Application of SAM-2 Vision Foundation Model in Apple Harvesting Robotics

Sunny Lu

Received August 31, 2024
Accepted September 30, 2024
Electronic access October 15, 2024

This study explores the innovative application of the Segment Anything Model 2 (SAM-2) visual foundation model in apple harvesting robots. We propose a method that combines SAM-2 with YOLOv8, utilizing the YOLOv8 object detection algorithm to identify anchor coordinates of apples and input this positional information into the SAM-2 model, thereby achieving high-precision apple recognition and segmentation. SAM-2's advanced memory mechanism significantly enhances segmentation perfor-mance in complex agricultural environments, while YOLOv8 provides initial localiza-tion cues. Considering practical applications, we optimized the deployment of the algo-rithm model on Jetson edge devices and employed TensorRT technology for accelerated inference, thus achieving real-time processing capabilities. Experimental results demon-strate that this method has achieved significant results in both accuracy and efficiency, providing robust visual perception capabilities for intelligent harvesting robots. Specifi-cally, compared to the traditional Mask R-CNN, our proposed SAM-2+YOLOv8 method improved Precision, IoU, and F1 Score by 8.91%, 4.98%, and 5.61% respec-tively, reaching high levels of 93.23%, 94.59%, and 92.34%. Compared to U-Net, these metrics also improved by 3.90%, 4.38%, and 3.01% respectively. These significant performance improvements not only demonstrate the superiority of this method over traditional algorithms but also provide a more reliable and efficient visual solution for apple harvesting automation.

Introduction

Domestic apple picking robots acquire information through a variety of sensing technologies that provide visual representations of the state of the apple tree and fruit, such as computer vision and laser radar (LiDAR) scanning. While LiDAR technology is widely used in certain industrial-grade applications, there are still limitations in home-grade apple picking robots. Firstly, the high cost of LiDAR devices makes them difficult for ordinary households to accept. According to our research data for 2023, mid-range LiDAR sensors suitable for home robots typically cost between 400*and*1000. For example, Velodyne's Puck LITE model is priced at approximately $999. In contrast, high-quality cameras suitable for computer vision typically range from $50 to $200, with the Raspberry Pi high-quality camera module costing only $50. Secondly, the high energy consumption characteristics of LiDAR are not conducive to developing portable or long-working home robots. Typical mid-range LiDAR sensors consume between 5 and 15 watts. For instance, Ouster's OS0-32 sensor has a typical power consumption of 14 watts. In comparison, camera modules used for computer vision usually consume between 0.5 and 2 watts, with the Raspberry Pi camera module V2 consuming only about 0.25 watts. Furthermore, in specific complex fruit tree environ-ments, vision-based methods sometimes demonstrate higher accuracy. This was con-firmed in Underwood et al.'s study[1]. Their research compared the application of LiDAR and visual sensors in fruit tree mapping, particularly in the context of almond orchards. The study found that while LiDAR performed excellently in measuring tree crown volume, vision-based methods showed higher accuracy in de-tecting flowers and fruits. Although Underwood et al.'s research focused on almond trees rather than apple trees, this study still holds significant reference value for the de-velopment of apple-picking robots.

Currently, deep learning has made significant research progress in the field of com-puter vision-based apple recognition and localisation, including classical convolutional neural networks (CNN)[2] and models focused on segmentation tasks such as U-Net[3]. These models have achieved positive results in improving recognition accuracy and reducing manual workload. For example, Bargoti and Underwood[4] used an improved Faster R-CNN model for orchard fruit detection, achieving an F1 score of 0.904 in the apple detection task using their own collected dataset. Their da-taset contained 1120 apple tree images, with 967 used for training and 153 for testing, with an image resolution of 1616×1232 pixels. Häni, Roy, and Isler[5] used a Faster R-CNN-based model for apple detection and achieved a 90.8% success rate using 103 images with a pixel resolution of 1920 × 1080. Li and Jia[6] proposed a novel green apple segmentation algorithm based on ensemble U-Net, specifically designed for precise apple fruit segmentation in complex orchard environments. On their collected orchard

dataset, the method achieved A mean Intersection over Union (mIoU) of 0.9553 in the green apple segmentation task. Their dataset contained 2000 green apple images, with 1600 used for training and 400 for testing, with an image resolution of 512×512 pixels. Compared to traditional U-Net, their ensemble method improved segmentation accuracy by about 3.2% and showed more stable performance under complex back-grounds, occlusions, and different lighting conditions.

Compared to LiDAR, these computer vision-based methods are not only less costly and energy intensive, but also better able to adapt to various complexities in the home environment. In recent years, visual foundation model (VFM), as a new generation of deep learn-ing model, cleverly captures rich visual features by pre-training on massive data[7]. VFM has achieved satisfactory results in the fields of natural images, video analysis and automatic navigation, which provide new ideas for the de-velopment of apple picking robots. For example, CLIPSeg can accurately understand semantic information in orchard environments[8], and ViNT can effectively adapt to a variety of robots' downstream navigation tasks[9], which is crucial for apple picking robots to localise and navi-gate in complex orchard environments. Florence's ability to classify[10], retrieve, detect, and visually query and answer from images to videos can help robots to better understand and analyse orchard scenarios. MathVista addresses aspects of mathematical reasoning and visual under-standing[11], and this capability can be applied to optimise picking paths and improve effi-ciency. Meanwhile, MetaAI Labs has made a breakthrough in the field of target seg-mentation with the introduction of the segment anything model (SAM)[12] . Similar to GPT-4 in the field of natural language processing (NLP)[13], SAM is a prompt-driven VFM, which is able to adapt to images under multiple imaging conditions, providing new possibilities for apple recognition and localisation. SAM-2 ex-pands on this to include video, enabling real-time segmentation of arbitrarily long vide-os. This added functionality allows SAM-2 to be applied to more scenarios that require video processing. Its flexible cue segmentation mechanism adapts to images with dif-ferent lighting conditions, fruit tree varieties, and fruit growth stages, eliminating the need to re-train like traditional deep learning methods when faced with different picking tasks. The approach significantly reduces the need for massive amounts of labelled data while improving the efficiency of apple picking robot development, but also enables users to enjoy more diversified, intelligent and efficient picking services[14, 15]. Despite significant progress in computer vision-based apple recognition and localiza-tion using deep learning, including the application of models such as CNN and U-Net, and the emergence of new-generation VFMs like CLIPSeg, ViNT, Florence, and SAM, there remain unresolved issues. These include:

