2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Assessing Monocular Depth Estimation Networks for UAS Deployment in Rainforest Environments

Srisai Anirudh Tangellapalli
Nimbus Lab
University of Nebraska-Lincoln
Lincoln, USA
atangellapalli2@huskers.unl.edu

Harman Singh Sangha
Agricultural & Biosystems Eng.
Iowa State University
Ames, USA
hssangha@iastate.edu

Joshua Peschel
Agricultural & Biosystems Eng.
Iowa State University
Ames, USA
peschel@iastate.edu

Brittany A. Duncan
Nimbus Lab
University of Nebraska-Lincoln
Lincoln, USA
bduncan@unl.edu

***Abstract*—The primary objective of this study was to utilize state-of-the-art deep learning-based monocular depth estimation models to assist UAS pilots in rainforest canopy data collection and navigation. Monocular depth estimation models provide a complementary technique to other depth measurement and estimation techniques to extend the range and improve measurements. Several state-of-the-art models were evaluated using a novel dataset composed of data from a simulated rainforest environment. In the evaluation, MiDaS outperformed the other models, and a segmentation pipeline was designed using this model to identify the highest areas of the canopies. The segmentation pipeline was evaluated using 1080p and 360p input videos from the simulated rainforest dataset. It was able to achieve an IoU of 0.848 and 0.826 and an F1 score of 0.915 and 0.902 at each resolution, respectively. We incorporated the proposed depth-estimation-based segmentation pipeline into an example application and deployed it on an edge system. Experimental results display the capabilities of a UAS using the segmentation pipeline for rainforest data collection.**

***Index Terms*—Deep Learning for Visual Perception, Vision-Based Navigation, Aerial Systems: Perception and Autonomy**

I. Introduction

Rainforest canopies are an important part of subtropical forest regions as they help support many plant and animal species in rainforests [1]–[4]. Rainforest studies help in understanding the effects on the biological processes and taking adequate control measures for preserving them. In these studies, data related to the water cycle, vegetation growth, etc., are collected at canopy and ground levels [5], [6]. Canopy sampling has many useful applications in increasing productivity and growth [7]. Most useful samples can only be found at the canopy level because of the increased exposure to sunlight, however, collecting samples from this region of the forest is difficult due to the height. One of the current practices includes using poles to increase reach but this method is limited due to the maneuverability of a pole longer than 10m [8]. Other methods are more involved while providing less comprehensive coverage because they require fixed infrastructure such as canopy cranes or canopy rafts [11], [12].

Fig. 1: Our data sampling drone conducting a mission in Children's Eternal Rainforest, Costa Rica.

Remote sensing has been a crucial part of forest management and monitoring, beginning with the early usage of crewed aerial imagery-based forest inventory to satellite image-based resource monitoring [13]. With the introduction of uncrewed aerial systems (UASs) in the early 2000s as a new tool for collecting imagery-based data, the turnaround time for remote sensing data collection was reduced to hours from days and months [14], [15]. The ability to collect samples from precise locations using UASs has been developed for various environments [8]–[10]. As these systems further develop, they can aid in obtaining soil, water, and leaf samples

This work was funded by the National Science Foundation under awards NRI-1925262 and NRI-1925368, additional support was provided by EDA under award 05-79-06226.

from predetermined and coordinated sites to provide sufficient coverage and spatially distributed data collection in locations where physical access is limited [16]. To efficiently collect canopy samples, the UAS must reach the sampling site and position itself accurately.

UASs perceive their surroundings by employing various kinds of sensors, such as LIDAR, depth cameras, etc. [13]. Physical depth measurement sensors such as LIDAR, while invaluable for their precision in mapping and object detection, confront significant challenges. Foremost among these is their inherent limitation in range, constraining their effectiveness in capturing data over long distances. Moreover, LIDAR sensors frequently encounter noise interference, which can distort readings and compromise the accuracy of the collected data. In addition to range limitations and noise interference, another challenge inherent to LIDAR sensors lies in the complexity of processing the data they generate. Despite their precision in capturing spatial information, LIDAR data is difficult to process effectively. Furthermore, the output typically yields a sparse depth map, presenting only discrete data points rather than a continuous representation of the environment. This leads to further hurdles in utilizing the data for other applications, requiring sophisticated algorithms and computational techniques to fully perceive the environment. To address these issues, alternative methods are currently being studied.

The recent advancements in computer vision due to deep learning led to the rise of powerful monocular depth estimation networks that can perceive depth from a single RGB image. Monocular depth estimation models offer a complementary perspective to LIDAR and other physical depth measurement sensors, potentially circumventing some of the limitations. By harnessing advanced computer vision techniques, monocular depth estimation models aim to infer depth information directly from 2D images, providing a denser and more continuous depth map compared to LIDAR's sparse output. Integrating these models alongside LIDAR sensors holds promise for enhancing overall perception systems, enabling more robust and comprehensive understanding of the surrounding environment. We investigated and evaluated state-of-the-art, deep learning-based, monocular depth estimation networks to aid UASs in rainforest canopy navigation and data collection. After evaluating several depth estimation networks, the *MiDaS* model performed the best in a rainforest environment [17]. MiDaS also achieved the highest average FPS among the tested networks.

