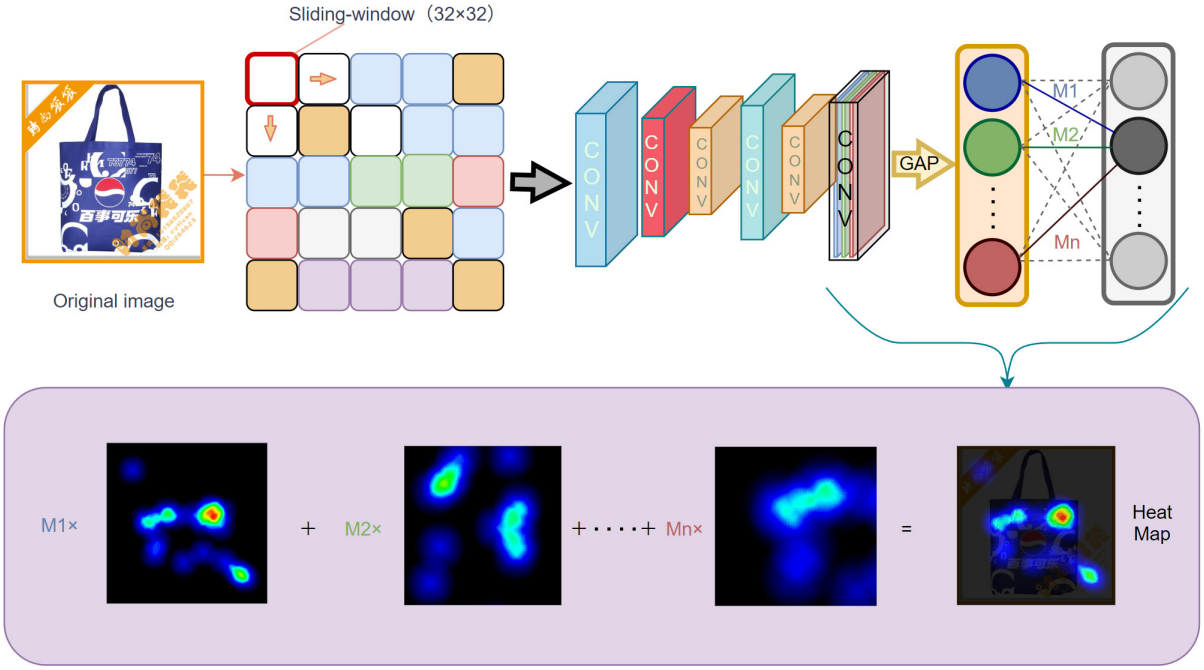


Fig. 4. Process of heat map generation

The spatial features are extracted by a convolutional neural network to generate a resolution decision mask $M \in \{0,1\}^{H \times W}$, which employs a depth-separable convolution to reduce the computational effort and introduces a Sigmoid activation function and a resource constraint term in the output layer to ensure that the percentage of high-resolution regions does not exceed a preset threshold.

The dynamic gating network G generates a resolution decision mask that can be expressed as:

$$M = G(C(Pi)); \theta G), \text{ s.t. } E[M] \leq Rmax \tag{7}$$

In equation (7), $Rmax$ is a preset maximum computational resource constraint. The red high-density regions (Hc>0.8) in the heat map are preferentially assigned to the high-resolution branch, while the blue low-density regions (Hc<0.3) are down-sampled to the low-resolution branch for processing, and the transition regions are realized as a soft selection through probabilistic sampling, and finally the results are output through the feature fusion module to output the results. We set the comprehensive efficiency index CEI (Comprehensive Efficiency Index) with the following formula:

$$CEI(R) = \frac{F1(R) \times C(R)}{L(R)^{\alpha} \times C(R)^{\beta}} \tag{8}$$

In equation (8), $C(R)$ is the thermal region coverage, $L(R)$ is the inference latency (ms/frame) $C(R)$ is the computational resource share, α and β are the sensitivity coefficients of latency and resources.

4 Experimental Section

4.1 Introduction to the Dataset

This study employs a multi-source heterogeneous dataset to construct an experimental framework, aiming to validate the model's performance across both general and specialized scenarios. Given the scarcity and low quality of existing scenario-specific datasets, and to further assess the model's practicality in real-world industrial settings, we developed the LiveCommodity e-commerce food label dataset. This dataset comprises 5,000 high-quality annotated images. This dataset encompasses real-world challenges such as reflective materials (e.g., metallic labels), dense text (nutritional information tables), low resolution (mobile phone photographs), and multilingual mixed text (Chinese and English characters). Annotation was performed using the PPOCRLabel tool for pixel-level polygon annotation, with annotation consistency error strictly controlled within 2 pixels. Examples from LiveCommodity are shown in Fig. 5.