1. Traditional algorithmic models are less efficient when dealing with large-scale da-tasets, especially in image segmentation tasks, which often leads to longer model train-ing and optimization cycles;
2. Insufficient adaptability of traditional models in complex outdoor environments;
3. Limitations of relying solely on object recognition methods for apple identification, which may result in insufficient accuracy in picking work, especially when dealing with partially occluded, irregularly shaped, or tightly clustered fruits;
4. The need to improve data utilization efficiency;
5. The challenge of balancing real-time performance and high accuracy on edge de-vices.

Furthermore, the method of combining object recognition algorithms with the latest segmentation large models to improve the speed and efficiency of recognition and segmentation, and applying this efficient method with edge computing devices in the field of home robots, still requires in-depth research and exploration. In this paper, we mainly use the latest SAM-2 model with YOLOv8 target detection model to assist the recognition segmentation of apples by a home apple picking robot, and the hardware side uses tensorRT technology on the Jetson edge end device to complete the accelerated recognition work. SAM-2 (Segment Anything Model 2) combined with YOLOv8 can better address the challenges of apple segmentation, mainly due to their complementary advantages. YOLOv8, as an efficient object detection model, can quickly locate apples in images, while SAM-2 excels at precise image segmentation. SAM-2's zero-sample learning capability allows it to adapt to a variety of complex forms of apples without the need for large amounts of scene-specific training data. This eliminates a significant amount of specialized work, such as annotating datasets for segmentation models, as well as training and parameter tuning of segmentation models. The combination of SAM-2 and YOLOv8 models can efficiently achieve a segmentation effect that is capable of segmenting irregularly shaped, partially shaded, and tightly clumped apples commonly found on apple trees. Compared to traditional object detection models, the SAM-2+YOLOv8 model may be more suitable for completing home-level apple picking tasks. In addition, thanks to the strong generalization ability of the SAM-2 large model, thus it can recognize and segment various apple varieties and handle complex outdoor environmental factors, such as variations in lighting conditions and shadows, which are the main challenges in apple picking scenarios.

To address the challenge of balancing real-time performance and high accuracy on edge devices, this research adopts Jetson edge devices on the hardware side and uses TensorRT technology to accelerate recognition work. TensorRT can

optimize the inference performance of deep learning models, significantly improving model running speed on edge devices while maintaining high accuracy. This approach enables complex SAM-2 and YOLOv8 models to achieve real-time inference on re-source-constrained edge devices, providing reliable performance assurance for home robots in practical applications.

To enable more efficient apple segmentation by the algorithmic models in this re-search, data was primarily collected from apple orchards in three different regions of Vancouver, Canada. In each orchard, 3-6 different areas were selected. Images were captured using a Raspberry Pi OV5647 optical image stabilization camera at different times of the day (morning, noon, evening) to capture apple image information under various lighting conditions.

This process yielded approximately 4,820 apple images. Through data augmentation techniques, this was expanded to 6,000 images, forming a new dataset of apples in production environments. These datasets are primarily used for training, validation, and testing of object recognition algorithm models.

These images were annotated by experts in image algorithm processing. Subsequent-ly, the annotated images were verified by agronomists to ensure the reliability of the training data. For data analysis, this study used accuracy, F1 score and, metrics to quantify the ac-curacy of segmentation. In addition, this paper measures the model's accelerated inference time using tensorRT on a Jetson device to evaluate its real-time performance.

This research aims to contribute to the advancement of smart home orchard manage-ment by providing an efficient, accurate, and user-friendly visual recognition and pick-ing system. The goal is to help homeowners better manage their fruit trees, achieve automated picking, and improve picking efficiency and precision. In the field of home robotics, this study combines object recognition algorithms with the latest large-scale segmentation models. This integration has the potential to address many challenges faced by existing technologies. Furthermore, home orchard management systems face unique challenges such as resource constraints, diverse environments, and us-er-friendliness, areas where research is still insufficient. With the development of edge computing devices, it has become possible to apply advanced visual recognition technologies to home scenarios. Research on how to optimize these technologies to adapt to the limitations of home devices is crucial.

By combining advanced computer vision technology with edge computing devices, this research method provides an innovative solution for the field of smart home gardening, which can significantly improve the management and harvesting efficiency of home orchards.

Methods

This study adopts experimental research design, the main experimental part is to collect relevant dataset, use the dataset to train the YOLOv8 detection algorithm, get the anchor coordinates (absolute pixel coordinate format) of the detection target through the target detection algorithm, input the direct detection coordinates into the SAM-2 vision base model, use the box-prompt part of the SAM-2 model to get the apple segmentation in-formation in the detected image. The overall model edge-end computation part uses the tensorRT technique of Jetson Xavier NX to accelerate the model inference speed, and to evaluate the performance of the SAM-2 vision Foundation model combined with the YOLOv8 target detection algorithm on the apple recognition and segmentation task. In this paper, a controlled experiment is designed to compare the two approaches on the same apple image dataset: the experimental group uses the integrated approach of SAM-2 and YOLOv8, while the comparison group uses a traditional computer vision algorithm which include: Mask R-CNN, U-Net, SAM+YOLOv5. The experiments were conducted under strictly controlled conditions using the same standardised dataset, And the dataset is divided into training set, validation set and testing set according to the ratio of 80%, 10%, 10%, and the segmented datasets were imported into the experi-mental and comparison groups respectively, thus ensuring the reliability and generalization ability of the results. This experimental design enables a direct comparison of the performance of the old and new algorithms and an objective assessment of the potential of the integrated SAM-2 and YOLOv8 approach in improving the effectiveness of the apple recognition and segmentation tasks.