After determining the best-performing model, we design a depth-based segmentation pipeline to identify tree clusters. A four-stage pipeline is proposed that utilizes MiDaS and conventional computer vision techniques to segment tree clusters in forest canopy. The pipeline is evaluated using a modified version of the synthetic dataset used to evaluate the depth estimation networks. Additionally, to verify the capability of the pipeline, an application is implemented to utilize the pipeline on a drone to help the pilot approach leaf sampling heights above a tree. Through analyzing the experimental results of the pipeline and the flight information, we demonstrated the capability of our proposed pipeline to aid in rainforest leaf sampling and navigation. In summary, our contributions in this work are listed as follows.

- We constructed a synthetic dataset generator for rainforest environments.
- We evaluated state-of-the-art depth estimation models and highlighted the need for more diverse datasets.
- We designed a depth-based segmentation pipeline to identify tree clusters within rainforest canopies.
- We developed and deployed the proposed segmentation pipeline which runs on the edge.

II. Related Work

In this section, we cover the importance of UAS in rainforest management. Through analyzing the current methods used to monitor rainforests, it is clear the usage of drones helps to cut the cost of intensive forest management and increase economic returns. We also review the usage of computer vision in autonomous navigation to learn the state-of-the-art in segmentation which is helpful for our pipeline. Finally, we review the current state of monocular depth estimation to find the best models and architectures for our pipeline.

A. Rain forest monitoring and management using uncrewed aerial systems

Early forest research in the 1900s was based on ground metering methods and bottom-up observations [18]. For collecting data on forest canopies, tall structures were made to reach the upper heights of the canopy. Constructing such structures took a lot of investment and observations were limited to a small area. Since the development of remote sensing technologies, new methods were found to monitor and manage forest canopies which were cheap and gave access to a larger area. Initial research used satellite imagery but using satellite remote sensing was found to be limiting when precise data collection was needed. Therefore, the introduction of UAS in the early 2000s in forest canopy research has revolutionized the data collected for forest canopy monitoring [13].

There are many studies where UAS were successfully deployed and crucial data was collected. They have been used in surveying forests, mapping canopy gaps, measuring forest canopy height, tracking forest wildfires, and forest management. Koh and Wich [19] surveyed and mapped tropical rain forests in Indonesia. Getzin [20] found a strong correlation between biodiversity and forest gap metrics obtained from drone remote sensing in independent forest regions in Germany. They suggest that UAS imagery is proficient in obtaining high-resolution images of the forest canopy. Forest canopy attributes are critical parameters of forest quantification. Chianucci [21] obtained accurate measurements of canopy structure using high-resolution true-color UAV images.

The drone remote sensing missions provided near real-time intelligence to support forest wildfire management. Ambrosia et al. [22]; Hinkley and Zajkowski [23] employed a large-long duration (24 h) fixed-wing drone for assisting forest wildfire management. Martinez-de Dios et al. [24]; Merino

et al. [25] conducted a series of experiments that indicated rotary-wing drones could effectively collect real-time data on forest wildfires. Simultaneous use of multiple drones allowed larger areas to be measured and obtain complementary views of wildfires. Management of forest plantations is similar to the practice of precision agriculture and can be promoted with drone remote sensing. Felderhof and Gillieson [26] used drone remote sensing to map tree canopy health in a macadamia plantation. A significant correlation was found between spectral radiometry and leaf nitrogen levels by field sampling. Charron et al. [8] proposed a novel tool that is directly installed on a drone to collect leaf samples. Data collection tools such as this would greatly benefit from our proposed pipeline to approach leaf sampling sites and navigate rainforest canopies.

B. Computer vision in autonomous navigation

Computer vision is an important computational task that plays a big role in the field of autonomous navigation. There have been many advancements made in this field, and deep learning-based methods have played a large role in these achievements, such as object detection and image segmentation networks. He et al. [27] built upon previous RCNN models to introduce a state-of-the-art object instance segmentation model, Mask-RCNN. This model accurately identifies objects such as cars and pedestrians in street-level photos and videos and also provides masks that can be used for pixel-level segmentation. Li et al. [28] proposes a real-time image segmentation model by leveraging smaller networks to achieve a high average FPS on difficult datasets. Milioto et al. [29] utilize LIDAR point clouds to segment the surrounding scene to be leveraged in autonomous driving. LIDAR and image segmentation have also been used in the field of tree segmentation. Fekete et al. [30] utilizes LIDAR footage collected by drones to segment trees in urban areas.

C. Monocular Depth Estimation

Deep neural networks have grown to be very popular in computer vision applications, such as image segmentation and classification, and now these networks are being applied to monocular depth estimation. Monocular depth estimation is considered to be the recovery of a pixel-level depth map from a single input RGB image [33], [34]. One of the first methods, Eigen et al. [35] proposed a two-stage network that provided the foundation for modern depth estimation networks. Since then, many different types of networks have been proposed, including CNNs, recurrent neural networks (RNNs), and generative adversarial networks (GANs) [36]–[38]. The proposed methods also utilized different training methods, including supervised, unsupervised, and semi-supervised [36].

