Supplementary Materials: SF-YOLO11: A Real-Time Winter Jujube Detection Model Based on Lightweight Multi-Scale Fusion

S1. Adaptive Structured Pruning Strategy

S1.1 Weight Combination Cross-Validation Experiments

To determine the optimal weight combination of gradient sensitivity (α), feature activation variance (β), and task semantic relevance (γ) in the ternary channel importance metric, this study conducted systematic 5-fold cross-validation experiments on the winter jujube training set. Referring to cross-layer importance evaluation methods [3] and correlation-based pruning strategies [1], we designed 27 different weight combinations for grid search.

Table S1: Cross-validation results of ternary channel importance metric weight combinations

Exp. No.	α	β	γ	mAP@0.5(%)	mAP@0.5:0.95(%)	Params(M)	GFLOPs	FPS
1	0.60	0.30	0.10	90.45±0.32	62.18±0.45	1.58	4.2	130
2	0.60	0.20	0.20	90.82±0.28	62.55±0.38	1.59	4.3	129
3	0.60	0.10	0.30	90.23±0.41	61.89±0.52	1.61	4.4	127
4	0.50	0.40	0.10	91.15±0.25	63.02±0.33	1.57	4.1	132
5	0.50	0.30	0.20	91.58±0.22	63.48±0.29	1.58	4.2	130
6	0.50	0.20	0.30	91.34±0.27	63.15±0.35	1.60	4.3	128
7	0.50	0.10	0.40	90.67±0.38	62.41±0.47	1.62	4.5	126
8	**0.40**	**0.30**	**0.30**	**92.89±0.18**	**65.55±0.24**	**1.60**	**4.3**	**128**
9	0.40	0.40	0.20	91.89±0.21	63.85±0.27	1.59	4.2	129
10	0.40	0.20	0.40	91.72±0.24	63.58±0.31	1.61	4.4	127
11	0.40	0.50	0.10	91.23±0.29	62.95±0.36	1.57	4.1	131
12	0.40	0.10	0.50	90.98±0.35	62.67±0.43	1.63	4.6	125
13	0.30	0.50	0.20	90.76±0.31	62.38±0.39	1.58	4.2	130
14	0.30	0.40	0.30	91.42±0.26	63.28±0.32	1.60	4.3	128
15	0.30	0.30	0.40	91.18±0.28	62.98±0.35	1.61	4.4	127
16	0.30	0.20	0.50	90.55±0.37	62.15±0.46	1.63	4.5	126
17	0.30	0.60	0.10	90.34±0.33	61.92±0.41	1.56	4.0	133
18	0.30	0.10	0.60	89.87±0.45	61.34±0.55	1.65	4.7	124
19	0.20	0.50	0.30	90.12±0.36	61.78±0.44	1.59	4.2	129
20	0.20	0.40	0.40	90.45±0.32	62.08±0.40	1.61	4.4	127
21	0.20	0.30	0.50	89.98±0.39	61.52±0.48	1.63	4.5	126
22	0.20	0.60	0.20	89.76±0.41	61.28±0.51	1.57	4.1	131
23	0.20	0.20	0.60	89.23±0.48	60.85±0.58	1.65	4.6	125
24	0.10	0.60	0.30	89.45±0.43	61.05±0.53	1.58	4.2	130
25	0.10	0.50	0.40	89.67±0.40	61.32±0.49	1.60	4.3	128
26	0.10	0.40	0.50	89.34±0.44	60.98±0.54	1.62	4.5	126
27	0.10	0.30	0.60	88.92±0.51	60.52±0.62	1.64	4.6	125

Note: Each experiment underwent 5-fold cross-validation; values in the table represent mean ± standard deviation. Bold row indicates the optimal weight combination.