Dataset

This study uses two different datasets for the SAM-2 model and the YOLOv8 algorithm. The reason for this setup is to take full advantage of the pre-training capabilities of SAM-2 while ensuring that the YOLO algorithm can be adapted to real-world apple recognition tasks.

Dataset description

1. For the SAM-2 model dataset part, this paper mainly uses the officially provided pre-trained model, which is trained on the SA-V (Segment Anything - Visual) dataset [16]. The reasons for choosing to use the official dataset include its large scale and diversity, the advantages of pre-training, and the potential for migration learning. The SA-V dataset is a vast and diverse visual repository specifically designed for image segmentation tasks. It encompasses millions of im-age samples, each accompanied by precise segmentation masks. What sets this dataset apart is its extensive coverage, ranging from everyday objects to complex scenes, encompassing a rich variety of object categories and environmental con-texts. The images are mainly derived from publicly available datasets such

as COCO, ADE20K, Open Images, etc., as well as web-crawled images using a semi-automated process to generate high-quality segmentation masks. The diversity of the dataset is reflected in the variety of objects in various lighting conditions, angles, scales and occlusions, improving the generalisation ability of the model. The size and diversity of the SA-V dataset endows the model with a strong zero-sample learning capability, enabling it to perform new segmentation tasks with-out task-specific training, which is beneficial for the apple recognition task in this paper. This study ingeniously employed a transfer learning strategy, using the SAM-2 model pretrained on the SA-V dataset as a foundation. This approach allows the generalized visual understanding capabilities acquired by the model from large-scale, diverse data to be precisely adapted and applied to the specific task of apple recognition. Through this knowledge transfer, the study significantly enhanced the accuracy and efficiency of the system in apple identification.

2. For the YOLOv8 target recognition algorithm, a specialised apple dataset was collected for this paper. The reasons for choosing to use an apple dataset from a real-world environment include task specificity, environmental adaptation, and im-proved recognition accuracy. The dataset contains a total of 6000 apple tree images from domestic apple tree orchards in different regions, with three different areas selected for each orchard. The images were taken at different times of day (morning, noon, and evening) to capture different light conditions.

Dataset Collection Process

The research in this paper follows a systematic sequence of steps, starting with a data collection phase, One of the main aspects of the data collection process was the use of a high resolution camera device, the Raspberry Pi OV5647 Optical Stabilizer Camera, to videotape the apple trees at different times of the day. Subsequently, the paper converts the videos into still images by means of a frame-by-frame extraction method. After ini-tial screening, 5834 images were acquired.

In the data preprocessing stage, this paper strictly controlled the image quality. Firstly, blurred images were eliminated, then samples with highly repetitive content were re-moved, and finally images without apple were excluded. After this series of screening, the final dataset of 4820 images was obtained. These images cover a wide range of lighting conditions and complex backgrounds, including a single image with dense fruit and mutual occlusion, as shown in the Fig.1 below, from left to right are the apple tree image data from morning, noon, and evening, randomly selected from three regional datasets.

The detailed statistics of the dataset are shown in the table below, which shows the number of images collected in the morning, noon and evening for the three regional or-chards respectively.

Data Augmentation

Faced with the deep learning models' insatiable appetite for massive training data and the limitations in scale of existing datasets, this study uses data augmentation. The re-search introduced advanced data augmentation techniques to effectively expand both the quantity and diversity of training samples, the initial data augmentation expanded 4,820 images to 9,640, doubling the data volume. However, to ensure data quality and representativeness, this study implemented a rigorous quality control process, ultimately obtaining 6,000 high-quality image data. The specific process is as follows:

1. Initial Data Augmentation: A series of data augmentation techniques were applied, including but not limited to random cropping, horizontal flipping, and rotation. Through these techniques, we successfully generated an average of one new augmented image for each original image, bringing the dataset size to 9,640.

2. Quality Screening: We invited experienced image research professionals to manu-ally review all 9,640 images (including original and augmented images). This step further eliminated approximately 3,640 images that were not suitable for the research ob-jectives.

3. Over-augmentation Prevention: To prevent bias caused by over-augmentation in the dataset, we implemented strict control measures. In the final set of approximately 6,000 images, we limited the number of augmented versions for each original sample, typically not exceeding 1-2 copies. This strategy aims to maintain dataset diversity while reducing potential learning bias that models might develop due to repeated sam-ples.

Through the rigorous data processing process described above, a dataset containing 6000 high-quality images was finally constructed in this study, in which the image res-olution was mainly 640 x 640 px in order to keep the same inputs for YOLOv8 and SAM-2, while the specification information of the cameras used is shown in Table 2 below:

However, increased data often leads to model overfitting. This study mainly judged overfitting by comparing the performance on training and validation sets: if the model performs extremely well on the training set but poorly on the validation set, it may indi-cate overfitting. We primarily used 3-fold cross-validation to evaluate the consistency of model performance across different data subsets, thereby determining if the model is overfitting. If overfitting exists, regularization techniques are used to reduce it. This ap-proach not only helps improve

Fig. 1: (Left) Picture of an apple tree in the morning, (Center) Picture of an apple tree at noon, (Right) Picture of an apple tree in the evening

Table 1 Information about Apple dataset for different time periods in different regions used for training, validation and testing

Collection time\region	Canadian region A orchard	Canadian region B orchard	Canadian region C orchard
Morning	1434	1146	1006
Noon	398	310	234
Evening	132	93	67

the model's detection accuracy but also reduces the risk of overfitting while enhancing the model's adaptability.

