Methods

S1 Style transfer of the dataset enhancement

In color-sensitive tasks such as ripeness detection, conventional data augmentation techniques including geometric transformations (e.g., rotation, flipping and cropping) and photometric perturbations (e.g., noise, blur, color jitter, and random grayscaling) may adversely affect model performance. Overly aggressive color manipulations can diminish the sensitivity of the model to discriminative hue variations (Zini et al., 2022). Thus, to mitigate the scarcity of low-light image data and improve generalization under varying illumination conditions, this study introduces the SaMam style transfer technique (Liu et al, 2025) to tomato ripeness recognition. Unlike conventional style transfer methods such as CycleGAN (Zhu et al., 2017), SaMam was selected for its stronger ability to preserve semantic structures while processing high-resolution agricultural images. GAN-based methods often introduce geometric distortions or texture artifacts that can obscure phenotypic features critical to ripeness assessment, such as skin texture and color gradation. In contrast, SaMam integrates State Space Models (Mamba) for efficient long-range dependency modeling (Gu and Dao, 2023), enabling the generation of globally consistent illumination effects without compromising the local morphological integrity of tomato fruits. This ensures that the augmented data remain physically realistic and suitable for training.

Using multiscene tomato images as the style references, we performed style transfer via SaMam to augment 600 images from the training set with three distinct lighting conditions. This process raised the proportion of complex lighting examples in the training data to 25%. The workflow of the SaMam-based style transfer and representative results are illustrated in Figs. S1 and S2, respectively.

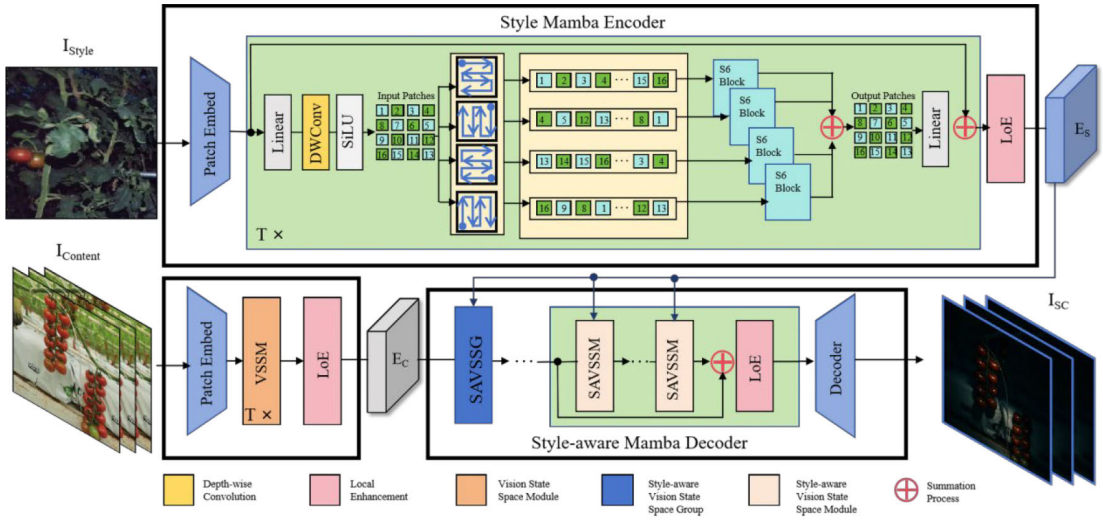

Fig. S1 Architecture and workflow of the SaMam style transfer model, employed for dataset augmentation under complex lighting conditions.

