Proceedings of the 5th International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/2026.31762

A Multi-scale Differentiated Feature Fusion Network for Real-time Tomato Maturity Detection

Zhening Xu[1], Yufei Lu[2], Geqi Xia[3], Heng Yi[4], Wei Chen[5*]

[1]Department of Vehicle Engineering, Nanjing Institute of Technology, Nanjing, China
[2]Department of Robotics Engineering, Nanjing Institute of Technology, Nanjing, China
[3]Department of Automotive Service Engineering, Nanjing Institute of Technology, Nanjing, China
[4]Department of Electronic Information Engineering, Nanjing Institute of Technology, Nanjing, China
[5]Department of Tianyin Lake Science and Technology Innovation College, Nanjing Institute of Technology, Nanjing, China
**Corresponding Author. Email: chenwei@njit.edu.cn*

Abstract. To address low accuracy and poor robustness in tomato maturity detection under complex agricultural environments (light variations, occlusion, multi-scale targets), this study proposes a lightweight multi-scale differential fusion network Tomato Net-MSF. Its core innovation is scale-customized feature enhancement modules: (1) Small-scale KSFA module (multi-core selection + spatial-spectral attention) suppresses background and improves small-target recognition; (2) Medium-scale C3k2l module (2*2 simplified convolution + dual-bottleneck stacking) enhances fusion stability; (3) Large-scale LSConv module (global perception + dynamic aggregation) refines target contours. Experiments on Laboro Tomato dataset show: mAP50 0.75 (8% higher than baseline YOLOv11), F1 0.72 (6% higher), and 50FPS inference speed (20ms/frame) on RK3588NPU (5-fold acceleration). The model optimizes multi-scale detection balance, integrating high precision, robustness and edge deployment capability, providing an efficient solution for real-time tomato maturity detection in complex scenarios.

Keywords: Computer vision, Edge computing, Tomato maturity detection, Agricultural robots

1. Introduction

Accurate tomato ripeness recognition is a core bottleneck for smart agricultural harvesters [1]. Field deployment faces key challenges: 28% small-target miss rate in multi-scale detection; <65% accuracy of HSV model for half-ripe fruits (hue 110-140°) under severe illumination; 17.2% false detection rate when branch/leaf occlusion >30%; and <0.68 F1-score for transitional ripeness, failing refined harvesting needs.Existing color-texture fusion methods lack generalization. YOLO series dominates agricultural detection for real-time performance (latency <22ms, speed >45 FPS) [2]. YOLOv4 (CSPDarknet, -20% computation), YOLOv8 (anchor-free, +3.8% accuracy), YOLOv11 (-35% parameters) evolve lightweight, but 72% small-target recall still mismatches

DOI: 10.54254/2753-8818/2026.31762

agricultural scenarios.Target detection optimization focuses on feature fusion, dynamic selection and large-kernel convolution (e.g., FPN/PAN, CSPNet). However, RGB-D fusion and knowledge distillation face cost/accuracy issues; existing grading systems confuse transitional ripeness [3].

2. Research methods

2.1. Overview of the overall architecture

Building upon the YOLOv11 detector framework, we modularized its backbone network to enhance tomato ripeness detection performance [4]. Tomato images extract features via improved CSP backbone, retain YOLOv11's main structure, replace some C3k2l with new modules (KSFA, C3k2l, LSConv) to adapt different scale processing. Multi-scale features fuse through FPN+PAN neck, new modules implanted by scale—shallow KSFA enhance small targets, mid-layer C3k2l stabilize fusion, high-layer LSConv refine representation [5]. Fused features feed into YOLOv11 detection head to complete ripeness classification and localization. The architecture follow the principles of reusing optimized structure, enhancing capabilities at key layers, and controlling overhead, new modules' details will be elaborated in next section.

2.2. KSFA module (small-scale branch)

Small-scale feature maps are mainly used to detect very small tomato targets in images, but they are also more susceptible to background clutter interference. For this reason, the Kernel-Selective Feature Attention (KSFA) module is introduced to improve the sensitivity of the small-scale branch to targets and the suppression of background noise.