The encoder-decoder architecture has contributed greatly to many depth estimation networks and is utilized by the methods evaluated in this paper [17], [33], [34], [40]. The encoder-decoder architecture is a two-stage network in which the encoder extracts important features from the input image at each layer and then the decoder upsamples the features into the required outputs [33]. The encoder-decoder architecture has been applied using both supervised and unsupervised models. Supervised methods are inherently at a disadvantage due to the requirement of ground truth training data [41]. To overcome this issue, the usage of synthetic data is starting to be adopted in training deep learning networks but it is not yet widely used [43]. Currently, there exist two datasets, NYU [44] and KITTI [45], which most state-of-the-art models utilize for training and evaluation, including the model selected in this study.

III. Methods & Materials

A. Depth Estimation Models

We evaluated several existing depth estimation networks to identify the most suitable network for use in a rainforest environment. The resulting model would be the primary model used in the proposed segmentation pipeline. The selected models are: *MiDaS* [17], *Monodepth2* [41], *GLPDepth* [34], *Big-To-Small* [39], *LapDepth* [40]. This subset was selected because the models contributed to the state-of-the-art of monocular depth estimation. The set of selected models includes supervised and self-supervised models. The networks were evaluated using the released public version of pre-trained weights with the configuration specified by their respective authors. The models were evaluated on a PC comprising an AMD Ryzen 9 5900X, 32 GB of RAM, and RTX 3070.

B. Synthetic Rainforest Dataset

The depth estimation models must be evaluated using a dataset comprising diverse forest canopies. Video data was collected at the Texas A&M Soltis Center located next to the Children's Eternal Rainforest in San Juan de Peñas Blancas, San Ramón, Costa Rica. Fig. 1 displays one of the data collection missions conducted in Costa Rica. After analyzing the collected rainforest canopy video data, several scenarios occur frequently that need to be represented in the evaluation dataset. The scenarios were: continuous canopy, canopy with noticeable gaps, and canopy with irregular height of trees. Unfortunately, there is a lack of sufficient real rainforest datasets that represent these scenarios and also contain the necessary depth information to evaluate the selected models. Thus, we developed a novel synthetic dataset generator to create a dataset that provides complete control of the environment.

The generator was implemented using Blender [42] and the included Python API. First, we modeled the trees and landscape and configured the environment to generate a photo-realistic dataset similar to the collected video data. Using Blender's weight painting tool, it was possible to control the random placement of trees to generate the previously mentioned scenarios. Finally, we developed scripts to automate the camera placement and render an RGB image and the corresponding z-depth map necessary for evaluation. The generator created a dataset comprising approximately 2000 images and several 30-second to 1-minute-long videos. To ensure the scenarios are accurately represented in the dataset, 100 images are manually generated for each scenario. The dataset includes forward-facing (45° below the horizon) and

downward-facing (directly facing the ground) images because the different perspectives provide different navigational data. However, only the downward-facing images are used for the segmentation pipeline because the primary perspective for leaf sampling is approaching from directly above the sample site. Examples from the generator are displayed in the top two rows of Figure 2.

C. Segmentation Pipeline

After selecting the primary depth estimation model, we developed a segmentation pipeline incorporating the network. The pipeline expects an input of downwards-facing RGB images or video and produces a segmentation mask that identifies the topmost layer of tree clusters in the canopy. The pipeline handles the input frame-by-frame and consists of four stages: preprocessing, depth estimation, post-processing, and segmentation. The preprocessing stage converts the input image to a normalized grayscale image and resizes the resolution (384×384) to match the input requirements of the depth estimation network. The resulting grayscale image is piped into the depth estimation network to generate a depth map. The depth map is resized back to the original input resolution, after which a Gaussian blur is applied and it is converted to a binary image using thresholding to only highlight the topmost layer of the canopy. This results in clusters that are separated into individual masks for use in leaf sampling or canopy navigation.

D. Evaluation

1) Depth Estimation Model Evaluation Metrics

The following evaluation metrics are utilized to compare the selected depth estimation networks. These evaluation metrics are commonly used across other similar works [36]. However, average FPS (frames per second) and model size were added because they indicate whether the model will be suitable for use in an edge system like a UAS. The model size was calculated by measuring the disk space taken up by the weights of each model in megabytes. The model evaluation metrics are defined as follows.

$$\textit{Accuracies: \% of } d_i \textit{ s.t. max } (\tfrac{d_i}{d_i^*}, \tfrac{d_i^*}{d_i}) = \delta < thr,$$

$$\textit{Abs Rel} = \tfrac{1}{|N|} \textstyle\sum_{i \epsilon N} \tfrac{|d_i - d_i^*|}{d_i^*},$$

$$\textit{Sq Rel} = \tfrac{1}{|N|} \textstyle\sum_{i \epsilon N} \tfrac{||d_i - d_i^*||^2}{d_i^*},$$

$$\textit{RMSE} = \sqrt{\tfrac{1}{|N|} \textstyle\sum_{i \epsilon N} ||d_i - d_i^*||^2},$$

$$\textit{RMSE log} = \sqrt{\tfrac{1}{|N|} \textstyle\sum_{i \epsilon N} ||log(d_i) - log(d_i^*)||^2},$$

$$\textit{Avg FPS} = \tfrac{1}{F} \textstyle\sum_{i \epsilon F} |t_i - t_i^*|$$

d_i is the predicted pixel value from the depth estimation model, and d_i^* is the ground truth pixel value. N is the total number of pixels in the depth maps which are useful. The depth values less than $1e^{-3}$ are ignored. thr is the threshold value to reach for the accuracy. The chosen values are 1.25, 1.25^2, 1.25^3. These values are commonly used in the evaluation of monocular depth estimation models [34], [36], [41]. t_i is the time captured after processing the current frame, and t_i^* is the time captured before the processing begins. F is the total number of frames in the input video.

