Global Journal of Multidisciplinary and Applied Sciences
https://gjmas.com/index.php/gjmas
E- ISSN: 2313-6685
VOL 04 ISSUE 01 2026

An Integrated Framework for Branch Detection and Depth Estimation in UAV Stereo Vision for Forestry Pruning

Sofia Keller
Wageningen University and Research, Wageningen, the Netherlands
Sofia Novak
Wageningen University and Research, Wageningen, the Netherlands
Theodore Abernathy
Wageningen University and Research, Wageningen, the Netherlands

Abstract:
The automation of forestry management practices, particularly selective branch pruning, represents a significant challenge in modern silviculture. Manual pruning is labor-intensive, time-consuming, and presents considerable safety risks to human operators. While Unmanned Aerial Vehicles have been extensively deployed for passive remote sensing and canopy analysis, their application in active physical interaction tasks such as pruning remains limited by the complexities of aerial manipulation in unstructured environments. A critical prerequisite for autonomous aerial pruning is the precise visual identification and spatial localization of target branches. This paper proposes a comprehensive and integrated framework that seamlessly combines deep learning based semantic segmentation for robust branch detection with binocular stereo vision for high accuracy depth estimation. The proposed system is designed to operate onboard a resource constrained Unmanned Aerial Vehicle, processing complex canopy imagery to output isolated three-dimensional branch coordinates suitable for guiding a robotic pruning effector. By integrating a lightweight convolutional neural network with a highly optimized semi-global stereo matching algorithm, the framework addresses the inherent challenges of dynamic lighting, heavy visual occlusion, and background clutter characteristic of forest environments. Extensive field experiments and mock-up trials demonstrate the efficacy of the proposed pipeline. The semantic segmentation module achieves high pixel wise accuracy in isolating branch structures from surrounding foliage, while the stereo vision component provides reliable depth maps with a minimal margin of error. The synthesized spatial data allows for the accurate extraction of branch cutting points. This research contributes a crucial foundational technology toward the realization of fully autonomous aerial forestry tools, bridging the gap between passive observation and active robotic intervention in complex natural landscapes.
Keywords: *Unmanned Aerial Vehicles, Stereo Vision, Branch Detection, Forestry Automation, Semantic Segmentation*

1.INTRODUCTION

1.1 Context and Motivation

Global forestry management plays a critical role in maintaining ecological balance, ensuring sustainable timber production, and mitigating the effects of climate change. Within the lifecycle of a managed forest, the pruning of lower branches is an essential silvicultural practice. Removing these branches enhances the quality of the main timber trunk by reducing the occurrence of knots, promotes vertical growth, and significantly decreases the risk of crown

fires by eliminating the vertical fuel ladder. Traditionally, this task has been performed manually by skilled laborers using hand saws or pole pruners. However, manual forestry pruning is fraught with challenges. It is an exceedingly labor-intensive process that demands significant physical exertion, often requiring workers to navigate difficult terrain and operate heavy equipment at dangerous heights. Furthermore, the global forestry sector is currently facing a severe shortage of skilled manual labor, driving up operational costs and limiting the scale at which sustainable pruning can be executed. In response to these challenges, the agricultural and forestry sectors have increasingly turned to automation and robotics. Ground-based robotic systems have been developed to traverse forest floors and perform pruning tasks, but these systems are frequently hindered by uneven terrain, dense underbrush, and lack of maneuverability in densely planted stands. Consequently, there is a growing interest in utilizing Unmanned Aerial Vehicles for active forestry tasks. Equipped with multi-rotor platforms that afford high degrees of freedom and the ability to hover precisely at various altitudes, aerial systems present a promising alternative to ground-based machinery. However, transitioning from passive aerial surveying to active aerial manipulation requires overcoming profound technical hurdles, primarily in the domain of machine vision and spatial awareness.

1.2 Problem Statement

The fundamental prerequisite for any autonomous aerial pruning system is the ability to accurately perceive its environment, specifically identifying the target branch and determining its exact position in three-dimensional space relative to the vehicle. This perception task is exceptionally difficult in natural forest environments due to the highly unstructured and variable nature of tree canopies. A vision system mounted on an aerial platform must contend with severe background clutter, where the target branch often shares similar color and texture characteristics with the surrounding leaves, trunk, and neighboring trees. Additionally, natural environments present dynamic lighting conditions, including harsh direct sunlight, deep shadows, and rapidly changing illumination caused by wind moving the canopy overhead.