$$\begin{aligned} \mathrm{U}_1 &= \mathrm{P}_{1*1}\left(\mathrm{DwConv}_{3*3}\left(\mathrm{X}\right)\right) \\ \mathrm{U}_2 &= \mathrm{P}_{1*1}\left(\mathrm{DwConv}_{5*5}^{(\mathrm{d}=2)}\left(\mathrm{DwConv}_{3*3}\left(\mathrm{X}\right)\right)\right) \end{aligned} \tag{1}$$

$$\begin{aligned} \widetilde{\mathbf{S}} &= \mathrm{Conv}_{1*1}\left[\mathrm{Avg}_{\mathrm{ch}}\left(\mathrm{U}\right);\mathrm{Max}_{\mathrm{ch}}\left(\mathrm{U}\right)\right] \\ \mathrm{S}_{\mathrm{i}} &= \sigma\left(\widetilde{\mathrm{S}}_{\mathrm{i}}\right) \end{aligned} \tag{2}$$

$$\mathbf{W_i} = \mathbf{Broadcast}\left(\mathbf{c_i}\right) \odot \mathbf{S_i} \quad (\mathbf{i}=1,2) \tag{3}$$

$$\mathbf{S} = \mathbf{Conv}_{1*1}\left(\sum_{\mathrm{i}=1}^{2} \mathbf{W_i} \odot \mathbf{U_i}\right) \tag{4}$$

$$\mathrm{X}' = \mathrm{X} \odot \sigma\left(\mathrm{S}\right) \tag{5}$$

2.3. C3k2l module (mesoscale branch)

Mid-scale features target medium tomatoes, the network's key link. Derived from YOLO's C3 with 2×2 residual kernels and dual-bottleneck stacking, C3k2l boosts fusion stability. Compared to original C3's 1×1-3×3-1×1, its 2×2 kernels cut parameters and sharpen edges; dual-bottlenecks enrich diversity. Retaining CSP basics (split-input concatenation), it splits gradients to reduce redundancy, captures fine textures, and achieves full feature extraction with minimal parameter growth.

DOI: 10.54254/2753-8818/2026.31762

2.4. LSConv module (large-scale branch)

Large-scale features target close-up/focused tomatoes, requiring attention to shape and edge details. The introduced LSConv aim to "see the big picture and analyze details", combining large kernels' global perception and small kernels' local refinement, with two sub-components: LKP and SKA.

Input features first pass LKP, which use large receptive field convolutions (e.g., 7*7 via dilated convolution/multi-layer stacking) to extract target contour. Large kernels integrate wider context, aiding global shape grasp, critical for ripe contour identification and avoiding adjacent adhesion. LKP's output then goes to SKA (drawing on Selective Kernel Attention), a dynamic small convolution aggregation module with the following formula:

$$\mathrm{F} = \phi\left(\mathrm{Conv}_{1\mathrm{ks}}\left(\phi\left(\mathrm{Conv}_{1*1}\left(\mathrm{X}\right)\right)\right)\right) \tag{6}$$

$$\mathrm{W} = \mathrm{reshape}\left(\mathrm{GN}\left(\mathrm{Conv}_{1*1}\left(\mathrm{F}\right)\right)\right) \in \mathrm{R}^{\mathrm{B}*\frac{\mathrm{C}}{\mathrm{G}}*\left(\mathrm{sks}^2\right)*\mathrm{H}*\mathrm{W}} \tag{7}$$

$$\mathrm{Y}_{\mathrm{b,c,h,w}} = \sum_{\mathrm{i}=1}^{\mathrm{sks}} \sum_{\mathrm{j}=1}^{\mathrm{sks}} \mathrm{X}_{\mathrm{b,c,h+i,w+j}} \bullet \mathrm{W}_{\mathrm{b,c,(i,j),h,w}} \tag{8}$$

$$\mathrm{X}' = \mathrm{BN}\left(\mathrm{Y}\right) + \mathrm{X} \tag{9}$$

2.5. Multi-scale module allocation strategy

KSFA, C3k2l and LSConv are assigned to small, medium and large scales respectively, a differentiated strategy matching tomato features and module capabilities; small-scale KSFA boosts SNR for tiny targets, large-scale LSConv balances big tomatoes' outline and details, medium-scale C3k2l stabilizes feature fusion, achieving "targeted solution"; ablation experiments show cross-scale misuse harm performance, e.g., LSConv on small scales increases cost and loses details, KSFA replacing C3k2l in medium scales causes insufficient fusion.

3. Experimental design

3.1. Data set sources

This study adopt open-source Laboro Tomato dataset, supporting tomato detection and segmentation across ripening stages. Collected via two cameras with different parameters under multi-angle/illumination conditions, it covers diverse growth environments and appearances; each tomato instance is annotated with bounding box, segmentation mask, ripeness (fully/half/unripe) and size (normal/cherry) tags. The dataset have diversity in ripeness, size combinations and collection conditions, with high research and application value in detection, segmentation and ripeness recognition.