2) Segmentation Pipeline Metrics

The segmentation pipeline is evaluated by comparing the results with manually annotated images from the generated synthetic dataset using common image segmentation metrics [31], [32].

$$\textit{Intersection over union (IoU)} = \tfrac{1}{N} \textstyle\sum_{i \epsilon N} \tfrac{|p_i \cap g_i|}{|p_i \cup g_i|}$$

$$\textit{Dice score } (F_1) = \tfrac{1}{N} \textstyle\sum_{i \epsilon N} \tfrac{2 \cdot |p_i \cap g_i|}{|p_i \cup g_i|}$$

$$\textit{Avg FPS} = \tfrac{1}{F} \textstyle\sum_{i \epsilon F} |t_i - t_i^*|$$

p_i is the binary mask output from the proposed pipeline, and g_i is the annotated mask. N is the total number of images in the dataset. To calculate the metrics, all of the individual segmentation masks related to an input image are combined into a single, whole binary mask. Then, the IoU and F_1 scores for each predicted and ground truth mask are calculated. Finally, the mean is calculated for both metrics. t_i is the time captured after processing the current frame, and t_i^* is the time captured before the processing begins. F is the total number of frames in the input video.

E. Leaf Sampling System

We designed and implemented an edge system to utilize the proposed depth-based segmentation pipeline that is deployed on a drone and assists the UAS pilot approach leaf sampling sites. The implementation of this system serves as a step in qualitatively validating the efficacy of our pipeline. We aim to expand upon this application based on the results we achieved in this study, and we will present a deeper, quantitative analysis of this application in a future work.

The system is deployed on an Intel®NUC 12 onboard a DJI Matrice 600 drone with Intel®RealSense™Depth Camera (D455) as the main camera. This RGB-D camera is used to simultaneously collect RGB video data and verify the depth map output from the segmentation pipeline. The depth-based segmentation pipeline was developed as a ROS package that runs on the NUC and publishes the resulting depth map and segmentation masks to corresponding ROS topics. A web application is designed to display several frames as a user interface to assist the pilot as they navigate the drone to the leaf sampling site. The frames include the original RGB camera input, estimated depth map, resulting segmentation masks, camera depth map, and an overlay of the original RGB camera input and segmentation masks. The overlay frame combines the original RGB video input and the segmentation masks to highlight the tree clusters that are closest for leaf sampling. The web application is deployed on a ground station connected to the same network as the drone and subscribes to the output topics from the drone.