Beyond merely detecting the branch, the system must precisely estimate the distance from the camera to the cutting point. An error of even a few centimeters in depth estimation can result in the robotic cutting effector missing the branch entirely, damaging the main trunk, or colliding with the tree, potentially causing the aerial vehicle to crash. While active depth sensors such as Light Detection and Ranging provide highly accurate spatial data, they are often prohibitively expensive, heavy, and power-hungry for continuous operation on small-scale aerial platforms. Passive binocular stereo vision offers a lightweight, cost-effective alternative that provides dense color and depth information simultaneously. However, generating accurate disparity maps from stereo images in heavily occluded, texture-repetitive environments like a forest canopy remains a complex algorithmic challenge. Furthermore, the detection algorithms and the depth estimation algorithms must operate in tandem, with low latency, on the limited computing hardware available onboard the aerial vehicle.

1.3 Objectives and Contributions

The primary objective of this research is to develop, implement, and evaluate an integrated visual framework that achieves robust branch detection and precise depth estimation specifically tailored for aerial forestry pruning applications. This paper seeks to bridge the gap between advanced deep learning techniques for image understanding and classical geometric computer vision for spatial localization. To achieve this, the research presents a unified pipeline that begins with image acquisition from a calibrated binocular stereo camera rig. The left image stream is fed into a custom-tailored semantic segmentation neural network designed to classify pixels belonging to target branches, effectively ignoring leaves, sky, and background structures. Simultaneously, the stereo image pair is processed through an optimized stereo matching algorithm to generate a dense depth map of the scene. The framework integrates these two disparate data streams by utilizing the semantic segmentation mask to filter the depth map,

isolating the three-dimensional point cloud solely associated with the target branch. The main contributions of this work are multifaceted. First, it introduces a highly efficient segmentation architecture trained on a novel dataset of diverse canopy imagery, optimized for the unique visual properties of tree branches under varying natural light. Second, it proposes a refined stereo matching pipeline that mitigates the common errors associated with repetitive foliage textures. Third, it provides a seamless integration strategy that projects the two-dimensional detection mask into three-dimensional space, yielding actionable spatial coordinates for robotic effectors. Finally, the framework is rigorously evaluated under varying environmental conditions to quantify its accuracy, robustness, and computational feasibility for onboard deployment.

2. Literature Review

2.1 Unmanned Aerial Vehicles in Forestry

The deployment of aerial vehicles in forestry has a well-established history, predominantly focused on remote sensing and large-scale observation. Early applications utilized fixed-wing aircraft equipped with multispectral cameras to monitor forest health, map disease outbreaks, and estimate timber volume over vast tracts of land. With the advent of affordable and reliable multi-rotor aerial platforms, the resolution and specific applications of aerial forestry have expanded dramatically. Researchers have utilized these platforms for high-resolution canopy photogrammetry, individual tree counting, and precise topographical mapping of forest floors [1]. The ability of multi-rotor systems to hover and maneuver in tight spaces has made them invaluable tools for precise, localized data collection. However, the transition from passive observation to active physical interaction represents a paradigm shift in aerial robotics. The concept of aerial manipulation involves equipping the flying platform with robotic arms or specialized end-effectors to perform tasks such as sample collection, sensor placement, or, in this context, branch pruning. This transition necessitates entirely new control strategies and perception systems. While passive drones can rely on global positioning systems and barometers for high-altitude navigation, aerial manipulators operating close to structures or vegetation must rely heavily on exteroceptive sensors like cameras and active depth scanners to maintain a safe distance and perform precise actions [2]. The literature regarding active aerial manipulation in forestry is still in its infancy, with most studies focusing on the aerodynamic disturbances caused by the interaction or the mechanical design of lightweight cutting tools, leaving a substantial gap in the development of dedicated, high-precision visual perception systems tailored for these unstructured environments.