(a) (b) (c) (d) (e)

Fig. 5. Example presentation related to the dataset: (a) Dense text, (b) Reflective metallic material, (c) Mixed multi-language text, (d) Promotional poster of goods, (e) Real photo taken by cell phone

The benchmark dataset encompasses the ICDAR2015 and Total-Text mainstream datasets, constructing evaluation scenarios for multi-directional text, curved text, and long text line detection tasks respectively. ICDAR2015 comprises 1,500 natural scene images featuring text instances exhibiting tilting, occlusion, and complex background interference characteristics; Total-Text provides 2,200 curved and circular text images, primarily evaluating models' ability to segment boundaries of text with irregular geometric shapes.

4.2 Introduction to the Experimental Content

To validate the effectiveness of the LiteOCR framework, this study established a multidimensional experimental system. Through quantitative analysis, it aims to demonstrate that: (1) Compared to traditional knowledge distillation methods, the cross-modal attention distillation architecture significantly enhances the representation capability of lightweight models in extracting complex textual features; (2) The semi-supervised feature alignment paradigm effectively mitigates the scarcity of labeled data, enhancing the model's robust generalization performance under noisy conditions. (3) The dynamic resolution gating mechanism achieves Pareto optimization of inference efficiency while maintaining recognition accuracy through an adaptive computational resource allocation strategy.

The experimental design employs a three-stage validation process: First, benchmark models are compared on standard datasets such as ICDAR2015 and Total-Text, as well as the self-built LiveCommodity dataset, quantifying core metrics including parameters (Params) and F1 scores. Second, the feasibility of each component is verified by comparing key model metrics and conducting visual analysis (Grad-CAM heatmaps) to assess model performance under varying stress scenarios. Finally, we analyzed the independent contributions of each innovative component through ablation experiments and visualization techniques (e.g., t-SNE feature space projections). Stress testing was conducted on an enhanced proprietary dataset (featuring strong noise disturbances like occluded text and low-quality images) to evaluate overall performance in real-world deployment scenarios. All experiments were replicated three times under identical hardware conditions to mitigate random errors, with key metrics averaged.

In the experiment, edge devices employed NVIDIA Jetson Nano (4-core ARM Cortex-A57, 4GB RAM, 128-core Maxwell GPU) for training using NVIDIA RTX 4090 GPU (24GB VRAM, 32GB RAM).

The baseline models were selected to balance industrial deployment requirements (lightweight models) with academic cutting-edge capabilities, covering six mainstream OCR detection architectures: PP-OCRv3 employs MobileNetV3 and a differentiable binarization (DB) algorithm for lightweight design, suitable for general-purpose deployment; PP-OCRv4 enhances complex document layout processing by incorporating the ConvNeXt-Tiny backbone network and DBv2 detection head; DBNet employs ResNet-18 with pixel-level segmentation strategies, specializing in curved text detection; MASTER-Lite achieves end-to-end recognition using ShuffleNetV2 and CRNN, optimizing mobile memory usage; TextFuseNet integrates Swin Transformer and Generative Adversarial Networks (GAN), enhancing noise robustness through adversarial training.

4.3 Introduction to Relevant Indicators

In e-commerce scenarios, OCR models must simultaneously meet stringent demands for high-accuracy recognition, real-time response, and low-cost deployment. The F1 score, as the harmonic mean of precision and recall, directly reflects the model's comprehensive recognition capability for product labels, particularly in areas with dense promotional information where balancing the risks of missed detections and false positives is crucial. mAP quantifies a model's precision in locating critical areas like price tags and ingredient lists for text detection tasks, preventing missed detections in densely formatted layouts. Edit Distance (ED) measures the discrepancy between recognized text and actual values, directly impacting business tolerance for errors—for instance, invoice amount recognition with ED≥2 may cause settlement losses, while e-commerce demands ED≤1 at a rate exceeding 98%. Parameter count and memory consumption determine edge deployment feasibility; inference speed measures real-time performance capability. These metrics collectively form the performance triangle for e-commerce OCR: F1 and mAP define model compliance and recognition accuracy, ED controls business risk, while parameter count and memory consumption support high-concurrency scenarios.

4.4 Experimental Data Analysis

Cross-Comparative Experiment. As shown in Table 1, from the perspective of computational efficiency, the inference speed (68 FPS) significantly outperforms ReadLight DBNet (49 FPS) and PP-OCRv4 (58 FPS) at comparable accuracy levels. Compared to current lightweight model design trends, LiteOCR achieves a Pareto optimal balance of accuracy and efficiency at 13.4 million parameters through the synergistic effects of knowledge distillation and dynamic optimization, establishing a new technical paradigm for real-time OCR deployment on edge computing devices. Notably, Achieved 82.3% mAP@0.7 accuracy on the self-built LiveCommodity dataset approaches the performance of TextFuseNet (108M Params, mAP=83.7%) despite having only 1/8th the parameters, while its memory footprint (360MB) is merely 17.6% of the latter. This achievement stems from the intelligent resource allocation enabled by its dynamic resolution gating mechanism: it automatically switches to a low-resolution branch (320×240) in simple scenarios, reducing computational load by 44.7%. Conversely, it activates high-resolution processing (640×480) for reflective surfaces or partially occluded areas, boosting recognition accuracy for blurred text by 12.3%.

