YOLO11-DAN: A Retail Commodity Detection and Generalization Method for Dense Shelf Scenarios

Authors: Xiangyu Gao

Date: 2026-03-05T17:48:30+00:00

Abstract

Conducting retail object detection in dense shelf scenarios encounters numerous challenges, such as highly dense targets, extremely similar appearances, and variable scales, which limit the precision of existing general detection models and result in insufficient cross-domain generalization capabilities. In response to this situation, a retail object detection model specifically designed for dense shelf scenarios is proposed, namely YOLO11-DAN (YOLO11 with Density-Aware Network).

First, a density-adaptive non-maximum suppression module is designed to achieve dynamic threshold adjustment through local density estimation, strengthening the suppression of false positives in sparse areas and minimizing missed detections in relatively dense areas. Subsequently, a three-stage optimization loss function suitable for dense scenarios is proposed, introducing density-sensitive weight adjustment, hard sample reinforcement learning, and scale-adaptive penalty mechanisms into the bounding box regression process to improve the localization accuracy of the model for overlapping targets and multi-scale commodities.

On this basis, a curriculum learning training framework built upon difficulty assessments across three dimensions—density, scale, and appearance—is constructed, combined with cross-domain data augmentation strategies to further enhance the model's generalization ability from retail commodity datasets (SKU-110K) to crowd-dense scenarios (CrowdHuman). Experimental results demonstrate that the proposed method achieves stable performance gains on the SKU-110K dataset: the mean Average Precision (mAP) increased by 2.4% compared to the baseline YOLO11, the recall rate improved by 5.3%, and other indicators also achieved varying degrees of improvement. On the CrowdHuman dataset, the mAP increased relatively by 6.9% and the recall rate improved by 5.8%.

These results fully verify the detection advantages of the model in dense scenarios and effectively demonstrate its cross-domain generalization capability and practical application effectiveness.

Full Text

Preamble

YOLO11-DAN: A Retail Object Detection and Generalization Method for Dense Shelf Scenarios

Gao Xiangyu, Mo Jianhua, Xiong Jun, Zhou Zhengwu, Zhou Yueshuai

(School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430000, China)

Corresponding Author: Gao Xiangyu, E-mail: 759448932@qq.com

Abstract: Conducting retail object detection in dense shelf scenarios presents numerous challenges, including highly crowded targets, extremely similar appearances, and significant scale variations. These factors limit the precision of existing general-purpose detection models and result in insufficient cross-domain generalization capabilities. To address these issues, this paper proposes a retail object detection model specifically designed for dense shelf environments, named YOLO11-DAN (YOLO11 with Density-Aware Network). First, a Density-Adaptive Non-Maximum Suppression (DA-NMS) module is designed to achieve dynamic threshold adjustment through local density estimation. This module strengthens the suppression of false positives in sparse regions while minimizing missed detections in relatively dense areas. Subsequently, a three-stage optimized loss function suitable for dense scenarios is proposed. This function introduces density-sensitive weight adjustment, hard example reinforcement learning, and a scale-adaptive penalty mechanism into the bounding box regression process, thereby improving the model' s localization accuracy for overlapping targets and multi-scale products. Building on this, a curriculum learning training framework is constructed based on difficulty assessments across three dimensions: density, scale, and appearance. Combined with cross-domain data augmentation strategies, this framework further enhances the model' s generalization capability from retail product datasets (SKU-110K) to crowded human scenarios (CrowdHuman).

Experimental results show that the proposed method achieves stable performance gains on the SKU-110K dataset: compared to the baseline YOLO11, the mean Average Precision (mAP) increased by 2.4%, while the recall rate improved by 5.3%, with other performance metrics also showing varying degrees of enhancement. On the CrowdHuman dataset, the mAP achieved a relative improvement of 6.9%, and the recall rate increased by 5.8%. These results fully demonstrate the detection advantages of the proposed model in dense scenarios, while effectively highlighting its cross-domain generalization capabilities and

practical utility.

Keywords: Object detection; YOLO; Density-adaptive network; Training framework optimization; Retail products

Citation Format: Gao Xiang-yu, Mo Jian-hua, Xiong Jun, Zhou Zheng-wu, Zhou Yue-shuai. YOLO11-DAN: Generalizable Retail Product Detection in Dense Shelf Scenes. *Computer Systems & Applications*, xxxx, xx(x): x–x. http://www.c-s-a.org.cn/1003-3254/xxxx.html

1 Introduction

Automated retail product detection is a key technology in the digital transformation of the retail industry. By utilizing computer vision algorithms to precisely identify and locate products on shelves, this technology facilitates intelligent inventory management, dynamic pricing, and automated checkout systems, providing significant data-level support for various retail applications [?].

