Hidden Fruit: A Multimodal Framework for Fruit Detection in Agricultural Environments

Harris Song
Henry Samueli School of Engineering, Computer Science
University of California, Los Angeles, United States
harris.song@ucla.edu

Abstract: Modern agricultural practices increasingly rely on automation to address labor shortages and enhance operational efficiency. However, accurately detecting fruits in dynamic, real-world environments remains challenging due to factors such as variable lighting, occlusions, and complex backgrounds. This paper presents a comprehensive framework for fruit detection that includes several key contributions. First, we present a large-scale and comprehensive multimodal dataset designed specifically for fruit detection, which serves as a valuable resource for advancing research in this area. Then, we develop an easy-to-mount handle-grip (portable handheld platform) that can be easily attached to a robot arm, facilitating both data collection and the practical application of fruit picking. Extensive experiments and analyses demonstrate the effectiveness and efficiency of our proposed method, setting a new benchmark for automated fruit detection and harvesting systems.

Keywords: Fruit detection, Multimodal dataset, Agricultural robotics, Real-time object detection

1 Problem Statement

Thermal cameras are widely used for their robustness in harsh conditions and ability to capture temperature-related information. This has led to many thermal datasets. MultiSpectralMotion [1] provides indoor and outdoor thermal images with ground-truth depth from handheld devices. ViViD++ [2] offers outdoor imagery from vehicle-mounted and handheld platforms. SubT-MRS [3] covers varied platforms under degraded conditions but lacks wildland forest imagery.

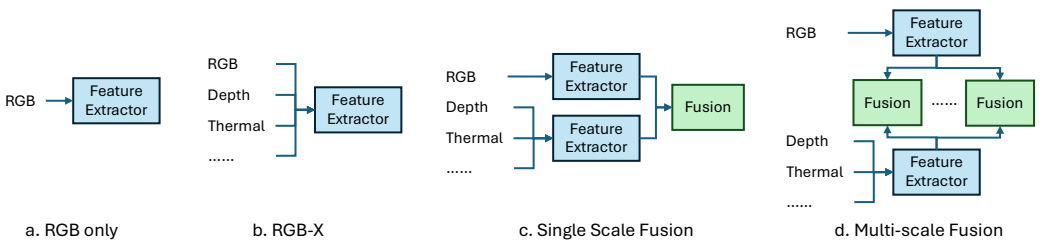

Figure 1: Comparison of different multimodal fusion methods. (a) Training a feature extractor using only RGB images. (b) Sharing a trainable feature extractor between RGB images and other modalities. (c) Single-scale feature fusion within the model. (d) Multiscale feature fusion.

Figure 2: Our multimodal handheld data acquisition platform. The device integrates a ZED 2i stereo camera for RGB, depth, and pose; an Allied Vision NIR camera; and a Teledyne FLIR thermal camera. All components are mounted on a rigid frame with an onboard Jetson Orin Nano for synchronized capture and storage..

Stereo datasets like STheReO [4], MS2 [5], and FIReStereo [6] mainly focus on urban driving. MS2 includes rain scenes but retains typical driving constraints. FIReStereo [6] supports depth estimation for small UAS in degraded environments.

2 System Design / Methodology

Hidden Fruits is a portable, multimodal data-capture rig designed for outdoor fruit detection in environments with variable lighting, clutter, and occlusion. The system synchronously collects RGB, near-infrared (NIR), long-wave infrared (LWIR), and depth imagery, along with 6-DoF pose information. All sensors are mounted on a compact platform centered around an NVIDIA Jetson Orin Nano or AGX host, which runs the capture pipeline and manages synchronization.

```
Dataset Root
 view_01/
    rgb_*.png
    thermal_*.bmp
    metadata_*.json
 view_02/
    ...
 view_08/
    ...

Views           : 8 (rotated every 45)
Frames per view : 121
Total frames    : 1936
Modalities      : RGB, Thermal, Pose
```

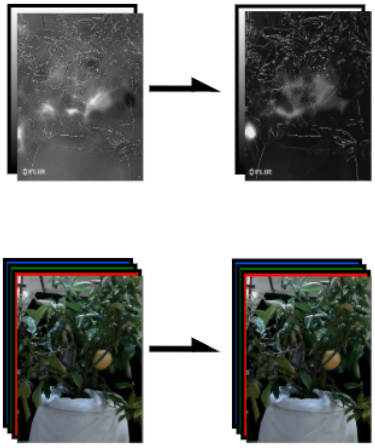

Figure 3: Overview of the dataset sample. Each capture includes synchronized RGB and thermal imagery along with metadata such as pose and timestamps.