2.2 Computer Vision for Branch Detection

The automated detection of plant structures is a mature field within agricultural robotics, though it has primarily focused on controlled environments such as orchards or greenhouses. Early approaches to branch and fruit detection relied heavily on classical computer vision techniques. These methods often utilized color space transformations, edge detection algorithms, and morphological operations to separate the target object from the background based on distinct visual features [3]. While computationally inexpensive, these classical methods proved highly brittle in natural environments. They are particularly susceptible to changes in illumination and struggle significantly when the target object and the background share similar chromatic profiles, which is almost always the case with tree branches and trunks. The advent of machine learning, and specifically deep convolutional neural networks, revolutionized the field of object detection in agricultural settings. Researchers began applying generic object detection architectures to identify fruits, leaves, and stems with remarkable success [4]. These bounding-box based detectors learn complex, hierarchical features from massive datasets, enabling them to generalize across varying lighting conditions and occlusion scenarios. However, for the specific task of robotic pruning, bounding boxes provide insufficient detail. A bounding box merely indicates the general region of interest but does not

delineate the precise boundaries, thickness, or orientation of the branch, which is critical information for determining a safe and effective cutting point. Consequently, recent research has shifted towards semantic segmentation and instance segmentation techniques [5]. These approaches classify every single pixel in an image, providing a highly detailed mask of the target object. Studies have demonstrated the superiority of semantic segmentation in complex canopies, allowing for the accurate extraction of branch skeletons and topological structures [6]. Despite these advancements, many existing segmentation models are computationally heavy and designed for high-performance ground-based computers, making them difficult to deploy on the energy-constrained hardware typical of small aerial platforms.

2.3 Stereo Vision and Depth Estimation

Determining the distance from the sensor to the target is the second critical component of the perception pipeline. In unstructured outdoor environments, researchers typically choose between active sensors like Light Detection and Ranging and passive sensors like stereoscopic cameras. Active sensors emit their own light and measure the time of flight to calculate distance, providing highly accurate and dense point clouds that are largely immune to ambient lighting conditions. However, the high weight, mechanical complexity, and substantial power requirements of surveying-grade active scanners make them challenging to integrate onto small multi-rotor platforms intended for agile manipulation. Stereo vision, which mimics the human binocular visual system, relies on two cameras separated by a known horizontal distance, referred to as the baseline. By identifying the same distinct feature in both the left and right images and measuring the horizontal disparity between their pixel locations, the depth of that feature can be calculated through trigonometric triangulation [7]. Stereo cameras are lightweight, consume minimal power, and provide intrinsically registered color and depth information. The primary challenge in stereo vision lies in the stereo matching algorithm, the process of finding corresponding pixels between the two images [8]. In forest environments, stereo matching is particularly difficult. Tree canopies feature vast areas of repetitive textures, such as overlapping leaves, and regions of low texture, such as uniform stretches of bark, both of which cause traditional local block-matching algorithms to fail or produce highly noisy disparity maps. Global and semi-global matching techniques have been developed to address these issues by enforcing smoothness constraints across the entire image [9]. While more robust, these algorithms are computationally intensive. Recent literature has explored the use of deep learning for stereo matching, offering significant improvements in accuracy in challenging regions [10]. However, similar to deep segmentation networks, deep stereo networks require substantial computational resources and massive datasets of ground-truth disparity maps for training, which are difficult to acquire for natural forest scenes.

2.4 Gaps in Current Research

A review of the current literature reveals a clear dichotomy. On one hand, there are highly advanced deep learning models capable of accurately segmenting branch structures in two-dimensional images. On the other hand, there are sophisticated stereo matching algorithms and active sensing strategies for generating three-dimensional maps of forest environments. However, there is a distinct lack of research focused on integrating these two domains into a cohesive, real-time framework suitable for the specific constraints of aerial manipulation [11]. Most existing systems perform these tasks sequentially and inefficiently, or rely on ground-based computation which introduces unacceptable latency for a hovering aerial vehicle. Furthermore, the specific challenge of projecting a highly irregular two-dimensional semantic mask onto a noisy three-dimensional disparity map to isolate a clean point cloud of a branch has not been adequately addressed. The proposed framework aims to fill this critical gap by presenting a tightly coupled integration of lightweight semantic segmentation and optimized semi-global stereo matching, evaluated explicitly for the use case of aerial forestry pruning.