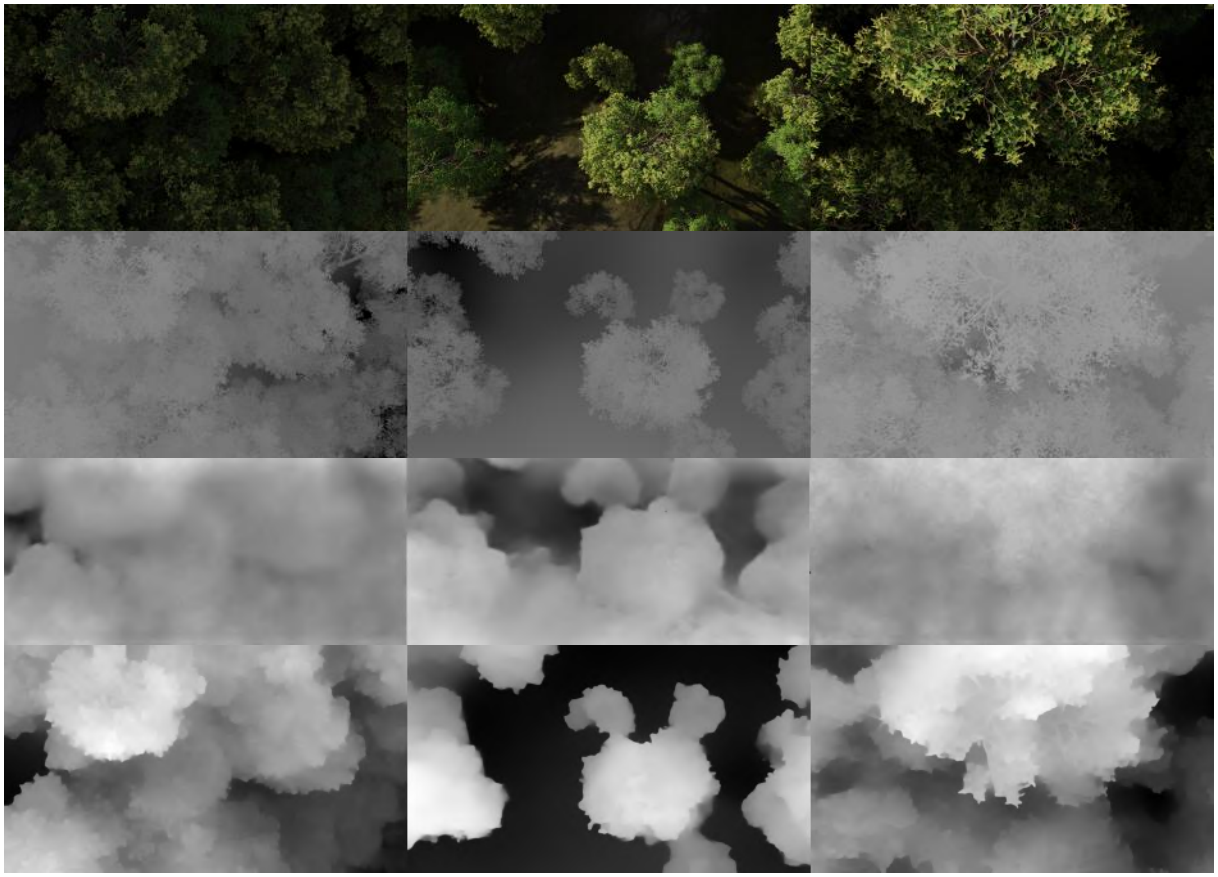

Fig. 2: Model outputs from the downwards-facing synthetic dataset (continuous, gaps, and irregular scenarios from left to right). The top two rows display example input and expected output images. The output images from the third row to the last are from LapDepth and MiDaS, the two best performing models.

TABLE I: Evaluation results for the entire forest dataset for every depth estimation model.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	AbsRel $\downarrow$	SqRel $\downarrow$	RMSE $\downarrow$	RMSE log $\downarrow$
BTS	0.01	0.21	0.55	77.13	65.42	0.50	1.98
GLPDepth	0.29	0.60	0.75	55.41	34.18	0.29	1.76
LapDepth	**0.42**	**0.69**	**0.80**	47.17	26.81	0.23	1.67
MiDaS	0.28	0.56	0.69	**18.28**	**5.61**	**0.22**	**1.54**
Monodepth2	0.24	0.49	0.67	22.66	8.46	0.26	1.68

TABLE II: Evaluation results for average FPS and model size.

Method	Avg FPS $\uparrow$	Model Size (MB)- $\downarrow$
BTS	1.1	549
GLPDepth	17.2	233
LapDepth	1.9	281
MiDaS	**20.0**	1280
Monodepth2	10.7	**56.6**

TABLE III: Evaluation results for the proposed pipeline using MiDaS.

Metric	1080p	360p
IoU score	0.848	0.826
F_1 score	0.915	0.902
Avg FPS	7.549	14.818

IV. Results & Discussion

1) Model Evaluation Results

Table I presents the evaluation results for the selected depth estimation models on the entire synthetic forest canopy dataset. The two best performing models were *MiDaS* and *LapDepth*. Table I provides insight into the ability of the models to generalize to a completely new dataset. The accuracy is low and error is higher than expected across all which indicates the models should be tuned to handle the new environment and dataset. However, another factor that could be impacting the error is the high precision of the synthetic dataset because it provides a very precise depth map as opposed to real depth measurement sensors. Additionally, none of the models were trained on a dataset that contained a rainforest environment. Most were trained using footage captured indoors and outside footage set in an urban environment. Moreover, the original training data heavily leaned towards a forward-facing angle which led to the models predicting the top of the image is further away and the bottom is closer. This highlights the need for a dataset like ours that can provide a new benchmark for future depth estimation models. Based on the results displayed in Fig. 2, the depth estimation maps from the MiDaS model (seen in the sixth row) have the most definition and clarity