However, real-world retail environments are characterized by extreme complexity, posing severe challenges to the practical implementation of related technologies. Specifically, these challenges manifest in several ways: products are often displayed densely, leading to severe occlusion and adhesion between items; products exhibit high inter-class similarity (within the same Stock Keeping Unit, or SKU) alongside intra-class diversity caused by promotional labels and lighting variations; and there are significant differences and variability in product dimensions, which makes small object detection particularly difficult [?]. These factors cause general-purpose detection models—such as those trained on the COCO dataset—to suffer substantial performance degradation when directly transferred to retail scenarios, making it difficult to meet the high precision and robustness requirements of practical applications.

Current object detection methods are generally categorized into two types: two-stage methods (such as R-CNN [?]) and single-stage methods (such as YOLO [?] and SSD). While two-stage methods typically offer higher precision, their slower processing speeds make it difficult to meet the requirements for real-time performance. Single-stage methods have become the preferred solution for real-time detection due to their effective balance between speed and accuracy. Within the context of retail scenarios, a vast body of research has focused on improving the YOLO framework. These enhancements include strengthening backbone networks by introducing attention mechanisms [?] and optimizing feature pyramids [?], refining loss functions (such as the IoU loss series [?]), improving post-processing algorithms (such as Soft-NMS [?] and Adaptive NMS [?]), and employing more advanced training strategies like curriculum learning [?].

However, most existing improvements are general optimizations that fail to provide targeted modeling for the core challenges of “high density and high similarity" inherent in retail scenarios. Furthermore, a systematic optimization framework centered on density information has yet to be established, and the cross-domain generalization capabilities of these methods require further enhancement. The emergence of the SKU-110K [?] dataset has significantly advanced research in dense object detection, yet current methods still struggle to meet the rigorous requirements for precision and generalization in complex retail environments [?].

To address the aforementioned challenges, this paper proposes a novel density-aware detection model: YOLO11-DAN (YOLO11 with Density-Aware Network). Its core innovation lies in transforming spatial density from a passive environmental description into an active optimization variable. This approach facilitates the construction of a density-aware mechanism that permeates the entire detection process. The primary contributions of this work can be summarized in the following three aspects:

1. We propose a Density-Adaptive Non-Maximum Suppression (DA-NMS) module. Unlike traditional NMS strategies, this module innovatively introduces real-time local density estimation to dynamically adjust suppression thresholds, balancing recall and precision in complex distributions.
2. We design a three-stage optimized loss function specifically tailored for dense scenarios, incorporating density-sensitive weight adjustment, a hard-sample focusing mechanism, and scale-adaptive penalties to improve localization accuracy.
3. We implement a curriculum learning strategy for robust feature extraction. By moving from simple to complex samples based on a three-dimensional difficulty evaluation, and incorporating cross-domain data augmentation, we enhance the model' s stability and generalization.

---

2 Proposed Method: YOLO11-DAN

2.1 YOLO11 Network Model Overview

The YOLO11 model framework consists of four primary components: the input terminal, the backbone network, the feature fusion network, and the detection head. The backbone network utilizes improved C2PSA and C3k2 modules as its core components. Specifically, the C2PSA module enhances feature selection capabilities by combining convolutional layers with attention mechanisms. Meanwhile, the C3k2 module achieves efficient feature reuse through the implementation of a Bottleneck structure and branch concatenation operations. With the introduction of the SPFF structure and pooling layers, the model' s multi-scale feature extraction capabilities have been further enhanced.

2.2 YOLO11-DAN Network Model Framework

The proposed collaborative optimization framework is integrated into a unified structure, as illustrated in [Figure 3: see original paper]. A primary limitation of standard models lies in post-processing suppression strategies, which remain static. To address this, the DA-NMS module is introduced at the end of the pipeline. Furthermore, the original bounding box regression loss is replaced with the newly proposed V3IoU loss function to provide clearer gradients for dense and challenging samples. Finally, we construct a curriculum learning framework to improve domain adaptation.

[Figure 3: see original paper] YOLO11-DAN network architecture diagram

2.3 Density-Adaptive Non-Maximum Suppression Module

Traditional NMS fails to utilize information regarding local spatial density. This module performs real-time estimation of local density to dynamically adjust the NMS threshold. [Figure 4: see original paper] presents a comparison between the proposed method and traditional NMS.

[Figure 4: see original paper] Schematic diagram of the core concept and workflow of the DA-NMS algorithm.