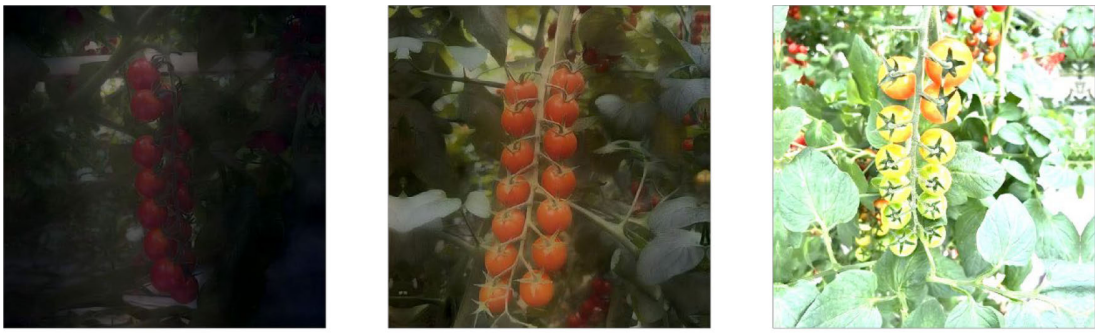

Fig. S2 Sample images generated by the SaMam model, simulating three challenging lighting scenarios.

S2 YOLOv12 algorithm

YOLOv12 represents a leading real-time object detection that significantly improves the balance between detection accuracy and processing speed (Tian et al., 2025). Building upon the hierarchical structure inherent to the YOLO series, it incorporates advanced attention mechanisms to enhance feature representation. As illustrated in Fig. S3, the model consists of three main components: a backbone, a feature fusion neck, and a detection head. The backbone retains the convolutional structure of YOLOv11 (Khanam and Hussain, 2024) in its initial stages, and its core introduces an area attention mechanism, which partitions the feature map into four equal regions along either the height or the width dimension. Each region undergoes independent attention module (A_2) computation before being fused, reducing the computational complexity from $O(n^2)$ to $O(n^2/4)$ while preserving a significant portion of the global receptive field.

To enhance feature fusion efficiency and mitigate training instability in large-scale models, YOLOv12 incorporates a residual efficient layer aggregation network (R-ELAN) with a scaling factor of 0.01 and a single-path bottleneck structure. This design reduces the number of parameters by 10% without compromising performance. In the feature fusion neck, a single-layer R-ELAN enhances the FPN structure by streamlining the structure and compressing the number of cross-scale fusion channels by 30%, thereby maintaining efficient multiscale detection capabilities. The detection head further optimizes efficiency by replacing standard convolutions with depth-wise separable convolutions and decoupling the classification and regression branches. Compared to contemporary models such as YOLOv11 and DETR (Carion et al., 2020), YOLOv12 achieves superior performance in both accuracy and detection speed within the domain of object detection.