Data Annotation

This enhanced dataset is used in this paper as the basis for training, validation and test-ing of the deep learning algorithm. Considering the traditional segmentation algorithms of the experimental group, experts in image algorithms were invited to perform the an-notation, mainly using LabelImg [17] and Labelme [18] tools for data an-notation, which manually added annotations to the apple objects in each image, and the labeled images were verified by agronomists. Finally, to strike a balance between the model's learning capacity and generalization performance, this study employed a classic dataset partitioning strategy. Specifically, the entire dataset was divided into three sub-sets in an 8:1:1 ratio: the majority training set for the model's learning process, a smaller validation set for hyperparameter tuning, and an equally sized test set for evaluating the model's actual performance. This partitioning method aims to ensure that the model can adequately learn data characteristics while effectively preventing overfitting phenomena.

Experimental Design

The model implementation phase uses the PyTorch framework to build the SAM-2 and YOLOv8 algorithms, the YOLOv8 hyperparameter information used in this is shown in the following table 3:

Which are trained on a high-performance workstation equipped with NVIDIA RTX 3060 GPUs , the YOLOv8 model training environment used in this is shown in the ta-ble below.

Due to the zero-shot transfer ability of the SAM-2 large model, this experiment pri-marily focuses on training the YOLOv8 algorithm model. The main experimental pro-cess is as follows:

1. Data Preparation: Collect the dataset mentioned above;

2. YOLOv8 Model Training: Train the YOLOv8 model using the prepared dataset, optimizing its apple detection capability. During training, pay attention to using the optimized hyperparameters from the aforementioned table 3 to improve the model's performance in specific scenarios;

3. Coordinate Conversion: Use the YOLOv8 official API 'results.pandas().xyxy' to adjust the YOLOv8 output to the absolute pixel coordinate format required by SAM-2: bounding box coordinates in the form of [x1, y1, x2, y2];

4. System Integration and Optimization: Load the pretrained SAM-2 model. Create a SAM-2 predictor object, input the adjusted coordinates along with the original image into the SAM-2 model as Box Prompts for the SAM-2 model, thereby generating precise segmentation masks;

5. Performance Evaluation: Use various evaluation metrics (such as Precision, IoU, etc) to measure the integrated system's performance in apple detection and segmentation tasks;

Table 2 Camera specification information

Sensor	Max Resolution	Video Modes	Focal Length	Fixed Focus
OmniVision OV5647	2592 x 1944 pixels	1080p @ 30fps, 720p @ 60fps	3.60 mm ± 0.01	1 m to infinity

Table 3 The YOLOv8 Hyperparameter Information

Parameter name	Parameter values
Batch	32
Epochs	100
Images	640
Lr0	0.01
Lrf	0.01

6. Practical Application Testing: Apply the optimized system to an actual smart apple picking robot for field testing and further optimization.

Meanwhile, Jetson Xavier NX, an advanced edge computing platform, was chosen as the test vehicle in this study, providing ideal conditions for evaluating the efficiency and accuracy of the model in real-world scenarios with resource constraints. This powerful embedded AI platform detects and segments apples in real-time in field environments and is paired with tensorRT technology to accelerate the computational process, thus validating the feasibility and efficiency of the algorithms on resource-constrained edge devices.

The performance evaluation session uses custom evaluation scripts to compute various metrics, where Precision, Recall, and F1 score are computed by the corresponding functions of the scikitlearn library, and the intersection-to-union ratio (IoU) is computed using a custom NumPy-based function. The computation of these evaluation criteria ensures the accuracy and comparability of model performance assessment. Finally, this study discusses its potential in real-world applications and evaluates the adaptability of the algorithms in edge computing environments. This well-designed and executed series of steps, combining a variety of specialised libraries, custom tools and advanced hard-ware platforms, ensures the systematic, accurate and reproducible nature of the study, providing a solid foundation for drawing reliable conclusions, while demonstrating the algorithms' potential for application in real-world agricultural environments.

Network Model of SAM-2

By pre-training on the SA-V large-scale dataset, SAM-2 learns rich image and annota-tion information, thus demonstrating good zero-sample generalisation ability. SAM-2 makes important improvements on the original SAM architecture by using a more effi-cient visual coder and a more powerful mask decoder. This is shown in the figure.

In terms of structural design, the SAM-2 model maintains the fundamental frame-work of its predecessor, comprising three key modules: an image encoder responsible for feature extraction, a prompt encoder that processes user inputs, and a mask decoder that generates precise segmentation results. The collaborative architecture of these three components is shown in Fig. 2 above. The working principle of each component and the component improvements in SAM-2 are described in detail below:

1. Image Encoder: SAM-2 adopts a more efficient visual backbone network, using an EfficientNet-based architecture instead of the original ViT. This change significantly improves the efficiency of the model, while maintaining a strong feature extraction capability. The new image encoder is better able to handle high-resolution inputs, e.g., a 1024×1024 image is processed to yield a feature map reduced to 64×64, maintaining a 16-fold downsampling rate. There are several versions of SAM-2's image encoder, each with a unique trade-off between computational requirements and model performance, including EfficientNet-B0, EfficientNet-B3, and EfficientNet-B5. The main differences between these versions are the depth, width, and resolution of the model. The increased size of the model directly improves its ability to resolve complex features in the input image, which significantly enhances the overall performance. However, larger models require more computational power, which will be a challenge with limited resources.