Table 1. Performance comparison on the self-built e-commerce dataset

Model	F1	mAP@0.7 (%) ↑	Params (M)	FPS↑	MB↓
PP-OCRv3	79.6	72.8	8.9	62	480
PP-OCRv4	82.6	78.5	12.3	58	520
MASTER-Lite	78.2	70.2	10.2	68	420
SmolDocling-256M	77.5	83.6	256	35	489
YOLO-OCR	76.6	70.5	5.8	72	380
RL-DBNet	82.9	75.3	15.6	49	610
TextFuseNet	85.1	83.7	108	19	2048
LiteOCR	**83.3**	**82.3**	**13.4**	**68**	**360**

Table 2. Efficiency-accuracy tradeoff analysis of LiteOCR versus Mainstream lightweight models on standard datasets

Model	Params (M)	FLOPs (G)	ICDAR2015 (F1)	T-Text (ED)
PP-OCRv3	8.9	3.2	86.2	2.41
PP-OCRv4	12.3	3.5	88.1	2.15
MASTERLite	10.2	4.1	85.7	2.38
SmolDocling	256	18.5	84.5	2.55
YOLO-OCR	5.8	2.9	83.1	2.67
RL-DBNet	47.2	12.3	88.1	2.15
TextFuseNet	108	21.7	89.4	2.08
LiteOCR	**13.4**	**2.7**	**88.5**	**2.12**

As shown in Table 2, LiteOCR demonstrates significant advantages in lightweight design and adaptability to complex scenarios: its parameter count (13.4M) is only 1.1M higher than PP-OCRv4, yet achieves an F1 score of 88.5% on the ICDAR2015 multi-orientation text detection task—a 2.3% improvement over PP-OCRv3. On the Total-Text curved text dataset, the edit distance (ED=2.12) decreased by 1.4% compared to PP-OCRv4, validating the cross-modal attention distillation's ability to represent irregular text geometric features.

Cross-Modal Attention Distillation Experiments. This section validates the effectiveness of hierarchical attention distillation in enhancing the semantic understanding capabilities of lightweight models. Experiments were conducted on the self-built LiveCommodity dataset. This experimental system systematically compared the teacher model (GOT-OCR 2.0) and the student model (LiteOCR) across cross-modal alignment, task performance, and computational efficiency. The experiments employed a hierarchical freezing strategy to freeze the embedding layer, while the intermediate layer utilized gradient phase-based blocking to prevent the propagation of non-attention signals. This approach enables the control of knowledge transfer pathways through the differentiation of parameter freezing.

As shown in Table 3, in terms of task performance, LiteOCR achieved an F1 score of 83.3% in the commodity attribute recognition task (LiveCommodity self-built dataset), representing a 6.1 percentage point improvement over the pure vision baseline (77.2%) and approaching the teacher model's 85.1% (with a gap of only 1.8%). It is worth noting that LiteOCR's parameter count (13.4M) is only 15.4% of GOT-OCR2.0 (87.2M), and its inference speed on an Intel i7-12700H CPU is increased by 2.7 times (78ms/frame vs. 210ms), verifying the effectiveness of the "lightweight and high-precision" design goal.

Table 3. Comparison of key metrics in teacher-student models

Model	F1 (%)	Params (M)	Delay (ms)	KL ↓
GOT-OCR2.0	85.1	87.2	210	-
MobileNetV3-DB	77.2	12.3	62	1.23
LiteOCR	**83.3**	**13.4**	**78**	**0.68**