3.2. Evaluation indicators

$$\mathbf{mAP}_{50} = \frac{1}{\mathbf{N}} \sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{AP_i}\left(\mathbf{IoU} = 0.5\right) \tag{10}$$

$$\mathbf{mAP}_{50:95} = \frac{1}{\mathbf{10N}} \sum_{\mathbf{t}=0.50}^{0.95} \sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{AP_i}\left(\mathbf{IoU} = \mathbf{t}\right),\ \mathbf{t} \in \left\{0.50, 0.55, \ldots, 0.95\right\} \tag{11}$$

DOI: 10.54254/2753-8818/2026.31762

$$\text{Pr ecision} = \frac{\text{TP}}{\text{TP+FP}}, \ \text{Re call} = \frac{\text{TP}}{\text{TP+FN}}$$
$$\text{F1} = \frac{2^{*}\text{Precision}^{*}\text{Re call}}{\text{Precision+Re call}} \tag{12}$$

4. Experimental results and analysis

4.1. Overall performance comparison

Table 1. Comparison of detection performance for different stage module combinations on YOLOv11n (n-scale)

Combination Name/Number	(P1/2)	(P3)	(P4)	(P5)	(F1 Score)	mPA50	mPA50:95
KSFA-LS | 101	KSFA	LSConv	LSConv	LSConv	0.68@0.434	0.72	0.54
KSFA-LS | 102	KSFA	KSFA	LSConv	LSConv	0.68@0.362	0.73	0.54
KSFA-LS | 103	KSFA	KSFA	KSFA	LSConv	0.69@0.361	0.73	0.56
KSFA-LS | 104	KSFA	KSFA	KSFA	KSFA	0.70@0.387	0.74	0.56
LSConv | 201	LSConv	LSConv	LSConv	LSConv	0.65@0.303	0.68	0.51
LSConv | 202	C3k2l	LSConv	LSConv	LSConv	0.68@0.343	0.70	0.53
LSConv | 203	C3k2l	C3k2l	LSConv	LSConv	0.68@0.361	0.71	0.55
LSConv | 204	C3k2l	C3k2l	C3k2l	LSConv	0.70@0.304	0.70	0.56
KSFA | 301	KSFA	C3k2l	C3k2l	C3k2l	0.68@0.364	0.72	0.54
KSFA | 302	KSFA	KSFA	C3k2l	C3k2l	0.68@0.365	0.72	0.55
KSFA | 303	KSFA	KSFA	KSFA	C3k2l	0.68@0.393	0.72	0.54
KSFA | 304	KSFA	KSFA	KSFA	KSFA	0.67@0.358	0.71	0.55
KSFA++LS | 401	KSFA	C3k2l	LSConv	LSConv	0.72@0.424	0.75	0.58
KSFA++LS | 402	KSFA	C3k2l	C3k2l	LSConv	0.68@0.385	0.73	0.56
KSFA++LS | 403	KSFA	KSFA	C3k2l	LSConv	0.69@0.339	0.73	0.55
YOLOv11(BS)	C3k2l	C3k2l	C3k2l	C3k2l	0.64@0.301	0.67	0.47

Table 2. Comparison of F1(max) and computational cost for different model combinations across five scales (n/s/m/l/x)

Combination Name/Number	n	s	m	l	x
LSConv | 204	0.70@6.2	0.73@21.8	0.76@41.8	0.76@85.4	0.79@133.6
KSFA++LS | 401	0.75@5.3	0.76@20.3	0.77@41.3	0.79@87.7	0.80@136.6
KSFA | 304	0.70@5.0	0.73@16.9	0.76@30.6	0.77@60.7	0.78@92.3
YOLOv11(BS)	0.67@6.4	0.72@21.6	0.75@62.4	0.74@89.6	- @195.5
average computing power requirement	5.7	21.23	44.025	80.85	139.575