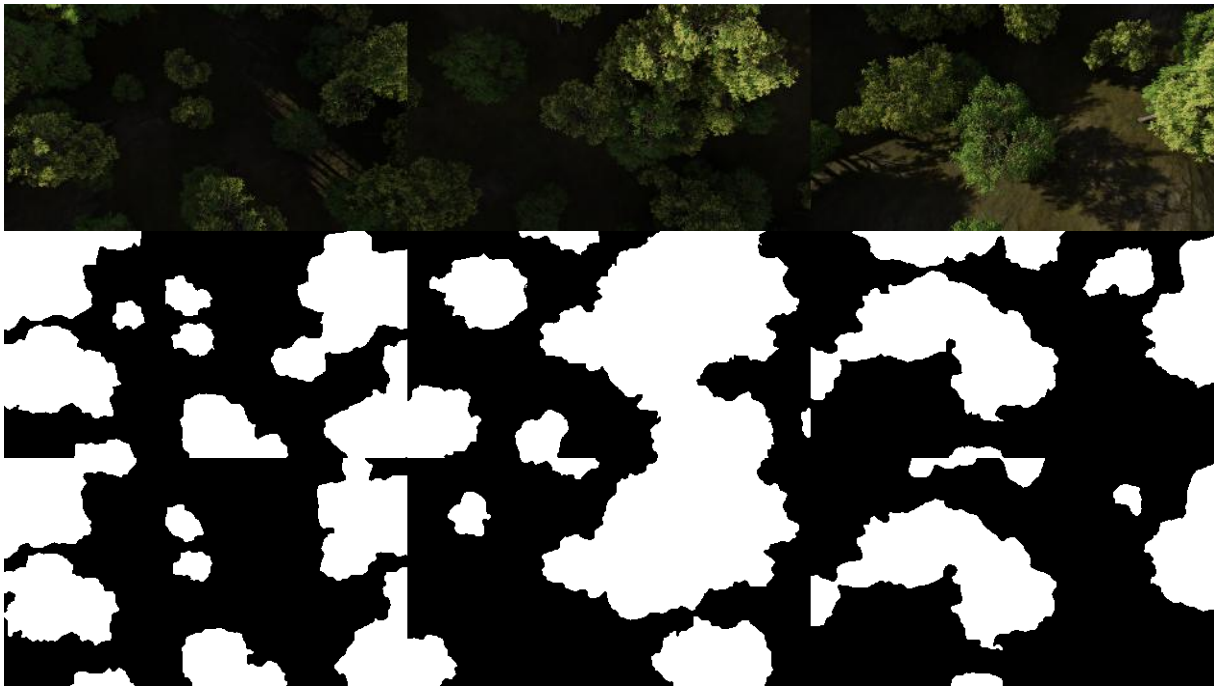

Fig. 3: The top row contains example inputs provided to the pipeline. The middle row displays the ground truth for corresponding images. The bottom row showcases the results from the proposed segmentation pipeline.

which is crucial for our pipeline. The MiDaS model was trained using zero-shot cross-dataset transfer which led to the model being able to generalize well to unseen datasets [17]. Additionally, regardless of the large model size as seen in Table II, the MiDaS model proves to be sufficiently optimized as it had the highest average FPS.

2) Pipeline results

Table III presents the evaluation results for the proposed depth-based segmentation pipeline on the entire synthetic forest canopy dataset at different resolutions. Fig. 3 displays some of the results obtained from the pipeline and the expected outputs. Based on the high IoU and F_1 scores in both input resolutions, the proposed pipeline displays a high ability to correctly segment tree clusters. However, there is room for improvement and there were several factors that impacted the scores, specifically in the fourth stage of the pipeline.

The fourth stage of the pipeline takes a binary image as input, and it segments the image based on connected clusters. The performance of this stage is highly dependent on the input image being properly binarized, such that no important details are lost in the conversion. However, this was fairly likely the cause for several masks to be mispredicted. To convert an image to binary, a threshold must be selected to classify a pixel as 0 or 1, and several algorithms help select an optimal threshold value. The selected thresholding method, Otsu's method, was not always able to correctly identify the optimal threshold. Otsu's thresholding technique is a variable thresholding technique that calculates the histogram of the input image and utilizes the results to determine the threshold. However, the input image has a large range of values and can contain several peaks in the resulting histogram. This can make choosing an optimal threshold difficult and likely the threshold needs to be adjusted manually for every new environment to consistently attain well-defined segmentation masks.

Table III also presents the results the pipeline achieved when processing a high-resolution (1080p) video and a low-resolution (360p) video. As shown in the table, the IoU and F_1 scores are not heavily impacted by the resolution of the input video. Moreover, reducing the input resolution drastically improves the average FPS. This is an important observation because it demonstrates a lower-resolution camera can be sufficient to utilize this pipeline. This approach not only reduces hardware complexity but also opens up possibilities for scalability. By leveraging multiple camera streams, it becomes feasible to achieve comprehensive 360-degree coverage, enhancing the system's ability to detect and navigate around obstacles effectively.

Other methods to improve the average FPS of the proposed method would be to use the other lighter or older versions of the MiDaS model which have significantly fewer parameters. This will lead to a faster inference time because the current bottleneck for the average FPS is the depth estimation network. However, it may significantly impact the accuracy of the resulting depth maps.

3) Leaf Sampling System Analysis

The leaf sampling system provides an example implementation and use case for the proposed pipeline. The system was deployed and tested at Havelock Research Farm, Lincoln, Nebraska, USA. We qualitatively analyze the results of the test flights to determine the efficacy of the proposed pipeline

and system. Our observation of data captured from the depth camera revealed significant noise levels when compared to the predicted depth results and segmentation masks generated by the pipeline. This discrepancy highlights the potential efficacy of using depth estimation models in mitigating the inherent limitations of depth cameras and similar depth measurement sensors. Despite our successful implementation allowing the application to execute tasks in real time, we encountered notable challenges regarding the stability of data transmission between the drone and the ground station. Communication between the drone and ground station took place over a WIFI router with sufficient wireless range, however, the processing and streaming of each frame was too computationally intensive for the NUC to maintain a stable data stream to the ground station. We attempted to use a lighter version of the MiDaS model but ran into significant accuracy issues in our pipeline, even though it did improve the consistency of the data stream.

V. Future Work

To improve the leaf sampling system, developing a lighter depth estimation model tailored to our novel dataset holds substantial promise for enhancing performance and accuracy. This targeted approach not only reduces the computational complexity and memory requirements of the model but also enables it to better capture and understand the intricacies of the rainforest environment. Furthermore, we plan to explore single-line depth estimation techniques, which combine the precision of LIDAR with the dense depth maps derived from monocular depth estimation. This approach presents a promising avenue for achieving high-fidelity depth perception while minimizing computational overhead. Additionally, bolstering the communication infrastructure between the drone and ground station through a stronger radio link will ensure consistent and stable data transmission, vital for real-time monitoring and control. Our current efforts are focused on exploring and implementing these changes to our leaf sampling system and presenting the quantitative evaluation results in a forthcoming paper.