2.3.1 Local Density Estimation Algorithm Given a set of raw detection boxes $B = \{b_i\}_{i=1}^{N}$, the center point of each box is calculated. To avoid high computational costs, the image plane is partitioned into a $G \times G$ grid (setting $G = 10$). The initial density d_{pq} is the count of box centers in grid cell (p, q). For each box b_i, we consider a 3×3 local neighborhood N_i to calculate the local density score ρ_i:

$$\rho_i = \frac{1}{|N_i|} \sum_{(p,q) \in N_i} d_{pq}$$

2.3.2 Dynamic Threshold Function We designed a nonlinear mapping function $T(\rho_i)$ to determine the IoU threshold τ_i:

$$\tau_i = T(\rho_i) = \tau_{base} - \alpha \cdot \text{sigmoid}\left(\gamma \cdot (\rho_i - \beta)\right)$$

where τ_{base} is the base threshold (0.5), α controls adjustment magnitude (0.15), β is the density offset, and γ is a scaling factor.

2.4 Three-Stage Optimized Loss Function

We propose V3IoU Loss to enhance performance where targets frequently overlap.

2.4.1 Loss Function Definition

$$L_{V3IoU} = \lambda_{IoU} L_{IoU} + \lambda_{center} L_{center} + \lambda_{aspect} L_{aspect} + \lambda_{dense} L_{dense}$$

The overall architecture is illustrated in [Figure 6: see original paper].

[Figure 6: see original paper] Overall architecture of the V3IoU loss function.

2.4.2 Density-Sensitive Weight Adjustment (Stage 1) The density penalty term L_{dense} imposes stronger geometric constraints in crowded regions:

$$L_{dense} = \rho_{norm} \cdot L_{center}$$

2.4.3 Scale-Adaptive Penalty (Stage 2) We adopt a differentiated aspect ratio penalty strategy. A coefficient $\omega_{small} < 1$ relaxes shape constraints on small targets, while $\omega_{large} > 1$ strengthens them for large targets.

2.4.4 Hard Sample Focusing Mechanism (Stage 3) We design an adaptive sample weight w_{hard}:

$$w_{hard} = 1 + \eta \cdot \exp(-\kappa \cdot \mathrm{IoU})$$

This assigns larger gradients to hard samples during backpropagation.

2.5 Density-Aware Curriculum Learning Framework

2.5.1 Three-Dimensional Sample Difficulty Evaluation Difficulty scores D are calculated across three dimensions: Density (d_{den}), Scale (d_{scale}), and Appearance (d_{app}), as shown in [Figure 7: see original paper].

$$D = \omega_1 \cdot d_{den} + \omega_2 \cdot d_{scale} + \omega_3 \cdot d_{app}$$

[Figure 7: see original paper] Schematic diagram of three-dimensional sample difficulty assessment.

2.5.2 Progressive Curriculum Scheduling Training is divided into K stages. A threshold function $\theta(e)$ gradually increases, allowing more difficult samples into the training pool over time, as shown in [Figure 8: see original paper].

[Figure 8: see original paper] Schematic diagram of the progressive curriculum learning scheduling strategy.

---

3 Experiments and Analysis

3.1 Experimental Setup

Experiments were conducted on an NVIDIA GeForce RTX 3090 GPU using PyTorch 2.3.1. We utilized the SKU-110K dataset [Figure 1: see original paper] for training and the CrowdHuman dataset [Figure 2: see original paper] for cross-domain generalization testing.

[Figure 1: see original paper] Original images from the SKU-110K dataset. [Figure 2: see original paper] Original images from the CrowdHuman dataset.

3.2 Ablation Study

presents the results of the ablation experiments on SKU-110K.

Ablation results of YOLO11-DAN modules.

The integration of all three modules achieved an mAP of 92.8%, a 2.4% increase over the baseline. The collaborative optimization allows the model to achieve a superior balance between recall and precision.

3.3 Comparative Experiments

We compared YOLO11-DAN against mainstream models on SKU-110K and CrowdHuman.

Comparative experimental results on the SKU-110K dataset.

On SKU-110K, YOLO11-DAN achieved an mAP50-95 of 61.8%, outperforming the baseline by 4.3%.

Comparative experimental results on the CrowdHuman dataset.

In zero-shot testing on CrowdHuman, YOLO11-DAN achieved an mAP50 of 72.3%, a 6.9% improvement over the baseline, demonstrating robust cross-domain generalization.

---

4 Conclusion

This paper proposes YOLO11-DAN, which transforms scene density into an active optimization variable. By integrating density-adaptive NMS, a three-stage loss function, and a curriculum learning framework, the model significantly improves detection performance in dense retail scenarios and exhibits strong generalization to other dense domains like crowd detection. Future work will explore applications in remote sensing and cell counting.

Note: Figure translations are in progress. See original paper for figures.