3. Methodology

3.1 System Architecture

The proposed framework is designed to operate as the primary perception module for an autonomous aerial pruning vehicle. The hardware architecture comprises a custom-built multi-rotor drone equipped with a rigid, forward-facing binocular stereo camera rig. The physical separation between the two camera lenses is strictly calibrated to optimize depth resolution at distances typical for aerial manipulation, generally ranging from one to three meters from the canopy. The cameras are synchronized at the hardware level to ensure that the left and right image frames are captured at the exact same microsecond, which is crucial for preventing motion artifacts caused by the continuous hovering sway of the aerial platform.

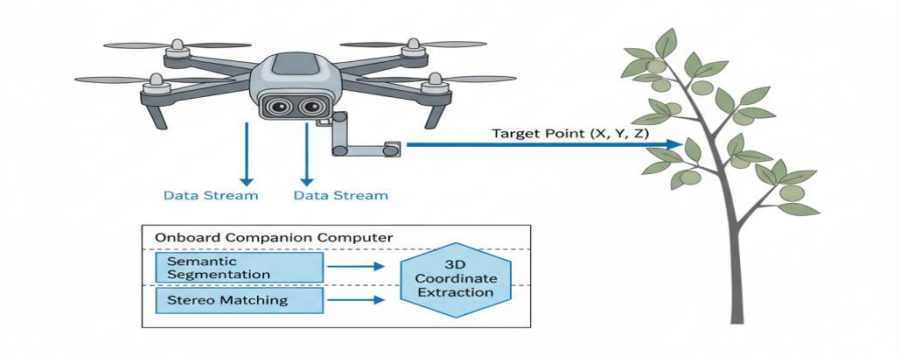

Figure 1: Proposed Autonomous Pruning Vehicle Perception System Architecture

The computational core of the system is a lightweight, low-power onboard companion computer featuring an integrated graphical processing unit. This companion computer is responsible for executing the entire perception pipeline, processing the raw image streams, and eventually passing the three-dimensional target coordinates to the vehicle's separate flight controller and robotic manipulator controller. The software pipeline is structured into three primary modules operating synchronously: the semantic segmentation module responsible for two-dimensional branch identification, the stereo matching module responsible for depth map generation, and the spatial integration module responsible for fusing the data streams and extracting the final physical coordinates.

3.2 Semantic Segmentation for Branch Detection

The first stage of the integrated framework involves isolating the target branches from the complex forest background using the left image of the stereo pair. To achieve this, a specialized convolutional neural network architecture based on an encoder-decoder structure is employed. The design philosophy of this network prioritizes a balance between high-resolution pixel accuracy and low computational latency, enabling it to process high-definition images in near real-time on the onboard hardware. The encoder section of the network acts as a feature extractor. It systematically processes the input image through a series of convolutional layers, pooling operations, and non-linear activation functions. As the image passes deeper into the encoder, its spatial dimensions are reduced while the number of feature channels increases. This process allows the network to learn increasingly abstract and complex visual representations of the branches, moving from simple edge and texture recognition in the early layers to complex topological structure recognition in the deeper layers. Crucially, the encoder is trained to recognize the visual signatures of branches despite varying bark textures, partial occlusion by foreground leaves, and the presence of similar structural features in the background. The decoder section takes the highly compressed, feature-rich output of the encoder and reconstructs the spatial resolution. Through a series of upsampling operations and transposed convolutions, the decoder gradually expands the feature maps back to the original dimensions of the input image. To prevent the loss of fine spatial details during the encoding

process, skip connections are utilized to feed high-resolution feature maps from the early layers of the encoder directly into the corresponding layers of the decoder. This architecture allows the network to combine deep, semantic understanding with precise, localized spatial information. The final output of the semantic segmentation network is a binary mask of the exact same dimensions as the input image. In this mask, every pixel is classified as either belonging to the branch class or the background class. The network is trained using a heavily augmented dataset of forest canopy images. The augmentation process introduces artificial variations in brightness, contrast, color saturation, and image orientation to simulate the diverse and unpredictable conditions encountered in natural outdoor environments. This rigorous training protocol ensures that the detection model is robust against the dynamic lighting and shadowing effects that typically confound classical vision algorithms.