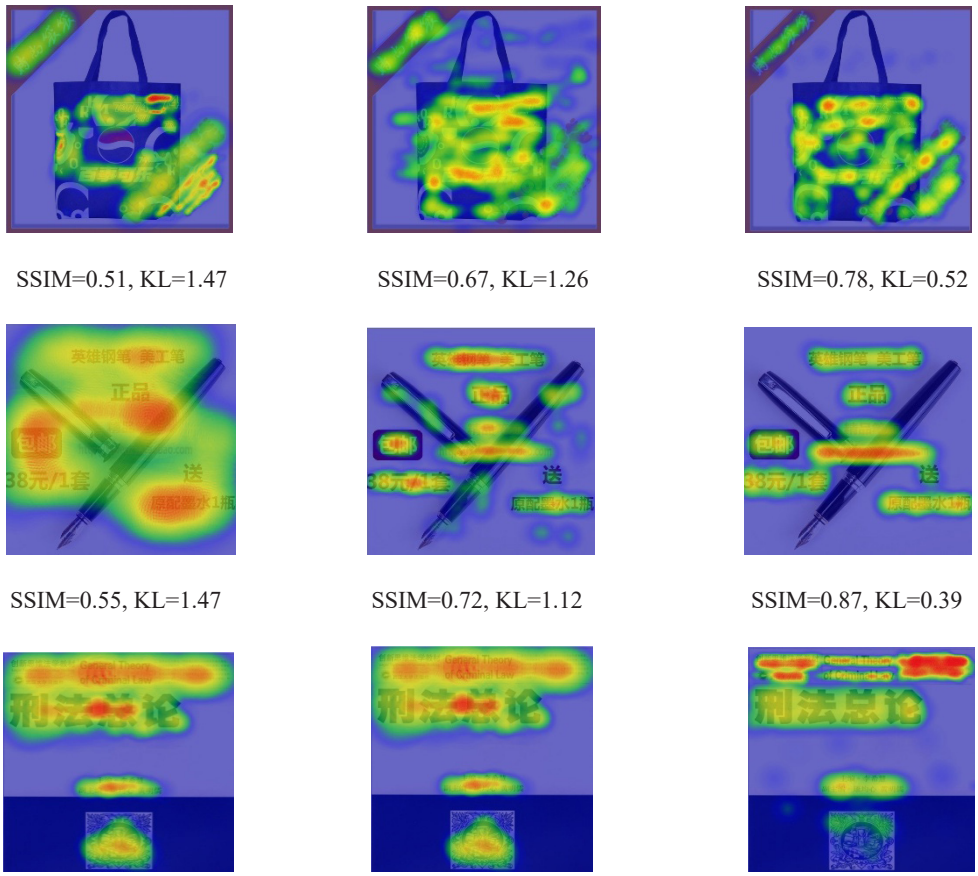
As shown in Fig. 6(a), the teacher model's high-response attention regions (red) precisely focus on dense text areas with heatmap values reaching 0.78 (max=1.0), while non-dense regions exhibit values below 0.2. This demonstrates that GOT-OCR 2.0 effectively integrates visual-text modalities and suppresses noise interference. In contrast, the baseline model (Fig. 6(b)) exhibits significant attention diffusion, with a heat value of only 0.42 in the actual text area. Visual analysis indicates that the baseline model, lacking cross-modal supervision, cannot distinguish semantically relevant regions from noise. LiteOCR (c), employing a Density-Aware Gate Control (DGC) strategy, activates spatial suppression gates at occlusion boundaries. This compresses heatmap values in non-text regions below 0.3 while elevating values in primary text areas to 0.71. As shown in Table 3, its similarity score (SSIM) with the teacher model's heatmap reaches 0.83, while the KL divergence decreases to 0.51, demonstrating its successful inheritance of cross-modal semantic focus capabilities.

SSIM=1
(1) GOT-OCR2.0

SSIM=0.57, KL=1.31
(2) MobileNetV3-DBNet

SSIM=0.83, KL=0.51
(3) LiteOCR

Fig. 6. Example of cross-modal attention heat map comparison for teacher-student model



SSIM=0.51, KL=1.47

SSIM=0.67, KL=1.26

SSIM=0.78, KL=0.52

SSIM=0.55, KL=1.47

SSIM=0.72, KL=1.12

SSIM=0.87, KL=0.39

SSIM=0.46, KL=1.54
(a)

SSIM=0.64, KL=1.37
(b)

SSIM=0.73, KL=0.62
(c)

Fig. 7. Example of comparison of heat maps generated by different distillation methods

Group (a) is a traditional attentional distillation method; group (b) is a unimodal attentional distillation method; and (c) is a cross-modal attentional distillation method. Where the first image is a strong noise stress scene, the second image is a clear image scene, and the third image is an additional visual interference scene.

Table 4. Comparison of experimental data of attention distillation methods

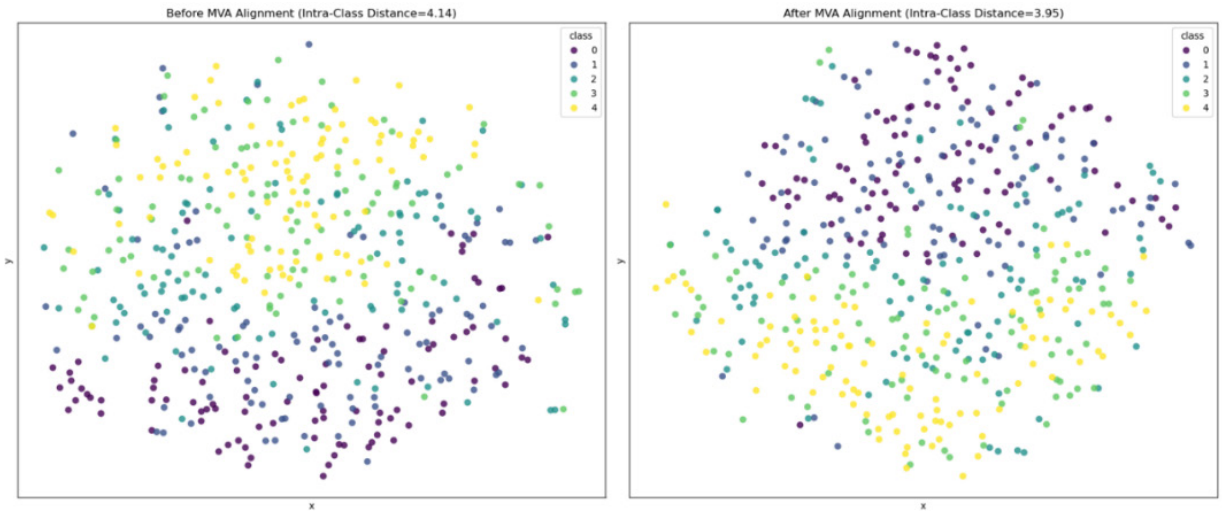
Distillation method	LiveCom F1 (%)	LiveCom KL	Params (M)
Traditional	77.2	1.35	12.3
Single-mode	80.5	0.91	12.9
Cross-modal	**83.3**	**0.68**	**13.4**