DOI: 10.54254/2753-8818/2026.31762

4.2. Complex scene performance

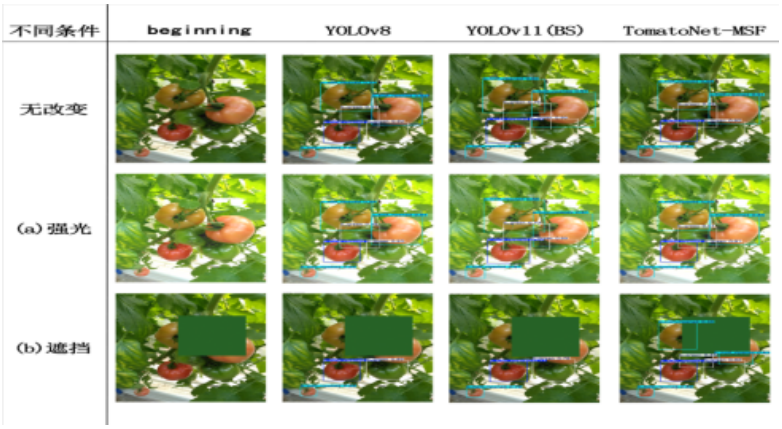

Figure 1. Visual comparison of detection under complex conditions

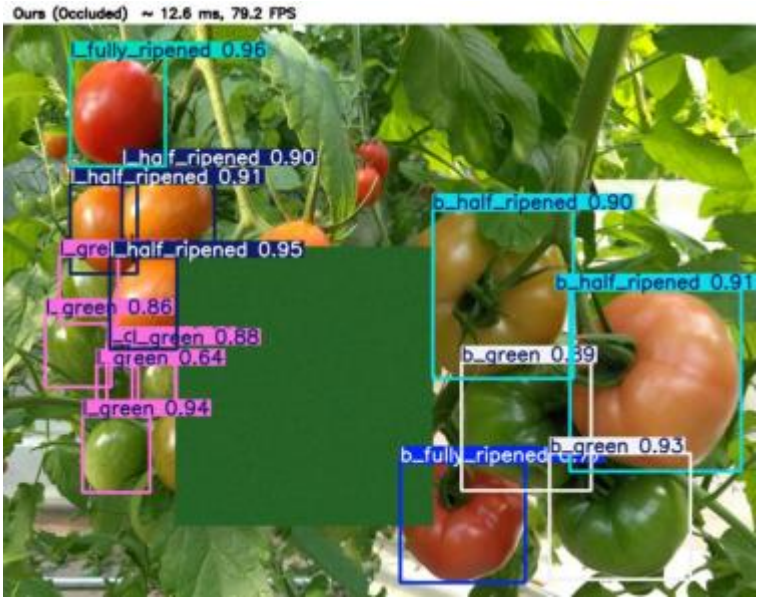

Figure 2. Tomato Net-MSF

Figure 2 demonstrates that the Tomato Net-MSF model accurately identifies different tomato varieties and their corresponding maturity levels.

4.3. Feature visualization

To analyze the impact of each module on feature extraction, researchers visualized feature maps at different scales using heat maps. The final Tomato Net-MSF model validated the effectiveness of the multi-scale differentiated module allocation strategy.

DOI: 10.54254/2753-8818/2026.31762

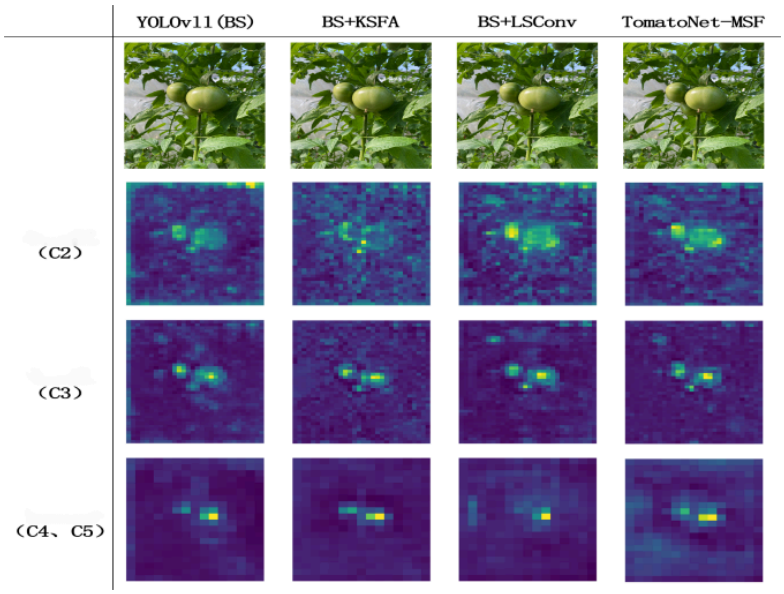

Figure 3. Comparison of response thermograms of different models in the multi-scale feature layer

4.4. Deployment of domestic edge computing platform

To verify the model's deployment performance on domestic SoC platforms, this study convert trained pt model to rknn via Rockchip's toolchain, develop an optimized multi-thread framework (supports custom pre/post-processing and visualization) and open-source it; deployment tests show excellent performance under limited computing power and overlapping fruits. To optimize RK3588 NPU quantization deployment, this study find YOLOv11n's SiLU has abnormal confidence output (limited to 0.5) in INT8 quantization due to asymmetry and scale adaptation issues, replacing SiLU with ReLU solves distortion with <0.8% accuracy loss, expanding confidence range and accelerating convergence, providing reliable basis for agricultural robots.