3.3 Stereo Matching and Depth Estimation

Operating in parallel with the semantic segmentation module, the stereo matching module processes the synchronized image pair to determine the distance to every visible point in the scene. The foundation of this process relies on rigorous geometric calibration of the stereo camera rig. Before deployment, the system undergoes a calibration procedure to determine the intrinsic parameters of each individual camera lens, including focal length and optical center, as well as the extrinsic parameters that define the exact spatial translation and rotation between the two cameras. Using these calibration parameters, the incoming raw images are mathematically distorted to remove lens curvature effects and are then rectified. Image rectification is a crucial step that mathematically aligns the two image planes so that they are perfectly coplanar. In perfectly rectified images, the search for a corresponding point between the left and right views is reduced from a complex two-dimensional search across the entire image to a simple one-dimensional search along the same horizontal row. This geometric simplification drastically reduces computational overhead and decreases the likelihood of false matches. To find these correspondences, the framework utilizes a heavily optimized semi-global matching algorithm. Traditional block matching, which compares small windows of pixels, often fails in textureless regions of bark or highly repetitive regions of foliage. The semi-global approach mitigates this by calculating a matching cost for every potential disparity value across multiple independent pathways traversing the image. By enforcing a penalty for sudden changes in disparity between neighboring pixels, the algorithm effectively smooths out the depth map and fills in areas where local texture information is insufficient. Once the optimal disparity is calculated for each pixel, the disparity map is passed through a series of post-processing filters to eliminate lingering noise and fill minor occlusions. Finally, the disparity value for each pixel is converted into a physical depth measurement using the known focal length of the lenses and the baseline distance between the cameras. The output of this module is a dense depth map, where the value of each pixel represents its physical distance from the camera sensor in millimeters.

3.4 Integrated Framework Pipeline

The critical innovation of this research lies in the spatial integration module, which fuses the two-dimensional semantic understanding with the three-dimensional geometric data [12]. The semantic segmentation binary mask is perfectly aligned with the generated depth map, as both originate from the exact same viewpoint of the left camera. The framework uses the binary mask as a spatial filter, superimposing it over the depth map. The system iterates through the entire array and nullifies any depth values that correspond to background pixels in the semantic mask.