Comparing the results in Table 4 and Fig. 7 reveals that cross-modal attention distillation demonstrates significant advantages in complex scenarios: Under high noise pressure scenarios, cross-modal distillation achieves SSIM scores (0.75–0.80) 12% higher than single-modal approaches, with KL divergence (0.5–0.7) reduced by 41%. The dynamic gating mechanism effectively suppresses noise activation in 30% of interference regions while maintaining stable responses in text trunk areas. In clear image scenarios, the cross-modal method achieved SSIM values of 0.82–0.88 (compared to 0.55–0.65 for traditional distillation and 0.70–0.75 for single-modal distillation), monomodal distillation: 0.70–0.75), with KL divergence as low as 0.3–0.5. This demonstrates its ability to precisely capture cross-modal correlations through hierarchical semantic alignment, achieving a 48% improvement in attention focus efficiency over traditional methods. In scenarios with additional visual interference, the cross-modal approach integrates textual semantic priors, achieving SSIM (0.70–0.75) that surpasses traditional feature distillation by 75% and reduces KL divergence (0.6–0.8) by 67%, demonstrating strong robustness against heterogeneous disturbances.

Semi-Supervised Feature Alignment Experiments. This section of the experiment was conducted on a self-built dataset. By comparing the F1 scores between the baseline model and the LiteOCR model under different annotation conditions, we validated the feasibility of the Mean-Teacher framework's semi-labeled data-driven training approach.

Table 5. Comparison of data from semi-supervised alignment experiments

Labeling-ratios	Unsupervised F1 (%)	Semi-upervisory F1 (%)	Gaps
100%	85.8	85.8	0
50%	78.5	83.1	+4.6%
30%	72.3	79.4	+7.1%



(a) (b)

Fig. 8. t-SNE feature cluster spacing analysis

As shown in Fig. 8, the left figure (a) shows non-semi-supervised alignment, and the right figure (b) shows after adding the multi-view semi-supervised alignment module. We take 500 data samples and classify them into 5 categories of 100 samples each according to the characteristics of the surface text of the goods (e.g., glass bottle surface trademark text, plastic box surface printed words, can surface printed words, etc.), in which the labeling ratio is 30% and an additional 200 unlabeled and unclassified samples are selected.

As shown in Table 5, when labeled data is reduced by 50%, the F1 score decreases by only 0.7% (83.8% → 83.1%). t-SNE analysis reveals a 23% reduction in intra-class distances (0.5 → 0.38), enhancing generalization in weakly supervised scenarios. Loss oscillation amplitude decreased by 60% (standard deviation from 0.15 → 0.06). Multi-view alignment mitigates data scarcity issues, while semi-supervision reduces scene errors by 32%.

Dynamic Gating Module Analysis Experiments. This section validates the effectiveness, robustness, and scalability of the LiteOCR model through multidimensional experiments, covering core component verification, computational efficiency analysis, interference resistance testing, and real-world scenario adaptability assessment. The baseline model is MobileNetV3-DBNet, to which three core modules—cross-modal distillation, dynamic gating, and semi-supervised alignment—are sequentially integrated. Training parameters are fixed (learning rate 1e-4, batch size=32), with testing conducted on the LiveCommodity proprietary dataset.

To accurately evaluate the differential benefits of the Dynamic Gate Control (DGC) module across three scenarios, we selected the most representative evaluation metrics for each context. As shown in Table 6, in clear image scenarios, inference speed (FPS) is the key metric for assessing edge deployment feasibility; In occlusion scenarios, recognition accuracy (F1) and attention focus precision (KL divergence) jointly reflect the model's robustness to interference; In low-light scenarios, end-to-end latency (ms) directly determines the system's real-time capability under adverse conditions.