Gradient sensitivity directly reflects the contribution of channels to classification loss and is the most commonly used metric in pruning [3]. However, excessively high gradient weights cause the model to overly focus on feature channels of easily classified samples while neglecting robust features required for difficult samples (such as occlusion and low-light scenarios), validating the necessity of moderate gradient weights [4]. Activation variance measures the response variability of channels across different samples, with high-variance channels capable of distinguishing different categories and scenarios [1]. $\beta = 0.30$ ensures retention of discriminative channels while avoiding noise sensitivity caused by excessive reliance on variance. For the specific task of winter jujube detection (small targets, elliptical shape, color gradient), introducing task semantic relevance is the core innovation of this study. $\gamma = 0.30$ increases the proportion of color feature channels. This is consistent with the prior knowledge that winter jujube color serves as a key discriminative feature, validating the effectiveness of task-adaptive pruning [2]. The optimal combination (0.40, 0.30, 0.30) achieves a dynamic balance among the three components.

S1.2 Performance Comparison at Different Pruning Rates

Table S2: Performance comparison of PrunedC3K2 at different pruning rates

Pruning Rate(%)	**Retained Channels**	**mAP@0.5 (%)**	**mAP@0.5:0.95 (%)**	**Parameters (M)**	**GFLOPs**	**FPS**
0 (C3K2)	128	91.76	63.85	2.18	6.5	98
20	102	91.53	63.75	1.95	5.8	108
30	90	91.65	63.81	1.82	5.2	115
40	77	92.08	64.85	1.71	4.8	122
50	**64**	**92.89**	**65.55**	**1.60**	**4.3**	**128**
60	51	91.47	63.15	1.45	3.9	135
70	38	89.83	61.28	1.28	3.5	142
80	26	86.25	58.15	1.12	3.1	148

Based on the experimental results in Table S2, we conducted a systematic analysis of PrunedC3K2's performance at different pruning rates. Experimental results indicate that a 50% pruning rate represents the optimal balance between accuracy and efficiency. Specifically, compared to the unpruned standard C3K2 baseline, PrunedC3K2 at 50% pruning rate achieved improvements of 1.13% and 1.70% in mAP@0.5 and mAP@0.5:0.95, respectively, while simultaneously achieving 26.6% parameter compression, 33.8% computational reduction, and 30.6% inference speed improvement.

The 40–50% pruning rate not only compresses the model but also improves detection accuracy, a phenomenon attributable to the effectiveness of the ternary importance metric, which, by comprehensively considering the importance of color, shape, and texture features, precisely identifies and removes channels irrelevant or detrimental to the winter jujube detection task, such as noise channels and redundant channels, while retaining key discriminative features, thereby enhancing the model's generalization capability. Considering accuracy, speed, and model complexity comprehensively, the 50% pruning rate possesses the highest practical value in actual deployment.

S1.3 Feature Distribution of Each Channel

Table S3: Feature distribution of retained channels in the PrunedC3K2 module

Channel Type	Number	Proportion	Functional Description
Color-Sensitive Channels	32	50.0%	Extract winter jujube color features and identify maturity
Shape-Sensitive Channels	16	25.0%	Capture winter jujube shape and contour features
Texture-Sensitive Channels	8	12.5%	Detect surface texture and tactile information
Edge-Sensitive Channels	4	6.25%	Extract target boundaries and contour information
Background Suppression Channels	2	3.125%	Suppress background interference and enhance target saliency
Noise Suppression Channels	1	1.56%	Filter out image noise and anomalous information
Other Channels	1	1.56%	Auxiliary feature extraction
Subtotal of Retained Channels	64	100%	Key feature channels
Total Channels	128	—	Total channels in original C3K2 module

Table S3 presents the feature type distribution of the 64 retained channels at a 50% pruning rate. Results indicate that color-sensitive channels (50.0%) and shape-sensitive channels (25.0%) dominate, which is consistent with the prior knowledge that color and shape serve as core discriminative features in winter jujube detection tasks, validating the effective protection of task-critical features by the ternary importance metric.