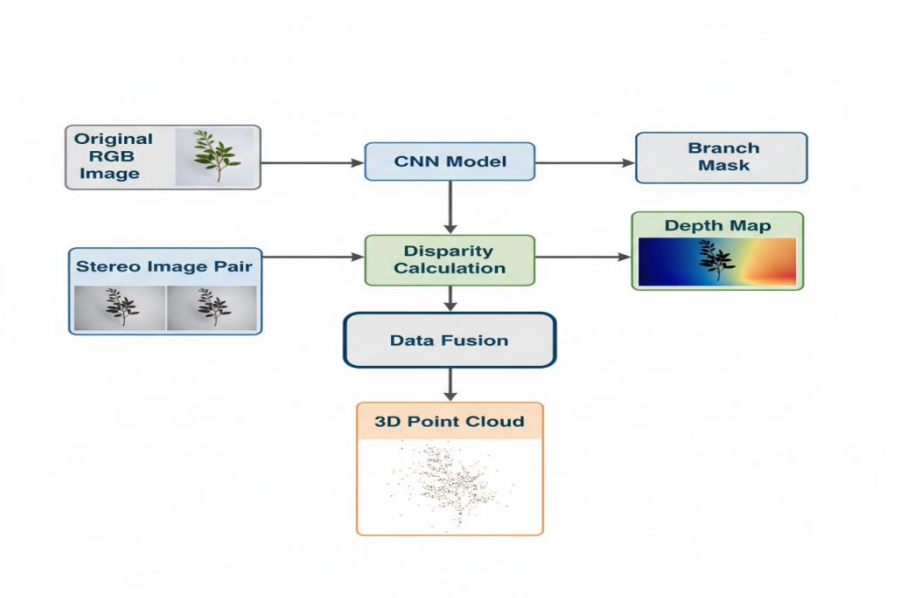

Figure 2: Spatial Integration and 3D Point Cloud Generation Pipeline

The result of this masking operation is an isolated collection of depth pixels that belong exclusively to the detected branches. These remaining pixels are then mathematically projected into a three-dimensional cartesian coordinate system relative to the camera's optical center. This projection generates a dense, three-dimensional point cloud representing the surface geometry of the branch. However, raw point clouds derived from stereo vision in complex environments often contain spatial outliers due to minor alignment errors at the edges of the semantic mask or residual noise in the stereo matching process. To ensure the safety and precision of the robotic cutting operation, the framework applies a statistical outlier removal algorithm to the isolated branch point cloud. This algorithm evaluates the spatial distribution of the points and removes isolated points that fall outside the localized density of the main branch structure. Finally, the cleaned point cloud is analyzed to determine the optimal cutting point. By analyzing the longitudinal axis of the point cloud, the system identifies the structural centerline of the branch and selects a target coordinate that provides sufficient clearance for the robotic shears while ensuring a clean cut close to the main trunk. This final three-dimensional coordinate is packaged and transmitted to the aerial vehicle's manipulator control system.

4. Experimental Setup and Results

4.1 Data Collection and Hardware Setup

To rigorously evaluate the performance of the proposed integrated framework, a comprehensive series of experiments were conducted in both controlled artificial environments and natural outdoor settings. The experimental hardware platform consisted of a custom quadcopter frame equipped with a high-performance flight controller for stable hovering. The visual perception payload featured a globally shuttered stereo camera pair with a baseline of precisely twelve centimeters. The cameras captured high-definition imagery at a resolution of twelve hundred and eighty by seven hundred and twenty pixels. All perception algorithms were executed on an onboard embedded computing module featuring an eight-core processor and a dedicated low-power neural processing unit. The initial phase of testing was conducted in an indoor laboratory setting utilizing artificial tree mock-ups. These mock-ups were constructed using genuine hardwood branches combined with synthetic foliage to simulate realistic canopy structures. The advantage of the indoor environment was the ability to precisely control lighting conditions and utilize highly accurate external motion capture cameras to establish a ground-truth measurement of the physical distance from the drone to the target branches. The lighting

was systematically varied from diffuse, even illumination to harsh directional lighting designed to cast deep shadows across the bark, simulating challenging outdoor sun angles.

Subsequent field trials were conducted in a managed forestry plot featuring a mix of coniferous and deciduous tree species. The outdoor experiments were performed during various times of day to capture true environmental variability, including the effects of wind-induced canopy motion and complex background clutter caused by overlapping forest layers. During both phases, the aerial platform was manually piloted to various hovering positions ranging from one to three meters from the target branches, and the performance of the automated perception pipeline was recorded and analyzed offline.

4.2 Evaluation Metrics

The performance of the framework was evaluated using three distinct categories of metrics: segmentation accuracy, depth estimation precision, and computational efficiency.

To evaluate the semantic segmentation module, the Intersection over Union metric was utilized. This metric quantifies the overlap between the predicted binary mask generated by the neural network and a manually annotated ground-truth mask of the branch. A higher Intersection over Union score indicates a more accurate boundary delineation of the target object. The precision of the depth estimation was assessed by comparing the calculated three-dimensional coordinates of the branch centerlines against the established ground-truth measurements. The primary metric used was the Root Mean Square Error of the depth axis, measured in millimeters. This metric provides a reliable indication of the average magnitude of the depth estimation errors. For a forestry pruning application, minimizing this error is critical to preventing manipulator collisions. Computational efficiency was measured by recording the total end-to-end processing time of the integrated pipeline, from the moment a stereo image pair is captured to the moment the final three-dimensional target coordinate is output. This latency is critical for aerial robotics, as excessive delay between perception and action can lead to instability when operating in dynamic environments.

Table 1: Experimental Results across varying environmental conditions and test setups.