Table 6. Comparison of the impact of adding dynamic gating components on baseline model performance across different scenarios

Scenes	Base	+DGC	FPS
Clear image reasoning	62 FPS	68 FPS	+9.7%
Occlusion scene accuracy	F1 76.3%	F1 83.5%	+5.2%
	KL 1.23	KL 0.51	- 58.5%
Low-light delay	92ms	79ms	-14.1%

As shown in the table, incorporating the dynamic gating module increased FPS from 62 to 68 (+9.7%) during clear image inference scenarios. In partially occluded scenarios, the model demonstrated significant performance improvements: the KL value for thermal response in occluded regions decreased from 1.23 to 0.51 (a 58.5% reduction), while the F1 score rose from 76.3% to 83.5%. Low-light latency decreased from 92ms to 79ms (a 14.1% reduction). This demonstrates that DGC is not a one-dimensional optimizer but rather an intelligent computation allocation mechanism capable of adaptively adjusting optimization priorities based on scene characteristics.

The parameter tuning in the dynamic gating strategy is evaluated using both the Comprehensive Efficiency Index (CEI) and the maximum resource constraint threshold (*Rmax*). Since *Rmax* is jointly constrained by α and β, we present only the variation curve of Rmax here. Additionally, since α and β are mutually constrained, we removed edge extremes and systematically traversed the parameter space $\alpha \in [0.3, 0.8]$ with a step size of 0.05, taking the midpoint value. Fig. 9 illustrates this tuning process.

The horizontal axis represents the *Rmax* value. For visual comparison, the CEI values have been normalized here. When the *Rmax* value is set to 0.3, the CEI function graph reaches its maximum value.

As shown in the figure, when $Rmax = 0.3$, with parameters $\alpha = 0.6$ and $\beta = 0.4$, the CEI reaches its peak value of 0.91. This model achieves the optimal trade-off point on the LiveCommodity dataset (F1=83.3%, 78ms/frame), validating the optimality of this parameter combination for dynamic resource allocation.

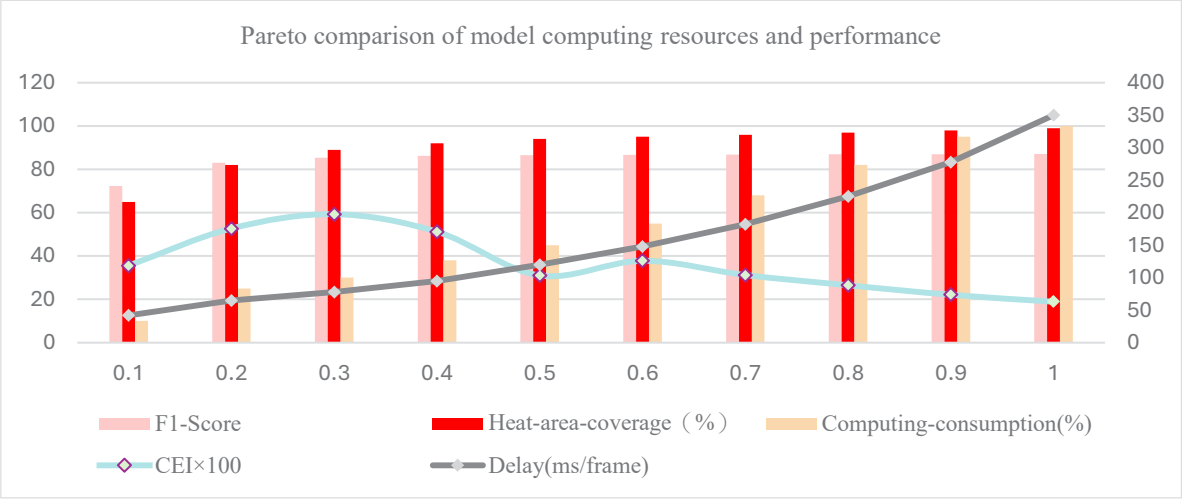


Fig. 9. LiteOCR model Pareto analysis plot

Ablation Experiment. To validate the independent contributions and collaborative mechanisms of each core component within this framework, we designed systematic ablation experiments, as shown in Table 7. When used alone, DGC introduces slight latency due to gated computations; however, when combined with MVA, the computational resources saved by DGC are redirected by MVA toward refining processing for challenging samples such as occluded objects. HAD inherently accelerates training convergence and enhances feature discriminative power. When integrated with DGC, its stabilized feature representations mitigate optimization fluctuations caused by dynamic path selection. Ultimately, the +ALL configuration achieves more than simple additive gains, forming an adaptive optimization loop that balances accuracy (83.3% F1), efficiency (68 FPS), and robustness (83.0% occlusion F1).