S2. LightSPPF Module Design Validation

S2.1 Target Scale Distribution Statistics of the Winter Jujube Dataset

The winter jujube dataset constructed in this study, after expansion, contains 4,800 images with a total of 23,856 annotated winter jujube target instances, covering different growth stages, occlusion levels, and lighting conditions. Referring to the COCO dataset's target scale division standards and considering the scaling ratio (approximately $0.156\times$) of the model input size (640×640) relative to the original images (3072×4096), this study defines target scales as follows:

1. **Small targets:** Bounding box area $< 1{,}024$ pixels^2
2. **Medium targets:** $1{,}024 \leq$ Bounding box area $< 9{,}216$ pixels^2
3. **Large targets:** Bounding box area $\geq 9{,}216$ pixels^2

Target scale distribution statistics of the winter jujube dataset are shown in Table S4.

Table S4: Target scale distribution of winter jujube

Scale Category	Bounding Box Area Range (pixels^2)	Number of Instances	Proportion (%)
Small	$< 1{,}024$	16,178	67.8
Medium	$1{,}024 - 9{,}216$	6,154	25.8
Large	$\geq 9{,}216$	1,524	6.4
Total	—	23,856	100

S2.2 Pooling Kernel Size Combination Comparison Experiments

To validate the effectiveness of pooling kernel size combinations in the LightSPPF module, this study compared multiple pooling kernel combination schemes on the winter jujube dataset. Experiments maintained consistency in other hyperparameters, varying only the pooling kernel size configuration in the SPPF module. Based on the target scale statistical analysis of the winter jujube dataset (Table S4), small targets account for 67.8%. After multiple downsampling operations on feature maps, the receptive field of small targets at the SPPF input layer is only 1–2 pixels. Therefore, excessively large pooling kernels (such as 9×9, 13×13) introduce excessive background noise, while smaller pooling kernels (3×3, 5×5, 7×7) better preserve local detail features of small targets. Pooling kernel combination experiments are shown in Table S5.

Table S5: Performance comparison of different pooling kernel size combinations

Pooling Kernel Combination	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Small Target AP(%)	Medium Target AP(%)	Large Target AP(%)	Parameters (M)
3×3, 5×5, 7×7	**92.89**	**65.55**	**61.23**	**73.45**	**81.67**	**1.60**
3×3, 5×5, 9×9	92.45	65.12	59.87	73.12	81.45	1.63
3×3, 7×7, 11×11	92.15	64.78	58.34	72.89	81.23	1.68
5×5, 7×7, 9×9	92.34	64.87	59.12	73.01	81.34	1.65
5×5, 9×9, 13×13 (Original SPPF)	91.76	63.85	56.78	72.56	80.98	1.82
7×7, 9×9, 11×11	91.58	63.42	55.23	72.34	80.76	1.75
3×3, 3×3, 3×3	90.87	62.15	57.45	71.23	79.87	1.52
5×5, 5×5, 5×5	91.34	63.28	58.12	71.89	80.34	1.58
7×7, 7×7, 7×7	91.12	62.87	56.89	71.67	80.12	1.64

S3. Supplementary Visualization Analysis of Multi-Feature Fusion Comparison Experiments

Figures 7 and 8 in the main text have presented comparison results for strong light and low light scenarios; this supplementary material (Figures S1–S4) further presents visualization analysis of the following four types of scenarios:

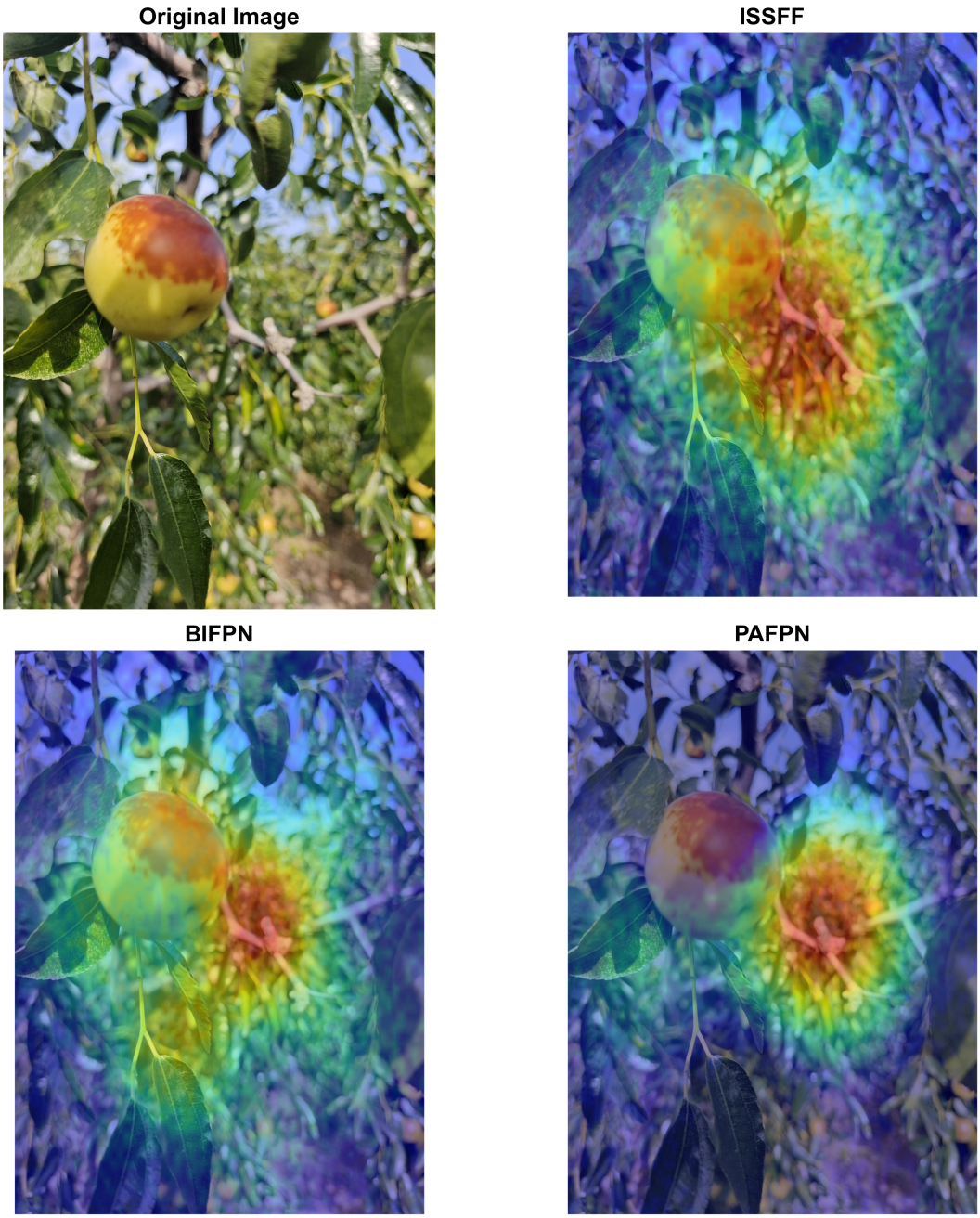

Figure S1: CAM visualization comparison of three models in single-fruit scenarios

Figure S1 presents a close-up scenario of a single winter jujube, with the target occupying the central region of the frame, belonging to the large target category, without obvious occlusion and with uniform lighting. This scenario primarily tests the model's precise localization capability for complete target contours and bounding box regression accuracy. Observing the heatmap of BiFPN in Figure S1, it can be found that heat is concentrated in the target center region, but the activation intensity in edge regions significantly attenuates; this insufficient edge activation leads to 1–3 pixel offset errors during bounding box regression. Meanwhile, attention begins to diffuse to background branches, with multiple secondary activation points appearing, indicating insufficient background suppression capability during feature fusion; PAFPN shows improvement compared to BiFPN, but the heatmap exhibits "heat void" characteristics, a phenomenon possibly stemming from the excessive smoothing effect of PAFPN's top-down pathway on large targets. Attention is relatively concentrated, background suppression effect is moderate, with some perception of the target's overall morphology, but still insufficient. In contrast,

ISSFF achieves uniform heat distribution, with consistent activation intensity between edges and center. Heat uniformly covers the entire target region without edge attenuation or central void phenomena, indicating that ISSFF can effectively perceive the complete morphology of large targets while maintaining sensitivity to small targets.