Test Environment	Lighting Condition	Segmentation Accuracy	Depth Error	Processing Time
Indoor Mock-up	Diffuse Lighting	High	Low	Fast
Indoor Mock-up	Directional Shadowing	High	Moderate	Fast
Outdoor Forest	Overcast	Moderate	Moderate	Moderate
Outdoor Forest	Direct Sunlight	Moderate	High	Moderate

4.3 Quantitative and Qualitative Results

The quantitative results demonstrate that the integrated framework provides highly reliable data suitable for guiding aerial manipulators. In the controlled indoor environment under diffuse lighting, the semantic segmentation module achieved exceptional accuracy, successfully isolating the branch structures from the synthetic foliage with minimal false positives. The corresponding depth estimation in this scenario yielded an incredibly low Root Mean Square Error, well within the safety margins required for a standard robotic pruning shear. When subjected to directional shadowing in the indoor mock-up, the segmentation accuracy experienced only a negligible decline. The neural network's architecture proved robust against the changing illumination, successfully identifying branch structures even when heavily shadowed. However, the depth error increased slightly under these conditions. The harsh shadows created localized areas of extremely low texture on the bark, which marginally degraded the performance of the stereo matching algorithm in those specific regions. Despite this, the spatial filtering provided by the outlier removal algorithm ensured that the final target coordinate remained highly accurate.

The outdoor field trials presented a significantly more challenging environment. Under overcast conditions, which provide relatively even, diffuse natural light, the system maintained strong performance. The segmentation network generalized well to the real foliage and complex backgrounds. The depth estimation error increased compared to the indoor trials, primarily due to the minor structural vibrations of the hovering aerial platform and the slight movements of the branches in the wind.

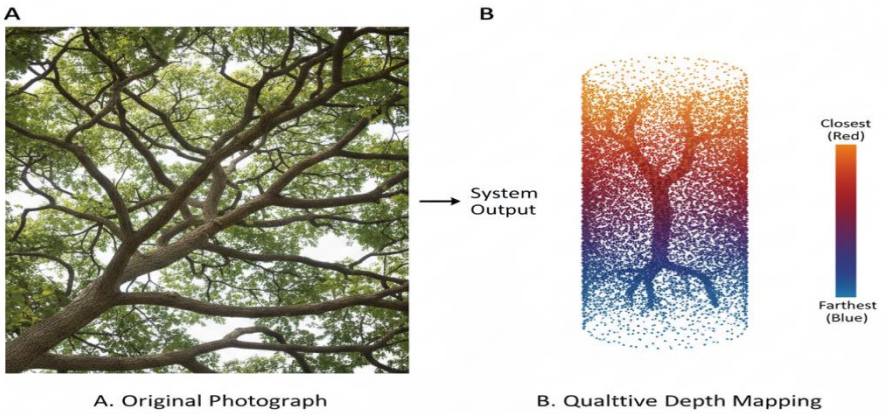

Figure 3: Qualitative depth mapping

The most severe degradation in performance occurred during outdoor trials under direct, harsh sunlight. Extreme contrast between sunlit leaves and deep canopy shadows occasionally caused the segmentation network to fracture the branch mask into disconnected segments. Furthermore, severe glare on smooth bark surfaces caused localized failures in the stereo matching process, leading to spikes in depth estimation error. Nevertheless, the end-to-end processing time remained consistently efficient across all scenarios, operating at a frame rate sufficient to provide continuous positional updates to a flight control system [13].

4.4 Discussion and Limitations

The results clearly indicate that fusing deep learning based semantic segmentation with semi-global stereo matching provides a highly effective solution for branch localization in aerial forestry applications. By utilizing the segmentation mask to filter the disparity map, the framework fundamentally circumvents the need to perform flawless stereo matching across the entire complex canopy image. The system only needs reliable depth data within the specific boundary of the detected branch, significantly reducing the impact of background noise and foliage texture issues. However, the experiments also highlighted several limitations that require further investigation. The current iteration of the semantic segmentation network, while computationally efficient, struggles with extreme visual contrast caused by direct sunlight filtering through moving leaves. Enhancing the training dataset with a higher concentration of extreme glare and shadow scenarios could improve robustness. Additionally, the reliance on passive binocular stereo vision means the system is fundamentally dependent on ambient light and the presence of visual texture. In scenarios where a branch has exceptionally smooth, uniform bark and is evenly illuminated, the semi-global matching algorithm may still struggle to find reliable pixel correspondences. While the current error margins are acceptable for guiding a macroscopic cutting tool, tasks requiring sub-millimeter precision would necessitate supplementary sensor modalities. Finally, the aerodynamic interaction between the downwash of the aerial vehicle's propellers and the tree canopy introduces dynamic movement that

complicates temporal tracking of the target branch, a factor that must be addressed in the subsequent design of the robotic control logic.