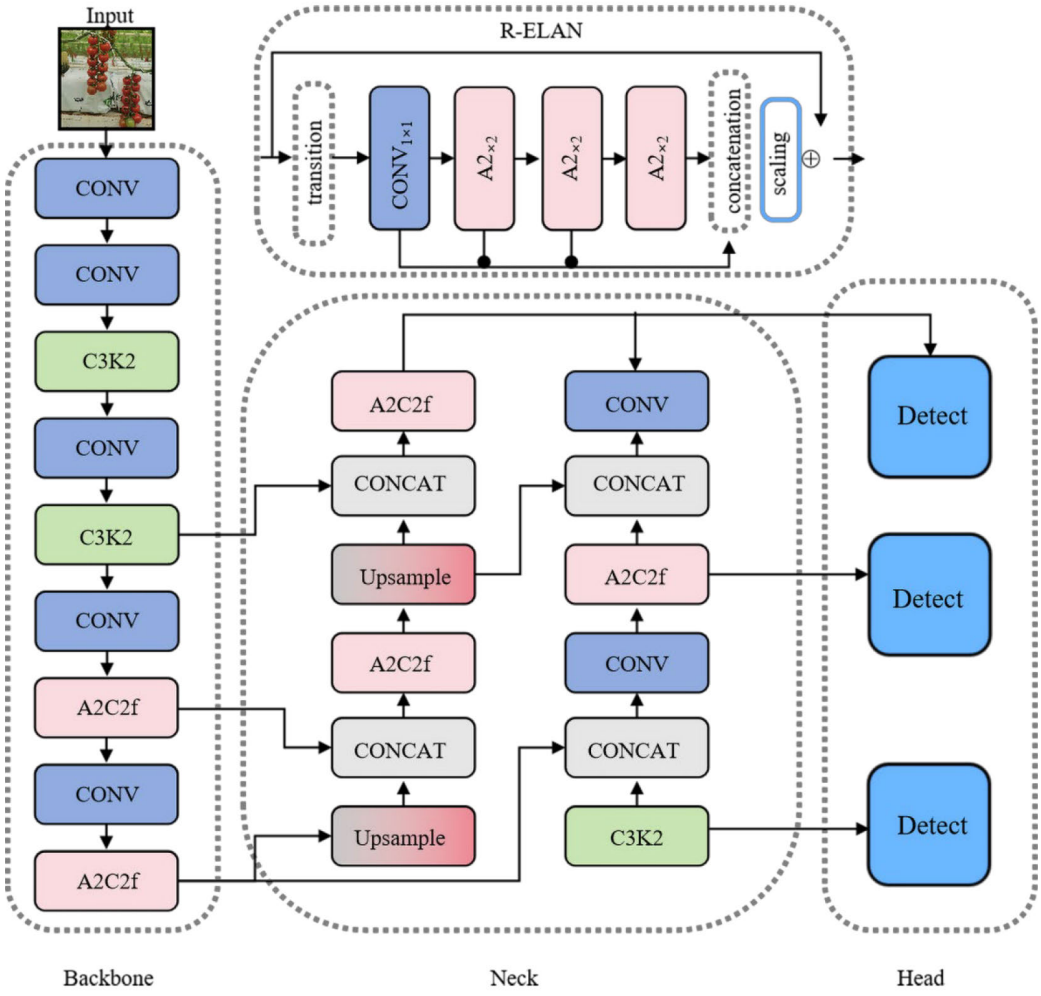

Fig. S3 Network architecture of the YOLOv12 model illustrating its three core components: backbone, neck and head.

S3 Training configuration details

Training was performed on a server equipped with dual NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM each), providing a total of 48 GB VRAM, and an AMD EPYC 7402 processor with 22 virtual cores under the Ubuntu 22.04 operating system. The development environment included PyTorch 2.2.0, CUDA 12.6 and Python 3.11.4. Model testing was carried out on a high-performance gaming computer with an NVIDIA RTX 4060 GPU (8 GB) and an Intel Core i7-14650HX CPU, while running Windows 11.

We trained the model on a custom tomato dataset with an input image resolution of 640×640 pixels, a batch size of 64, and 300 epochs. The optimization used a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. As the dataset had been previously augmented for illumination variance using the SaMam technique, no further color-space augmentations were employed. During training, spatial augmentations including random perspective (offset ±0.0005), horizontal flips (50% probability), and rotations (up to 15°) were applied. In addition, a patience of 100 was set to prevent overfitting. This configuration aimed to enhance model generalization and training efficiency.

References

Carion N, Massa F, Synnaeve G, et al., 2020. End-to-end object detection with transformers. arXiv:2005.12872.
https://doi.org/10.48550/arXiv.2005.12872

Gu A, Dao T, 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv:2312.00752. https://doi.org/10.48550/arXiv.2312.00752

Khanam R, Hussain M, 2024. YOLOv11: An overview of the key architectural enhancements. arXiv:2410.17725. https://doi.org/10.48550/arXiv.2410.17725

Liu H, Wang L, Zhang Y, et al., 2025. SaMam: style-aware state space model for arbitrary image style transfer. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA.
https://doi.org/10.48550/arXiv.2503.15934

Tian Y, Ye Q, Doremann D, 2025. YOLOv12: attention-centric real-time object detectors. arXiv:2502.12524. https://doi.org/10.48550/arXiv.2502.12524

Zhu J Y, Park T, Isola P, et al., 2017. Unpaired image-to-image translation using Cycle-Consistent adversarial networks. arXiv:1703.10593.
https://doi.org/10.48550/arXiv.1703.10593

Zini S, Gomez-Villa A, Buzzelli M, et al., 2022. Planckian Jitter: countering the color-crippling effects of color jitter on self-supervised training. Paper presented at the 2023 International Conference on Learning Representations (ICLR). Kigali, Rwanda.
https://doi.org/10.48550/arXiv.2202.07993