VI. Conclusion

The primary objective of this study was to evaluate and utilize the state-of-the-art monocular deep learning-based depth estimation methods to assist UASs in forest canopy navigation and monitoring. By using monocular depth estimation, LIDAR sensors on UASs can be replaced as they are expensive and heavy to carry as a payload on a UAS. Monocular depth estimation simply requires an RGB input, and it is capable of constructing a dense depth map of the given scene. A synthetic dataset was created using Blender to gather ground truth depth data to evaluate depth estimation networks. After evaluating several networks, the MiDaS model outperformed the other models in nearly all aspects. We proposed a pipeline using the best-performing depth estimation model, MiDaS, for tree cluster segmentation. However, all models encountered high errors highlighting the need for more depth estimation networks targeting under-represented outdoor environments, such as rainforests. The pipeline was comprised of four stages and used RGB video as input. The pipeline's output was a mask of the tree clusters in each video frame. The pipeline was evaluated using a set of manually annotated images from the generated synthetic dataset. The pipeline achieved high IoU and F_1 scores and displays the capability to run in sufficient average FPS. However, a new depth estimation network targeting the rainforest environment would significantly improve pipeline accuracy and viability. Finally, the pipeline was incorporated into an edge system to showcase the feasibility and usefulness of the proposed depth-based segmentation pipeline. This study displays the capabilities of monocular depth estimation in UAS navigation and proposes a novel segmentation pipeline to aid in rainforest leaf sampling and navigation.

References

[1] May, R. The dimensions of life on earth. *Nature And Human Society*. pp. 30-45 (2000)

[2] Novotny, V., Basset, Y., Miller, S., Weiblen, G., Bremer, B., Cizek, L. & Drozd, P. Low host specificity of herbivorous insects in a tropical forest. *Nature*. **416**, 841-844 (2002)

[3] Sechrest, W., Brooks, T., Fonseca, G., Konstant, W., Mittermeier, R., Purvis, A., Rylands, A. & Gittleman, J. Hotspots and the conservation of evolutionary history. *Proceedings Of The National Academy Of Sciences*. **99**, 2067-2071 (2002)

[4] Myers, N., Mittermeier, R., Mittermeier, C., Da Fonseca, G. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature*. **403**, 853-858 (2000)

[5] Oliveira, E., Marimon-Junior, B., Marimon, B., Iriarte, J., Morandi, P., Maezumi, S., Nogueira, D., Aragão, L., Silva, I. & Feldpausch, T. Legacy of Amazonian Dark Earth soils on forest structure and species composition. *Global Ecology And Biogeography*. **29**, 1458-1473 (2020)

[6] Sheil, D. Forests, atmospheric water and an uncertain future: the new biology of the global water cycle. *Forest Ecosystems*. **5**, 1-22 (2018)

[7] Coyle, D. & Coleman, M. Forest production responses to irrigation and fertilization are not explained by shifts in allocation. *Forest Ecology And Management*. **208**, 137-152 (2005,4), https://www.sciencedirect.com/science/article/pii/S0378112704008291

[8] Charron, G., Robichaud-Courteau, T., Vigne, H., Weintraub-Leff, S., Hill, A., Justice, D., Bélanger, N. & Desbiens, A. The DeLeaves: A UAV device for efficient tree canopy sampling. *Journal Of Unmanned Vehicle Systems*. **8** (2020,5)

[9] Ore, J., Elbaum, S., Burgin, A. & Detweiler, C. Autonomous Aerial Water Sampling. *Journal Of Field Robotics*. **32**, 1095-1113 (2015), https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21591

[10] Ore, J. & Detweiler, C. Sensing water properties at precise depths from the air. *Journal Of Field Robotics*. **35**, 1205-1221 (2018), https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21807

[11] STORK, N. Australian tropical forest canopy crane: New tools for new frontiers. *Austral Ecology*. **32**, 4-9 (2007), https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1442-9993.2007.01740.x

[12] Basset, Y. Arthropods of Tropical Forests: Spatio-Temporal Dynamics and Resource Use in the Canopy. (Cambridge University Press,2003), https://books.google.com/books?id=LD3nznHaXPcC

[13] Tang, L. & Shao, G. Drone remote sensing for forestry research and practices. *Journal Of Forestry Research*. **26**, 791-797 (2015)

[14] Ambrosia, V., Hutt, M. & Lulla, K. unmanned airborne systems (UAS) for remote sensing applications. (TAYLOR & FRANCIS LTD 4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON,2011)

[15] Shahbazi, M., Théau, J. & Ménard, P. Recent applications of unmanned aerial imagery in natural resource management. *GIScience & Remote Sensing*. **51**, 339-365 (2014)

[16] Doi, H., Akamatsu, Y., Watanabe, Y., Goto, M., Inui, R., Katano, I., Nagano, M., Takahara, T. & Minamoto, T. Water sampling for environmental DNA surveys by using an unmanned aerial vehicle. *Limnology And Oceanography: Methods*. **15**, 939-944 (2017)

[17] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. & Koltun, V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions On Pattern Analysis & Machine Intelligence*. **44**, 1623-1637 (2022,3)

[18] Birnbaum, P. Canopy surface topography in a French Guiana forest and the folded forest theory. *Tropical Forest Canopies: Ecology And Management*. pp. 293-300 (2001)