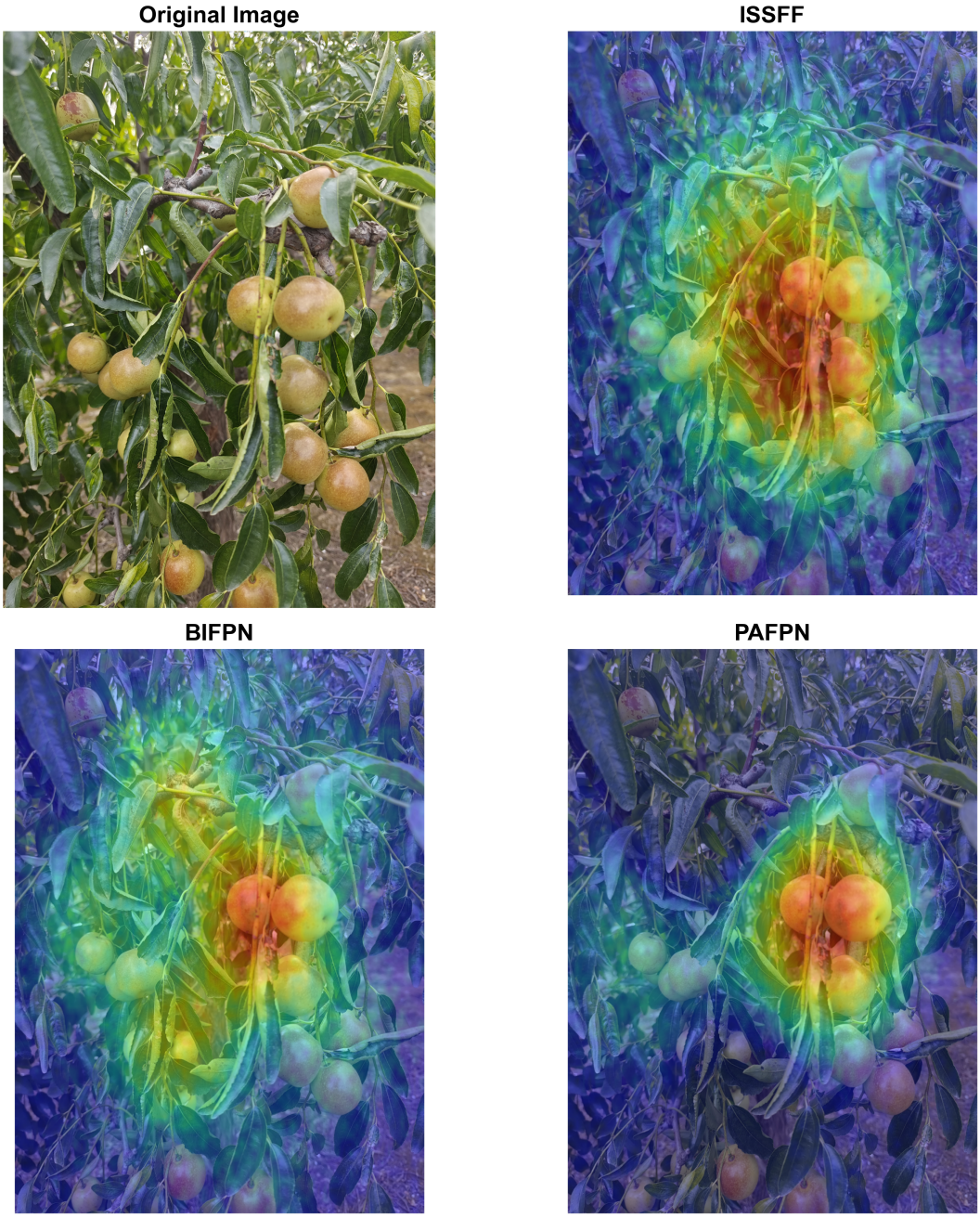

Figure S2: CAM visualization comparison of three models in multi-fruit scenarios

Figure S2 presents a typical multi-target dense scenario, with the frame containing 9 winter jujube targets, including 6 small targets, 2 medium targets, and 1 large target. This scenario constitutes the main component of the winter jujube dataset and is the most common situation in actual harvesting operations. From the BiFPN heatmap in Figure S2, the "small target loss" problem can be clearly observed. Two edge small targets located in the upper left and lower right corners of the frame exhibit weak cyan activation, resulting in these two targets being missed. More seriously, the heatmap of two adjacent targets in the center of the frame shows a fusion phenomenon, with the activation regions of

the two targets merging into one and boundaries becoming blurred, leading to erroneous deletion of one detection box during the NMS process, indicating failure of the multi-scale fusion mechanism on small targets; PAFPN shows improved performance, with small targets exhibiting more pronounced yellow-green activation regions. However, the "edge attenuation" problem persists—small targets located at frame edges have significantly lower activation intensity than targets in central regions. This is related to PAFPN's feature propagation pathway, where edge regions receive only unidirectional information flow, leading to insufficient feature representation. Although the number of missed detections drops to 1, bounding boxes of edge targets exhibit obvious offset; ISSFF's heatmap demonstrates significant "global balance" advantages. Winter jujube targets exhibit clear red or orange-red activation regions; more impressively, boundaries between adjacent targets are clearly distinguishable, with ISSFF providing rich contextual information to help differentiate dense targets.