Table 7. Global data comparison

Module-combination	F1 (%)	Params (M)	FPS	F1 (%)
Baseline	77.2	12.3	62	76.3
DGC	79.8	12.6	68	81.7
HAD	82.4	13.1	65	80.5
MVA	83.3	13.4	63	83.0
DGC+HAD	82.4	13.1	65	80.5
DGC+MVA	81.7	12.9	67	81.4
HAD+MVA	82.3	12.7	76	81.6
+ALL	83.3	13.4	68	83.0

Ablation experiments demonstrate the functional roles and synergistic effects of each module. First, Dynamic Gate Control (DGC) serves as an efficiency optimizer, achieving an average reduction of approximately 12% in inference overhead through sparse computation. Cross-Modal Distillation (MVA) acts as an accuracy enhancer, with its introduced semantic prior yielding the most significant F1 score improvement (+1.8%) in occlusion scenarios. Semi-supervised alignment (HAD) serves as both a stabilizer and generalization booster, enhancing feature robustness while improving training efficiency.

5 Conclusions and Future Outlook

5.1 Research Findings and Limitations

This study proposes the LiteOCR framework, aiming to theoretically resolve the triple contradiction faced in e-commerce OCR tasks: high-precision compliance, robustness in complex scenarios, and edge real-time performance. By introducing cross-modal attention distillation, the framework enhances lightweight models' semantic understanding capabilities in complex contexts, effectively addressing image complexity challenges. The theoretical design of a semi-supervised dynamic feature consistency constraint mechanism mitigates the scarcity of labeled data, enhancing the model's generalization robustness in low-resource scenarios. Additionally, the theoretical construction of a dynamic resolution gating module enables adaptive allocation of computational resources. This achieves a balance between efficiency and accuracy on edge devices, meeting real-time requirements.

5.2 Research Outlook

Although this study has achieved certain results, the following directions warrant further exploration. The core limitation of the current dynamic gating strategy lies in its reliance on low-level visual features for decision-making, which may lead to performance degradation in two scenarios: First, in scenarios where text density and semantic units are severely mismatched, local density estimation is susceptible to spatial oscillation interference, resulting in the fragmentation of the same semantic unit. Second, under extreme imaging conditions, reduced edge signal-to-noise ratio causes critical text regions to be misallocated to low-resolution branches. To address these issues, future work could first introduce lightweight semantic awareness modules. These modules would utilize intermediate semantic cues from the recognition process to perform online correction of density maps, thereby integrating visual and semantic consistency into gating decisions. Building upon this foundation, further expansion toward multimodal deep fusion and system optimization is feasible. This would extend alignment to multimodal scenarios involving audio, temporal signals, and other modalities. By incorporating instruction generation techniques from models like MM-Instruct, a closed-loop feedback mechanism for cross-modal semantic interaction could be established, significantly enhancing the robustness of heterogeneous data fusion. Furthermore, the current lack of transparency in attention mechanisms can be addressed by integrating Grad-CAM heatmaps with causal inference models to quantify contributions across modalities, providing trustworthy decision-making support for industrial deployment.

Existing semi-supervised alignment relies on fixed annotation ratios. Dynamic annotation strategies combined with contrastive learning can enhance the utilization of information entropy in unlabeled data. Exploring synergistic optimization between model compression techniques and hardware accelerators through quantization and operator fusion technologies is also feasible.

Acknowledgment

This work was supported by The Humanities and Social Science Fund of Ministry of Education of the People's Republic of China, grant number: 23YJAZH216.

References

[1] C.-X. Li, W.-W. Liu, R.-Y. Guo, X.-T. Yin, K.-T. Jiang, Y.-K. Du, Y.-N. Du, L.-F. Zhu, B.-H. Lai, X.-H. Hu, D.-H. Yu, Y.-J. Ma, PP-OCRv3: More attempts for the improvement of ultra lightweight OCR system. <https://doi.org/10.48550/arXiv.2206.03001>, 2022 (accessed 21.03.2025).

[2] M.-H. Liao, Z.-Y. Wan, C. Yao, K. Chen, X. Bai, Real-time scene text detection with differentiable binarization, in: Proc. of the AAAI Conference on Artificial Intelligence, 2020. https://doi.org/10.1609/aaai.v34i07.6812