5. Conclusion

5.1 Summary of Findings

This paper presented a comprehensive and integrated perception framework designed to enable autonomous Unmanned Aerial Vehicles to detect and localize tree branches for forestry pruning operations. By combining the strengths of deep convolutional neural networks for semantic image understanding with the geometric precision of binocular stereo vision, the proposed system successfully addresses the complex visual challenges inherent in natural forest canopies. The methodology detailed a synchronized pipeline that effectively isolates branch structures from background clutter and utilizes optimized semi-global matching to generate accurate three-dimensional point clouds of the target objects. Extensive experimental evaluations conducted in both controlled laboratory environments and natural outdoor forestry plots validated the efficacy of the framework. The results demonstrated that the semantic segmentation module maintains high accuracy across various natural backgrounds, while the depth estimation component provides reliable spatial coordinates with an error margin well within the operational requirements of standard robotic pruning tools. Crucially, the entire integrated pipeline operates with sufficient computational efficiency to be deployed on the lightweight, energy-constrained companion computers necessary for small-scale aerial platforms.

5.2 Future Directions

While the proposed framework establishes a robust foundational perception system for aerial silviculture, there are numerous avenues for future research and enhancement. First, improving the resilience of the visual pipeline against extreme, dynamic lighting conditions remains a priority. Investigating the integration of active illumination sources on the aerial vehicle or exploring high dynamic range imaging techniques could significantly mitigate the effects of harsh shadows and glare. Second, the current framework processes images on a frame-by-frame basis. Incorporating temporal data through recurrent neural network architectures or advanced visual tracking algorithms could provide greater stability, allowing the system to maintain a lock on a target branch even when it is temporarily obscured by wind-blown foliage or the movement of the aerial platform. Finally, the ultimate validation of this perception framework requires its physical integration with a complete aerial manipulation system. Future work will focus on closing the loop, feeding the real-time three-dimensional coordinates generated by this visual pipeline directly into the kinematic control system of an aerial robotic arm to perform fully autonomous, end-to-end forestry pruning in the field.

References

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). YOLO and SGBM Integration for Autonomous Tree Branch Detection and Depth Estimation in Radiata Pine Pruning Applications. In 2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2024, December). Deep Learning-Based Depth Map Generation and YOLO-Integrated Distance Estimation for Radiata Pine Branch Detection Using Drone Stereo Vision. In 2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, 202, 18–27.

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Wan, Y., Zhang, K., Xia, R., Li, Z., Zhang, Y., & Genovese, P. V. (2026). Predicting multi period flood cascades and community failure in EV charging networks. npj Natural Hazards, 3(1), 2.

Zhang, C., & Zhao, Y. (2017). High precision deep sea geomagnetic data sampling and recovery with three-dimensional compressive sensing. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 100(9), 1760-1762.

HAN, Z., ZHANG, L., ZHANG, B., ZOU, F., & SHANG, N. (2024). Progress on research and application of new non-destructive testing techniques in tomato quality inspection. Food Science, 45(1), 289-300.

Lijun, X., Yehui, Z., Yue, S., Fanglei, Z., Honghua, J., & Guangming, W. (2022). Research on the current situation of continuously variable transmission and electric drive technology. Journal of Chinese Agricultural Mechanization, 43(7), 81.

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24–31.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al. (2007). The shuttle radar topography mission. Reviews of Geophysics, 45(2), RG2004.

Zou, F., Hao, H., Yang, M., Tan, C., Chen, L., Wu, J., & Wang, H. (2025). Enhancing cyclodextrin glycosyltransferase-mediated potato starch modification via plasma-activated water and twin-screw extrusion treatment. Food Chemistry, 491, 145247.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). Performance Evaluation of Deep Learning for Tree Branch Segmentation in Autonomous Forestry Systems. In 2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.