2. Cue Encoder: SAM-2 retains the cue encoder design of the original SAM and is capable of handling multiple types of cues, including dots, boxes and text. Specifically, the cue encoder includes sparse cues and dense cues. Sparse prompts include point-prompts, box-prompts, and text-prompts, while dense prompts are mainly masks; SAM-2 also uses positional encoding to represent points and boxes, where points are encoded by two learnable tokens to specify

Table 4 The YOLOv8 model training environment

Property Name	Property Value
System version number	Ubuntu 18.04
Python release version	Python 3.8
Deep Learning Framework	PyTorch 1.13.1
Target detection framework	YOLOv8

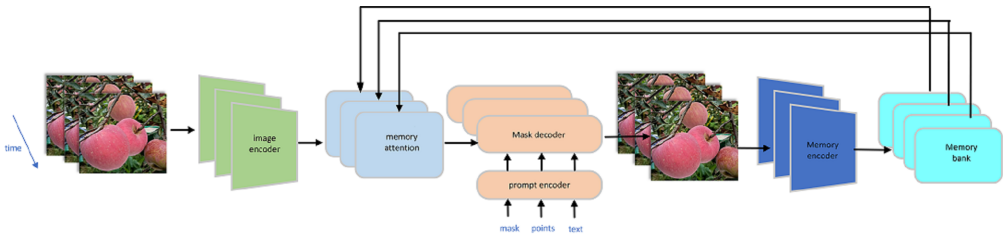

Fig. 2 Network structure diagram of SAM-2

the foreground and background, and the bounding box, on the other hand, is defined by the coordinates of its diagonal vertices. As shown in Fig. 3, the article is tested using apple images from an existing dataset, and the results of the test images for point cues (blue dots denote the selected region) and box cues are clearly shown as examples. In addition, SAM encodes free-form text using a pre-trained text encoder from CLIP[19]. Due to the experimental design principle, this study mainly uses the box-prompt approach for the design.

3. Mask Decoder: SAM-2 has improved the mask decoder with a more efficient de-coding strategy. The new decoder is able to generate high quality masks more quickly while reducing computational overhead.

   In addition to the original three main structures of the SAM model retained above, SAM-2 also introduces several innovative components and mechanisms, among which the three key modules related to the memory mechanism are particularly noteworthy: Memory Attention, Memory Encoder, and Memory Bank. Together, these modules constitute the powerful memory system of SAM-2, which These modules together form a powerful memory system of SAM-2, which significantly improves the performance of the model in processing video sequences and complex dynamic scenes. These modules are described in detail below:

4. Memory Attention: Memory Attention is a core innovation in SAM-2, this module enables the model to efficiently utilise previously stored information when processing new inputs. Its main function is to establish a link between the current frame and the stored memory, enabling selective retrieval of information. The working principle is based on the Query-Key-Value (QKV) mechanism, where the features of the current frame are used as the query and the stored memories are used as keys and values. The advantage of this mechanism is that it can effectively extract relevant content from his-torical information and improve the model's understanding of temporal relevance. Memory Attention allows the model to intelligently retrieve and utilise relevant historical information based on the current content when processing new frames, which greatly enhances the model's performance in the video segmentation task, especially when dealing with scenes with strong continuity or occlusion.

5. Memory Encoder: Memory Encoder is responsible for encoding the information of the current frame into a format that can be efficiently stored and retrieved. Its main function is to compress and encode the input information for efficient storage and subsequent use. The structure usually contains multiple layers of Transformer or Convolutional Networks that are used to extract and encode key features. The output of Memory Encoder is a compact but information-rich representation that is used to update the Memory Bank. Its role is to transform complex visual information into a more abstract and information-intensive form. This encoding not only reduces storage requirements, but also facilitates subsequent information retrieval and utilisation. Through effective encoding, SAM-2 is able to retain the key information of each frame without being distracted by unnecessary details, thus ensuring the retention of key visual features while maintaining efficient processing power.

6. Memory Bank: The Memory Bank is a dynamically updated storage structure used to save and manage the information of history frames. Its main function is to store and manage the encoded history information. Structurally, Memory Bank is usually implemented as a queue or

Fig. 3 An example application of point-prompt and box-prompt.

dynamically re sizable tensor. Its update mechanism uses a first-in-first-out (FIFO) strategy, where new information enters and the oldest information is removed. As the 'long-term memory' of SAM-2, the Memory Bank allows the model to maintain access to historical information while processing long sequences of video. Its dynamic update mechanism ensures that the stored information is always the most up-to-date and relevant, while avoiding infinite memory growth. This design allows the SAM-2 to maintain memory of important historical information while processing long video sequences, as well as updating and adapting to new scene changes in a timely manner, resulting in efficient and consistent performance in video segmentation and object tracking tasks.

YOLOv8 Algorithmic Model

The architectural design of YOLOv8 adheres to the typical framework of deep learning object detection models, comprising four key components: an input data layer, a feature extraction backbone, a feature fusion network, and a task-specific output layer. The structure is shown in Fig. 4 below.

As shown in Fig. 4 above , the input feeds the apple image into the network. The Backbone part of the network uses the Darknet53 architecture, which includes the Convolutional Unit (Conv), the Spatial Pyramid Pooling - Fast (SPPF) module that implements local features and the Cross Stage Partial Network Module (C3) that combines the features of the Cross Stage Partial Network Module (C3) of YOLOv7 and the Cross Stage Partial Network Module (C3) of YOLOv7, which is a highly efficient aggregated network. Network Module Cross Stage Partial Network Module (C3) in YOLOv5 and Cross Stage Partial Network Module - Fast (C2f) that combines the features of the efficient aggregation network in YOLOv7. The Neck network (Neck) part of the network is designed with the CSP2 structure of Cross Stage Partial Network (CSPNet), which enhances the feature fusion capability of the network, and the Head network (Head) is mainly implemented by the Path Aggregation Network Path Aggregation Network (PAN) and Predictive Head to achieve feature fusion and target detection in apple images.