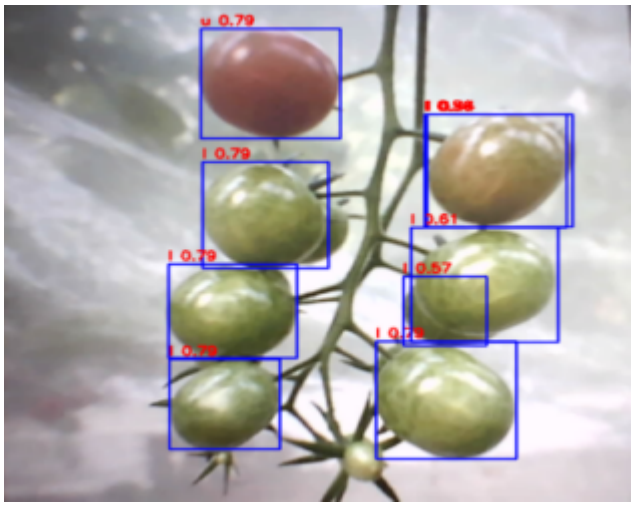

Figure 4. Demonstration of real-time edge detection with RK3588

The proposed Tomato Net-MSF (KSFA+C3k2l+LS architecture) in this study achieve the best full-scale performance, its F1 Score outperform the baseline YOLOv11n under all model scales while GFLOPs reduce significantly; for example, under the n model, the F1 Score reach 0.72@5.3,

DOI: 10.54254/2753-8818/2026.31762

which is more than 6% higher than YOLOv11n's 0.68@6.4, and the computational cost decrease by about 17%, it also show significant scalability advantages at large scales—under the x model, the baseline YOLOv11n cannot run in the local environment due to excessive computing power requirements (195.5 GFLOPs), while Tomato Net-MSF still maintain stable performance of 0.80@136.6; the improvement of a single module is limited, KSFA or LSConv perform close to Tomato Net-MSF at some scales but lack overall stability and robustness, making it difficult to balance multi-scale target detection, in addition, Tomato Net-MSF balance lightweight and accuracy, under small scales (n, s), it not only improve accuracy but also reduce computational cost, proving its strong application potential in edge computing and resource-constrained scenarios.

5. Conclusion

This study proposes Tomato Net-MSF (improved for tomato ripeness detection), integrate differentiated multi-scale modules into YOLOv11. Small-scale KSFA enhance small-target detection; mid-scale C3k2l stabilize feature fusion, large-scale LSConv balance global perception and detail. Differentiated allocation optimize accuracy-efficiency balance.Experiments show Tomato Net-MSF outperform YOLOv11n (baseline) in mAP/F1 on Laboro Tomato dataset, robust to strong light/occlusion. It achieve 50 FPS (≈20 ms/frame) on RK3588 NPU (~5x faster), with high practical deployment potential.Limitations remain: limited data diversity, poor ultra-low power platform adaptation, insufficient greenhouse verification. Future research include dataset expansion, model lightweight, multi-task fusion and cross-modal perception to advance fruit detection in intelligent harvesting.

Funding information

This work was supported by the Jiangsu Province College Students' Innovation and Entrepreneurship Training Program under Grant No. 202411276017Z. The corresponding project title is "Design and Control Method Research of Tomato Picking Robot Based on Machine Vision". Additionally, we would like to thank Professor Chen Wei for his careful guidance during the research and Chen Wei Laboratory for providing technical support.

References

[1] KHAN A, HASSAN T, SHAFAY M, et al. Tomato maturity recognition with convolutional transformers [J]. Science Reports, 2023, 13: 22885.
[2] WU Q, HUANG H, SONG D, et al. YOLO-PGC: A tomato maturity detection algorithm based on improved YOLOv11 [J]. Applied Sciences, 2025, 15(9): 5000.
[3] GOUIDER C, SEDDIK H. YOLOv4 and branch attention: An improved approach to real-time object detection [C]//2022 IEEE Information Technologies & Smart Industrial Systems (ITSIS). Paris, France: IEEE, 2022: 1-6.
[4] WANG C Y, LIAO H Y M, WU Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE, 2020: 1571-1580.
[5] VIṆA A. Comparing Ultralytics YOLO11 vs previous YOLO models [EB]. Ultralytics, 2025-04-02.