[19] Koh, L. & Wich, S. Dawn of drone ecology: low-cost autonomous aerial vehicles for conservation. *Tropical Conservation Science*. **5**, 121-132 (2012)

[20] Getzin, S., Wiegand, K. & Schöning, I. Assessing biodiversity in forests using very high-resolution images and unmanned aerial vehicles. *Methods In Ecology And Evolution*. **3**, 397-404 (2012)

[21] Chianucci, F., Disperati, L., Guzzi, D., Bianchini, D., Nardino, V., Lastri, C., Rindinella, A. & Corona, P. Estimation of canopy attributes in beech forests using true colour digital images from a small fixed-wing UAV. *International Journal Of Applied Earth Observation And Geoinformation*. **47** pp. 60-68 (2016)

[22] Ambrosia, V., Wegener, S., Zajkowski, T., Sullivan, D., Buechel, S., Enomoto, F., Lobitz, B., Johan, S., Brass, J. & Hinkley, E. The Ikhana unmanned airborne system (UAS) western states fire imaging missions: from concept to reality (2006–2010). *Geocarto International*. **26**, 85-101 (2011)

[23] Hinkley, E. & Zajkowski, T. USDA forest service–NASA: unmanned aerial systems demonstrations–pushing the leading edge in fire mapping. *Geocarto International*. **26**, 103-111 (2011)

[24] Dios, J., Merino, L., Caballero, F. & Ollero, A. Automatic forest-fire measuring using ground stations and unmanned aerial systems. *Sensors*. **11**, 6328-6353 (2011)

[25] Merino, L., Caballero, F., Martinez-De-Dios, J., Maza, I. & Ollero, A. An unmanned aircraft system for automatic forest fire monitoring and measurement. *Journal Of Intelligent & Robotic Systems*. **65**, 533-548 (2012)

[26] Felderhof, L. & Gillieson, D. Near-infrared imagery from unmanned aerial systems and satellites can be used to specify fertilizer application rates in tree crops. *Canadian Journal Of Remote Sensing*. **37**, 376-386 (2011)

[27] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask r-cnn. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 2961-2969 (2017)

[28] Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S. & Tong, Y. Semantic flow for fast and accurate scene parsing. *European Conference On Computer Vision*. pp. 775-793 (2020)

[29] Milioto, A., Vizzo, I., Behley, J. & Stachniss, C. RangeNet++: Fast and Accurate LiDAR Semantic Segmentation. *IEEE/RSJ Intl. Conf. on Intelligent Robots And Systems (IROS)*. (2019)

[30] Fekete, A. & Cserep, M. Tree segmentation and change detection of large urban areas based on airborne LiDAR. *Computers & Geosciences*. **156** pp. 104900 (2021), https://www.sciencedirect.com/science/article/pii/S0098300421001928

[31] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R. & Blaschko, M. Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. *International Conference On Medical Image Computing And Computer-Assisted Intervention*. pp. 92-100 (2019)

[32] Kamnitsas, K., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Rueckert, D. & Glocker, B. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*. **36** pp. 61-78 (2017)

[33] Bhat, S., Alhashim, I. & Wonka, P. Adabins: Depth estimation using adaptive bins. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4009-4018 (2021)

[34] Kim, D., Ga, W., Ahn, P., Joo, D., Chun, S. & Kim, J. Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. *CoRR*. **abs/2201.07436** (2022), https://arxiv.org/abs/2201.07436

[35] Eigen, D., Puhrsch, C. & Fergus, R. Depth map prediction from a single image using a multi-scale deep network. *Advances In Neural Information Processing Systems*. **27** (2014)

[36] Zhao, C., Sun, Q., Zhang, C., Tang, Y. & Qian, F. Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*. **63**, 1612-1627 (2020)

[37] Wang, R., Pizer, S. & Frahm, J. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 5555-5564 (2019)

[38] Kumar, A., Bhandarkar, S. & Prasad, M. Monocular Depth Prediction Using Generative Adversarial Networks. *2018 IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops (CVPRW)*. pp. 413-4138 (2018)

[39] Lee, J., Han, M., Ko, D. & Suh, I. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation. *CoRR*. **abs/1907.10326** (2019), http://arxiv.org/abs/1907.10326

[40] Song, M., Lim, S. & Kim, W. Monocular Depth Estimation Using Laplacian Pyramid-Based Depth Residuals. *IEEE Transactions On Circuits And Systems For Video Technology*. **31**, 4381-4393 (2021)

[41] Godard, C., Mac Aodha, O., Firman, M. & Brostow, G. Digging Into Self-Supervised Monocular Depth Estimation. *Proceedings Of The IEEE/CVF International Conference On Computer Vision (ICCV)*. (2019,10)

[42] Community, B. Blender - a 3D modelling and rendering package. (Blender Foundation,2018), http://www.blender.org

[43] Mayer, N., Ilg, E., Fischer, P., Hazirbas, C., Cremers, D., Dosovitskiy, A. & Brox, T. What makes good synthetic training data for learning disparity and optical flow estimation?. *International Journal Of Computer Vision*. **126**, 942-960 (2018)

[44] Nathan Silberman, P. & Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. *ECCV*. (2012)

[45] Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T. & Geiger, A. Sparsity Invariant CNNs. *International Conference On 3D Vision (3DV)*. (2017)