3 Evaluation Results

Our initial handheld platform was built on an acrylic base and included an Intel RealSense D435i for RGB-D sensing, a commercial FLIR thermal camera, and an Allied Vision NIR camera. This version enabled early multimodal trials but lacked tight integration.

Table 1: Comparison between different datasets in our experiments.

Dataset	# Images	Modality			Environments				
		RGB	Depth	Thermal	Concealed	Occluded	Nighttime	Interior	Precipitation
ACOD-12K [7]	12,148	✓	✓		✓	✓	✓		✓
MFNet [8]	1,569	✓		✓	✓	✓	✓		
PST900 [9]	894	✓		✓	✓	✓			
NYU Depth V2 [10]	1,449	✓	✓					✓	
SUN RGB-D [11]	10,335	✓	✓					✓	
HiddenHeatedFruit	1936	✓		✓	✓	✓		✓	

We then transitioned to a more compact and integrated handheld rig, which incorporates a Stereo-Labs ZED 2i for RGB, depth, and pose; an Allied Vision NIR camera; a Teledyne FLIR thermal camera; and an NVIDIA Jetson Orin Nano for on-board processing.

3.1 Our Multimodal Dataset

Collecting high-quality datasets for fruit detection in outdoor settings is challenging due to variability in lighting, occlusions from foliage, inconsistent thermal signatures, and tight physical constraints in orchards. Our goal is to address these challenges through a synchronized, multimodal dataset that fuses RGB, depth, NIR, thermal, and pose data.

Calibration between modalities is critical. RGB and thermal sensors have different intrinsic properties and perspectives. We apply offline calibration using checkerboards—heated for thermal alignment and retroreflective for NIR—and fine-tune extrinsics using OpenCV-based reprojection techniques. The data acquisition process follows a structured pipeline. First, all frames from RGB, depth, thermal, NIR, and pose are timestamped and stored per session. In post-processing, we align modalities using frame timestamps and calibrated projection models. Optionally, 3D fruit annotations are generated via semi-automated tools

4 Discussion and Reflections

The HiddenObject framework has the potential to improve object detection across multiple real-world domains. In security and surveillance, enhanced multimodal sensing can support more reliable threat identification and decision-making. In industrial automation, it may help reduce human error, increase safety, and improve workflow efficiency. The framework also supports search-and-rescue operations by enabling more accurate localization in visually complex environments. In agriculture, our fusion-based approach can aid in detecting fruits and crops even when partially or fully obscured by foliage, improving yield estimation, reducing waste, and supporting automated harvesting. It may also assist in real-time crop health monitoring, weeding, and pruning—contributing to more sustainable and precise farming practices.

5 Team Member Contribution

Thank you to Postdoc Andy (Tuan-Anh) Vu for the mentorship, along with Professor Jawed Khalid, the Computer Structures Lab in the Mechanical and Aerospace Engineering. Thank you to Professor Yuchen Cui on the additional support with the Computer Science department. For team member contribution, Harris was the sole contributor to this project. He developed the codebase, collected and processed the dataset, wrote the paper, produced the demonstration video, and built the project website.

References

[1] W. Dai, Y. Zhang, S. Chen, D. Sun, and D. Kong. A multi-spectral dataset for evaluating motion estimation systems. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5560–5566. IEEE, 2021.

[2] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung. Vivid++: Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022.

[3] S. Zhao, Y. Gao, T. Wu, D. Singh, R. Jiang, H. Sun, M. Sarawata, Y. Qiu, W. Whittaker, I. Higgins, Y. Du, S. Su, C. Xu, J. Keller, J. Karhade, L. Nogueira, S. Saha, J. Zhang, W. Wang, C. Wang, and S. Scherer. Subt-mrs dataset: Pushing slam towards all-weather environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22647–22657, 2024.

[4] S. Yun, M. Jung, J. Kim, S. Jung, Y. Cho, M.-H. Jeon, G. Kim, and A. Kim. Sthereo: Stereo thermal dataset for research in odometry and mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3857–3864. IEEE, 2022.

[5] U. Shin, J. Park, and I. S. Kweon. Deep depth estimation from thermal image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1043–1053, 2023.

[6] D. Dhrafani, Y. Liu, A. Jong, U. Shin, Y. He, T. Harp, Y. Hu, J. Oh, and S. Scherer. Firestereo: Forest infrared stereo dataset for uas depth perception in visually degraded environments. *IEEE Robotics and Automation Letters*, 10(4):3302–3309, 2025. doi:10.1109/LRA.2025.3536278.

[7] L. Wang, J. Yang, Y. Zhang, F. Wang, and F. Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17201–17211, June 2024.

[8] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. doi:10.1109/IROS.2017.8206396.

[9] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network, 2019.

[10] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[11] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. doi:10.1109/CVPR.2015.7298655.