Jeston edge-side computing device

To ensure the reliability of the experiments and the accuracy of performance evaluation, this study employed an NVIDIA GeForce GTX 3060 graphics processor as the core computing platform for executing the training process and performance testing of the algorithmic model. However, due to the limitations of practical application scenarios, especially the need to consider from the aspects of implementing decision-making, low-latency response, energy saving and cost, it is necessary to select suitable edge-side embedded devices as hardware carriers. In particular, due to the consideration of the deployment environment of home-grade apple picking robots in which high-performance computing power is required for real-time processing of image recognition and segmentation tasks, while also taking into account the low-power design for long working hours, and the device needs to be sufficiently small and light-weight for easy integration into mobile robot platforms. In addition, factors such as re-liability, rich interface support, real-time responsiveness, and cost-effectiveness also need to be considered. In light of these requirements, a comparison of various edge-side embedded devices reveals that the NVIDIA Jetson Xavier NX

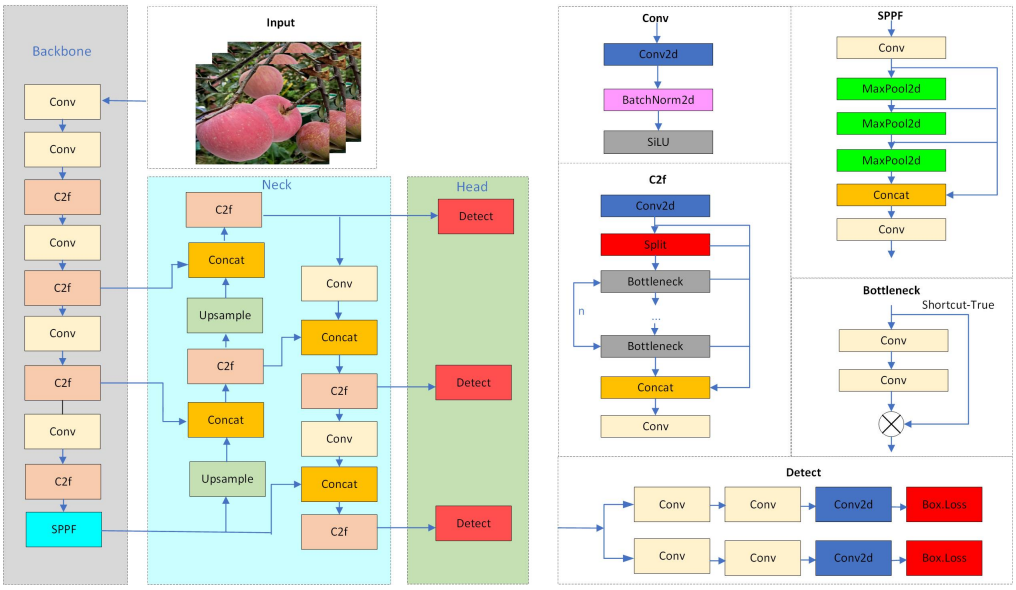

Fig. 4 YOLOv8 network structure

modular system has a smaller form factor and is used for ultra-high-acceleration computing power. Moreover, the various software frameworks and tools optimised by Jetson can achieve high re-al-time and low latency. In view of the above advantages, this paper mainly selects the NVIDIA Jetson Xavier NX module as the core control unit, where the physical diagram of Jetson Xavier NX is shown in Fig. 5:

In order to break the model performance bottleneck and accelerate the inference process, TensorRT technology is used in this study. This advanced deep learning optimisation tool aims to enhance the efficiency of model execution in real-world application scenarios. By applying TensorRT acceleration, this paper is able to significantly improve the inference speed of the apple detection model while maintaining high detection accuracy. TensorRT's optimisation techniques, such as computational graph optimisation, accuracy calibration, dynamic tensor memory management, and kernel auto-tuning, enable the model to make full use of the hardware resources of the Jetson Xavier NX. This study experimentally demonstrates that the model optimized with TensorRT achieved a significant breakthrough in inference efficiency. This performance enhancement is of great significance for building systems capable of real-time detection and precise localization of apples. The optimized model not only meets the stringent real-time requirements of household apple-picking robots but also lays a foundation for improving the overall system's response speed and operational efficiency. Combining the powerful hardware capability of Jetson Xavier NX and the optimisation technology of TensorRT, this paper successfully constructs an efficient and real-time apple detection system, which provides a reliable visual perception capability for home-grade apple picking robots.

Results

Evaluation metrics for image segmentation

In order to verify the effectiveness of the SAM-2 model combined with the YOLO tar-get recognition algorithm in apple picking in this paper, the effectiveness of the method is compared and investigated using evaluation metrics such as Precision, Recall, Inter-section over Union (IoU) and, F1-Score. The definitions of these evaluation indicator formulas are shown below:

$$Precision = \frac{TP}{TP+TF}$$

$$Recall = \frac{TP}{TP+TF}$$

$$F1-Score = \frac{2 \times Precision \times Recall}{precision+Recall}$$

$$10U = \frac{TP}{TP+FP+FN}$$

Where Ture Positive (TP) is the number of true positives that split into positive sam-ples that are actually also positive, False Positive (FP) is the number of false positives that split into positive samples that are actually also negative, False Negative (FN) is the number of false negatives that split into negative samples that are actually positive, and Ture Negative (TN) is the

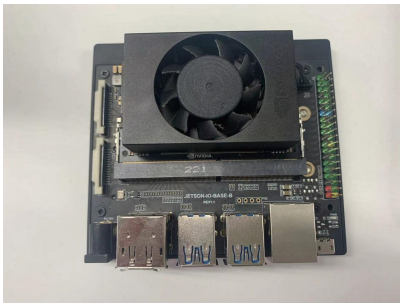

Fig. 5 Jetson Xavier NX entity diagram

number of true negatives that split into negative samples that are actually also negative. The number of true negative samples.

Eq. (1): Precision denotes the ratio of correctly recognized apple pixels among all pixels predicted to be apples; Eq. (2): Recall denotes the ratio of correctly recognized apple pixels among all real apple pixels; Eq. (3): F1-Score is a composite value, which can balance the Recall and Precision to some extent performance; Equation (4): IoU mainly calculates the intersection and concatenation ratio of the real apple region and the predicted apple region.