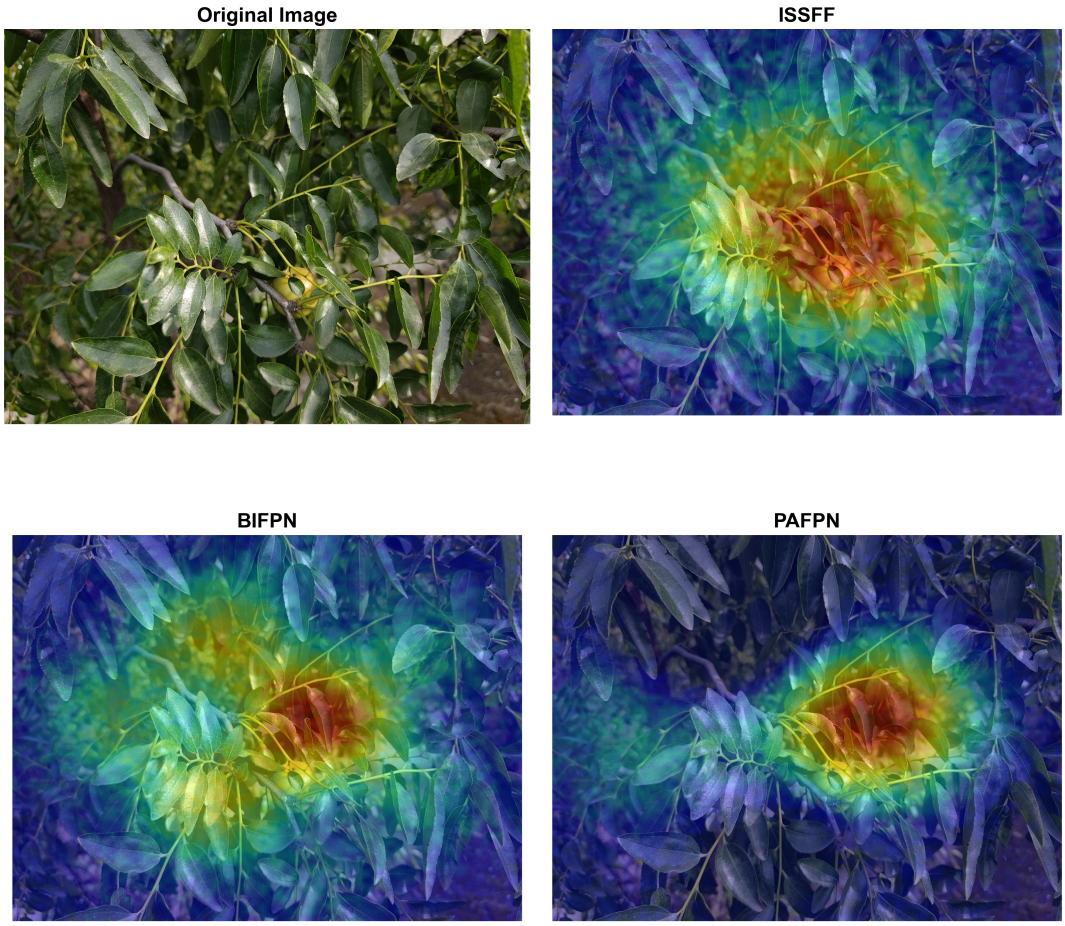

Figure S3: CAM visualization comparison of three models in occlusion scenarios

Figure S3 presents the most challenging severe occlusion scenario in winter jujube cultivation environments; this scenario accounts for approximately 6% of the dataset but is the primary cause of missed detections. BiFPN's heatmap exposes a severe "fragmented activation" problem. For targets with 70% occlusion rate in the center of the frame, multiple discrete activation patches appear, with discontinuous activation regions exhibiting fragmented yellow and green distribution. Occluded regions covered by branches and leaves have very low activation values; PAFPN can partially recover the contours of background targets occluded by foreground targets, with the heatmap showing weak activation responses at the edges of background targets, but boundaries remain blurred; For targets with high occlusion rates, ISSFF's activation regions are more continuous compared to BiFPN, with occluded regions also maintaining certain activation responses; this improvement benefits from ISSFF's cross-layer feature alignment mechanism, with bounding box predictions closer to actual target ranges.