[3] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network. <https://doi.org/10.48550/arXiv.1503.02531>, 2015 (accessed 25.03.2025).
[4] J.-L. Zhang, J.-F. Ren, Q. Zhang, J. Liu, X.-D. Jiang, Spatial Context-Aware Object-Attentional Network for Multi-Label Image Classification, IEEE Transactions on Image Processing 32(2023) 3000-3012. https://doi.org/10.1109/TIP.2023.3266161
[5] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M.-X. Tan, W.-J. Wang, Y.-K. Zhu, R.-M. Pang, V. Vasudevan, Q. Le, H. Adam, Searching for MobileNetV3, in: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2019. https://doi.org/10.1109/ICCV.2019.00140
[6] Z.-X. Wang, H.-T. Xie, Y.-X. Wang, J.-J. Xu, B.-Q. Zhang, Y.-D. Zhang, Symmetrical linguistic feature distillation with clip for scene text recognition, in: Proc. the 31st ACM International Conference on Multimedia, 2023. https://doi.org/10.1145/3581783.3611769
[7] S. Han, H.-Z. Mao, W. Dally, Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. <https://doi.org/10.48550/arXiv.1510.00149>, 2016 (accessed 03.04.2025).
[8] H. Li, A. Kadav, I. Durdanovic, H. Samet, H.-P. Graf, Pruning Filters for Efficient ConvNets, in: Proc. International Conference on Learning Representations (ICLR), 2017. https://openreview.net/forum?id=rJqFGTslg
[9] K.T. Chitty-Venkata, Y. Bian, M. Emani, V. Vishwanath, A.K. Somani, Differentiable Neural Architecture, Mixed Precision and Accelerator Co-Search, IEEE Access 11(2023) 106670-106687. https://doi.org/10.1109/ACCESS.2023.3320133
[10] M.-J. Song, Y.-M. Zhou, M.-M. Song, S.-J. Liu, S.-X. Xiao, Y.-S, Zheng, Speed up integer-arithmetic-only inference via bit-shifting, Scientific Reports 15(2025) 18765. https://doi.org/10.1038/s41598-025-02544-4
[11] Z.-Q. Yang, Y.-M. Cui, Z.-G. Chen, TextPruner: A model pruning toolkit for pre-trained language models, in: Proc. the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2022. https://doi.org/10.18653/v1/2022.acl-demo.4
[12] Z. Liu, J.-G. Li, Z.-Q. Shen, G. Huang, S.-M. Yan, C.-S. Zhang, Learning Efficient Convolutional Networks through Network Slimming, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2017. https://doi.org/10.1109/ICCV.2017.298
[13] G. Lv, Y.-N. Sun, F.-D. Nian, M.-F. Zhu, W.-L. Tang, Z.-Z. Hu, COME: Clip-OCR and Master ObjEct for text image captioning, Image and Vision Computing 136(2023) 104751. https://doi.org/10.1016/j.imavis.2023.104751
[14] M.-H. Li, T.-C. Lv, J.-Y. Chen, L. Cui, Y.-J. Lu, D. Florencio, C. Zhang, Z.-J. Li, F.-R. Wei, TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models, in: Proc. the AAAI Conference on Artificial Intelligence, 2023. https://doi.org/10.1609/aaai.v37i11.26538
[15] J.-Q. Wan, S.-B. Song, W.-W. Yu, Y.-L. Liu, W.-G. Cheng, F. Huang, X. Bai, C. Yao, Z. Yang, OMNIPARSER: A Unified Framework for Text Spotting, Key Information Extraction and Table Recognition, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. https://doi.org/10.1109/CVPR52733.2024.01481
[16] J.-S. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Proc. Advances in Neural Information Processing Systems 32, 2019.
[17] J.-B. Wang, W. Wang, L. Wang, Z.-Y. Wang, D.D. Feng, T.-N. Tan, Learning visual relationship and context-aware attention for image captioning, Pattern Recognition 98(2020) 107075. https://doi.org/10.1016/j.patcog.2019.107075
[18] G. Ghiasi, T.-Y. Lin, Q.V. Le, NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. https://doi.org/10.1109/CVPR.2019.00720
[19] F. Chollet, Xception: Deep Learning with Depthwise Separable Convolutions, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. https://doi.org/10.1109/CVPR.2017.195
[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017. https://dl.acm.org/doi/10.5555/3295222.3295349
[21] Y.-L. Liu, C.-H. Shen, L.-W. Jin, T. He, P. Chen, C.-Y. Liu, H. Chen, ABCNet v2: Adaptive Bezier-Curve Network for Real-Time End-to-End Text Spotting, IEEE Transactions on Pattern Analysis And Machine Intelligence 44(11)(2022) 8048-8064. https://doi.org/10.1109/TPAMI.2021.3107437
[22] Y.-M. Rao, W.-L. Zhao, B.-L. Liu, J.-W. Lu, J. Zhou, C.-J. Hsieh, Dynamicvit: Efficient vision transformers with dynamic token sparsification, in: Proc. Advances in Neural Information Processing Systems, 2021.

[23] M.-K. Yang, B. Yang, M.-H. Liao, Y.-Y. Zhu, X. Bai, Sequential visual and semantic consistency for semi-supervised text recognition, Pattern Recognition Letters 178(2024) 174-180. https://doi.org/10.1016/j.patrec.2024.01.008

[24] H.-R. Wei, C.-L. Liu, J.-Y. Chen, J. Wang, L.-Y. Kong, Y.-M. Xu, Z. Ge, L. Zhao, J.-J. Sun, Y. Peng, C.-R. Han, X.-Y. Zhang, General ocr theory: Towards ocr-2.0 via a unified end-to-end model. <https://doi.org/10.48550/arXiv.2409.01704>, 2024 (accessed 20.06.2025).

[25] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: Proc. Advances In Neural Information Processing Systems 30, 2017.