These evaluation metrics not only measure the model's recognition accuracy of ap-ples, but also reflect the practicality of the model in actual picking scenarios. For exam-ple, a high Precision implies a low misidentification rate, which reduces misoperation on non-apple parts; a high Recall indicates that the model is able to find most of the ap-ples that need to be picked, which improves the completeness of the picking; and the IoU evaluates the model's ability to accurately locate the shape and position of individ-ual apples, which is crucial for the subsequent robotic grasping operation. F1-Score serves as a reconciled average of Precision and Recall, providing a balanced assessment of the model's overall performance.

In order to comprehensively evaluate the performance of the apple segmentation method proposed in this study, a series of comparative experiments were conducted on a uniform dataset. This experimental design ensures the comparability and reliability of the results and enables this study to objectively measure the effectiveness of the pro-posed method. Through these rigorous tests, this study has obtained the following per-suasive experimental data, which will help to analyse in-depth the strengths and poten-tial room for improvement of this study's approach.

Description: Table 5 demonstrates the performance comparison of different models in the apple recognition task. It is obvious from the data that the SAM-2+YOLOv8 com-bination achieves the best performance in all evaluation metrics. Compared with tradi-tional segmentation algorithms, SAM-2+YOLOv8 achieves significant improvements in all aspects. Specifically, compared to Mask R-CNN, SAM-2+YOLOv8 improves Precision by 8.91 percentage points (from 84.32% to 93.23%), Recall rate increased by 5.31 percentage points (from 89.28% to 94.59%), IoU by 4.98 percentage points (from 89.61% to 94.59%), and F1 Score by 5.61 percentage points (from 86.73% to 92.34 per cent). Compared to U-Net, SAM-2+YOLOv8 increases Precision, IoU, and F1 Score by 3.90, 4.38, and 3.01 percentage points, respectively. Notably, SAM-2+YOLOv8 also achieves a small but meaningful improvement over other SAM and YOLO combi-nations. Compared to SAM-1+YOLOv8, Precision improved by 3.11 percentage points, the IoU metrics increased by 2.21 % and the F1 score increased by 2.13 %. In addition, compared to the SAM-2+YOLOv5 method, this study achieves a 1 to 2 % improve-ment in all evaluation metrics, demonstrating overall performance improvement, sug-gesting that the introduction of YOLOv8 has brought additional performance improve-ments. These data clearly show that the SAM-2+YOLOv8 combination not only out-performs traditional methods in apple identification tasks, but also achieves significant improvements in precision, localisation accuracy and overall performance, providing a more reliable visual identification solution for apple picking automation. In order to show the segmentation results of the research methods more clearly, this paper will use the test dataset for the demonstration of the segmentation effect, selecting two of the segmentation methods for comparison, where the demonstration of the com-parison results are shown in Fig. 6 and Fig. 7 below

According to the U-Net segmentation result graph and the YOLOv8+SAM-2 seg-mentation result graph, it can be clearly seen that in the same test image, due to the oc-clusion and other reasons in the actual detection, the U-Net network model is not good at segmenting apples, and it can only identify about 60% of apples, relatively, the seg-mentation accuracy of the YOLOv8+SAM-2 model is about 90%, In the comparison of deep learning algorithms, the YOLOv8+SAM-2 model demonstrates excellent segmen-tation performance, significantly outperforming the Mask R-CNN. This advantage is mainly due to its ability to extract overall semantic features, which allows the model to efficiently capture and process the semantic information of the image, resulting in out-puts of significant quality.

Table 5 Modelling and assessment indicators

Model	Precision%	Recall%	IoU %	F1 Score%
Mask R-CNN	84.32	89.28	89.61	86.73
U-Net	89.33	89.33	90.21	89.33
SAM-1+YOLOv8	90.12	90.3	92.38	90.21
SAM-2+YOLOv5	91.34	91.3	92.89	91.32
SAM-2+YOLOv8	93.23	91.47	94.59	92.34

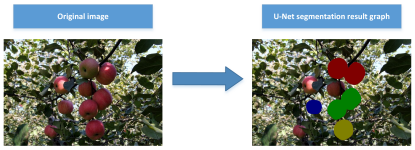

Fig. 6 U-Net Segmentation Visualisation Map

Discussion

This study pioneered the fusion of two advanced algorithms, SAM-2 (Segment Any-thing Model 2) and YOLOv8, to build a completely new vision system, an innovative combination specifically designed to meet the needs of intelligent apple picking robots. Through a comprehensive comparison of multiple deep learning models, including Mask R-CNN, U-Net, and other SAM variant combinations, the study finds that SAM-2 + YOLOv8 performs best in apple detection and segmentation tasks. Preliminary results indicate that the combination of SAM-2 and YOLOv8 has shown potential under identical test conditions. On our test dataset, compared to traditional Mask R-CNN and U-Net, this combination has shown improvements in metrics such as Precision and Intersection over Union (IoU) relative to other tested models. In particular, there has been a significant increase in accuracy compared to traditional segmentation algorithm models like Mask R-CNN and U-Net, and notable progress has been made in real-time processing capabilities. Meanwhile, this study uses TensorRT to optimize the SAM-2+YOLOv8 model to achieve near real-time inference speed while maintaining high accuracy segmentation. Moreover, the SAM-2 + YOLOv8 combination does not require a large amount of training data for segmentation models in specific scenarios. This eliminates the need for extensive professional work such as labeling datasets required for segmentation models and training and parameter tuning of segmentation models. This combination substantially reduces the workload in developing home apple picking systems.

The model shows excellent generalization ability to effectively cope with the complex and changing environment in the orchard can effectively cope with the complexity of different light conditions and the presence of partial shading during the apple picking process, such as different lighting conditions and partial occlusion, and these results are consistent with the initial hypothesis that the SAM-2 vision Foundation Model combined with the YOLOv8 target detection algorithms can effectively identify, localise, and accurately segment apples when integrated with a picking robot system. The YOLOv8 algorithm's strength in rapidly localising apples is complemented by SAM-2's excellence in fine segmentation, further improving the efficiency and accuracy of the overall system. In contrast, the results of this paper echo previous studies that have identified a tangible link between advanced vision algorithms and automated harvesting efficiency, further cementing the crucial role of vision Foundation Modelling combined with traditional target detection methods in improving agricultural automation.