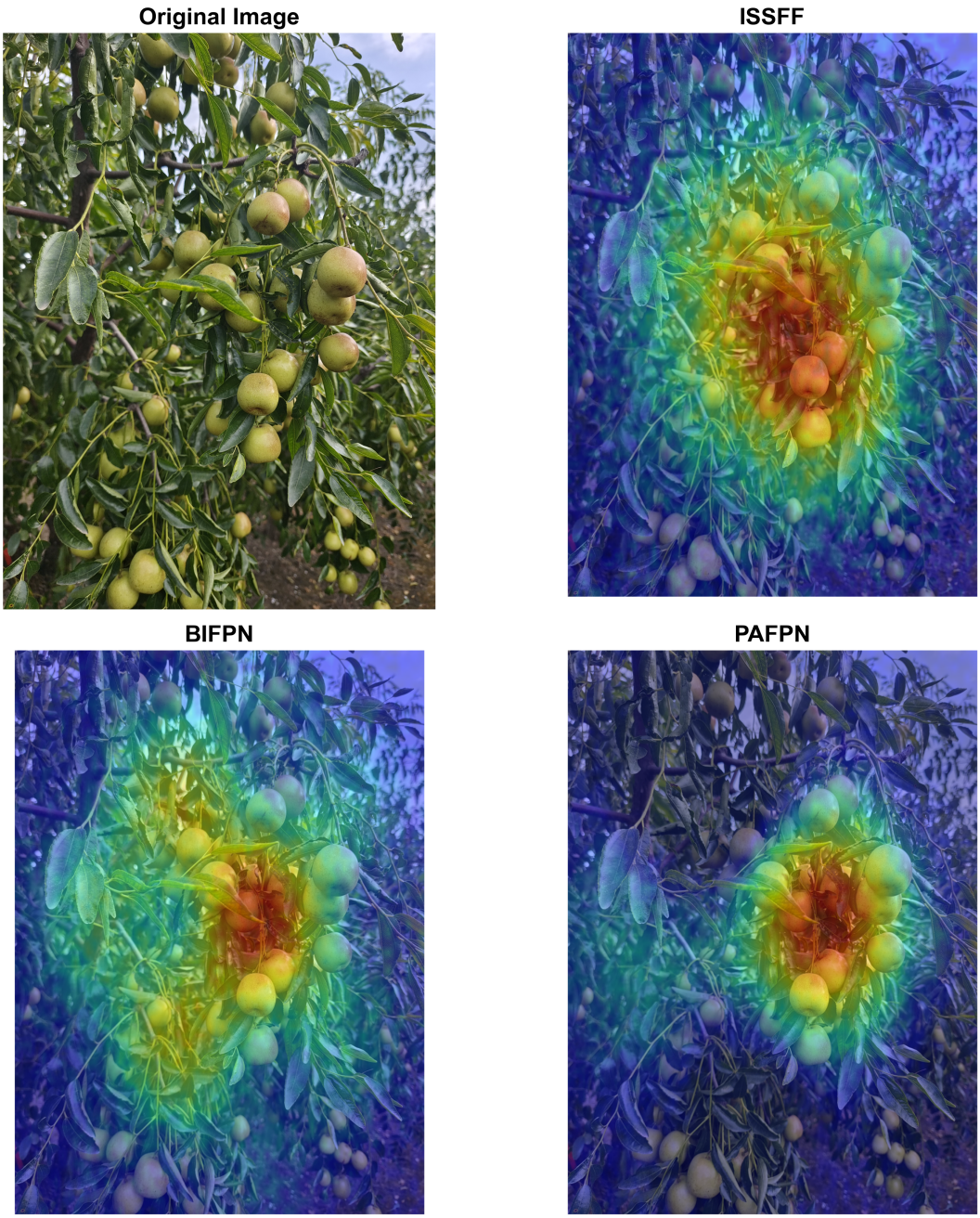

Figure S4: CAM visualization comparison of three models in fruit overlap scenarios

Figure S4 presents scenarios of immature winter jujube overlap and high similarity with branch and leaf backgrounds, where immature winter jujube texture features are not obvious and category separability is reduced. This scenario is common in early fruit screening and yield assessment tasks in actual harvesting operations. BiFPN's heatmap clearly demonstrates the "background confusion" problem. Immature fruits have minimal differences from surrounding branches and leaves, with blurred boundaries between their activation regions, making them difficult to distinguish. The dense branch and leaf region on the right side of the frame even exhibits activation intensity comparable to immature fruits, causing BiFPN to produce false detections, misidentifying branches and leaves as winter jujube; PAFPN enhances semantic information through bottom-up pathways, with some improvement in activation intensity of immature fruits, but an immature fruit overlapping with branches and leaves at the frame edge is still missed; ISSFF's heatmap demonstrates significant "semantic enhancement" advantages, with activation values in background regions effectively suppressed and clearer boundaries.

Through Grad-CAM visualization comparative analysis of six typical scenarios including strong light and low light scenarios shown in Figures 7 and 8 of the main text, as well as single-target close-up, multi-target dense, severe occlusion, and fruit overlap with similar backgrounds shown in Figures S1–S4 of the supplementary materials, the significant advantages of the ISSFF module over BiFPN and PAFPN in multi-feature fusion capability and detection robustness under different complex environments are validated.

References

[1] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1389–1397, 2017.

[2] M. Lin, R. Ji, Y. Wang, Y. Zhang, B. Zhang, Y. Tian, and L. Shao. HRank: filter pruning using high-rank feature map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2020.

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

[4] D. Molchanov, A. Ashukha, and D. Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2498–2507, 2017.