The innovative approach of this study not only demonstrates the potential of large models for application in home smart devices, but also opens up new avenues for accu-rate object recognition and segmentation in complex home environments. The research results demonstrate the possibility of efficiently running advanced AI models on con-sumer-grade hardware, provide new ideas for low-cost and high-performance smart home solutions, and are expected to promote the widespread application and innovation of AI technology in daily life.

However, there are some limitations in this study. Firstly, the experiments mainly focused on apple recognition, and the ability to generalise to other fruits or objects needs to be further verified. Second, although good results were achieved in laboratory and controlled home environments, performance in more diverse and complex real home scenarios requires more testing. In addition, although the TensorRT optimisation significantly improves the performance, the energy consumption and long-term stability of the model need to be further investigated. Finally, given the complexity of the SAM-2 model, performance may vary on different configurations of Jetson devices, which may affect the applicability of the system in products at different price points. Recognising these limitations helps this paper to evaluate the research results more objectively and points the way to future improvements and extensions. This study also faces some challenges. The first and foremost is

Fig. 7 YOLOv8+SAM-2 Segmentation Visualisation

how to balance the performance and accuracy of the model with limited computational resources. Com-pressing a large visual model such as SAM-2 to a scale suitable for home devices while maintaining its strong segmentation capability is a major technical challenge. In addition, ensuring the stability of the system under different lighting conditions, complex backgrounds, and partial occlusion is also a challenge. Another challenge is to optimise the model for different types and maturity levels of fruits in order to extend the application range of the system.

Going forward, this research opens up new possibilities for the application of visual foundation model in smart devices in the home, which is expected to drive further development and innovation in smart home technology. Future research directions could focus on pre-trained models that directly utilise target recognition models, such as the pretrained YOLOv8 or other advanced target detection models. This approach could significantly reduce the effort of dataset collection and training, speed up the development process, and concentrate on how to effectively combine these pre-trained models with SAM-2 and fine-tune them for specific home scenarios. This not only reduces development costs, but also improves the generalisation of the models to a wide range of fruit and household item recognition. In addition, exploring lightweight model compression and quantisation techniques, as well as optimisation strategies for the Jetson platform, will help to further improve the performance of the system on household devices. And in combination with migration learning techniques, investigating how to allow the system to quickly adapt to new object categories will make this technology more flexible and practical. These directions will not only drive the rapid development of smart gardening devices for home use, but may also provide an efficient and low-cost devel-opment paradigm for a wider range of consumer-grade AI applications, accelerating the popularisation and application of AI technology in daily life.

Conclusion

This study achieved significant progress in the vision system for home apple-picking robots by combining SAM-2 with YOLOv8. Experimental results show that, compared to traditional methods, this integrated approach improved the Intersection over Union (IoU) and F1 score by approximately 4.98% and 5.61%, respectively. Furthermore, by applying TensorRT optimization on the NVIDIA Jetson Xavier NX platform, this study achieved an increase in model inference speed while maintaining high-precision re-al-time fruit detection and segmentation capabilities. These performance improvements not only validate the effectiveness of the proposed method but also demonstrate the potential for applying large-scale visual AI models on consumer-grade hardware. The outcomes of this research have potential positive impli-cations for the practicality and reliability of home smart devices and may provide new research directions for the development of smart home technology.

Acknowledgments

SAM-2 data copyright Meta AI, available from https://sa-v.org/.

References

1 J. Underwood, C. Hung, B. Whelan and S. Sukkarieh, *Computers and Electronics in Agriculture*, 2016, **130**, 83–96.

2 K. Fukushima, *Biological Cybernetics*, 1980, **36**, 193–202.

3 O. Ronneberger, P. Fischer and T. Brox, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, 2015, pp. 234–241.

4 I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez and C. McCool, *Sensors*, 2016, **16**, 1222.

5 N. Häni, P. Roy and V. Isler, *Journal of Field Robotics*, 2020, **37**, 263–282.

6 Q. Li, W. Jia, M. Sun, S. Hou and S. Zheng, *Computers and Electronics in Agriculture*, 2021, **180**, 105900.

7 Y. Yuan, International Conference on Machine Learning, 2023, pp. 40519–40530.

8 T. Lüddecke and A. Ecker, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7086–7096.

9 D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose and S. Levine, *arXiv preprint arXiv:2306.14846*, 2023.

10 W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu and Y. Qiao, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14408–14419.

11 P. Lu, H. Bansal, J. Liu, C. Li, H. Hajishirzi and J. Gao, *arXiv preprint arXiv:2310.02255*, 2023.

12 A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson and R. Girshick, Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

13 A. Thirunavukarasu, D. Ting, K. Elangovan, L. Gutierrez, T. Tan and D. Ting, *Nature Medicine*, 2023, **29**, 1930–1940.

14 P. Rajpurkar, E. Chen, O. Banerjee and E. Topol, *Nature Medicine*, 2022, **28**, 31–38.

15 P. Hamet and J. Tremblay, *Metabolism*, 2017, **69**, S36–S40.

16 N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Radle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar and C. Feichtenhofer, *arXiv preprint arXiv:2408.00714*, 2024.

17 Tzutalin, *LabelImg: A graphical image annotation tool*, `https://github.com/tzutalin/labelImg`, 2015.

18 B. Russell, A. Torralba, K. Murphy and W. Freeman, *International Journal of Computer Vision*, 2008, **77**, 157–173.

19 J. Wang, K. Chan and C. Loy, Proceedings of the AAAI Conference on Artificial Intelligence, 2023, pp. 2555–2563.