Semantic Segmentation for Path Identification in Vineyards Using Deep Learning

Fama NGOM[1], Huaxi (Yulin) ZHANG[2]

***Abstract*— This study aims to employ semantic segmentation to identify and delineate paths within vineyard images using deep learning models. The goal is to assess the accuracy and efficiency of these models in differentiating between paths and vine areas, thereby facilitating robotic navigation and route management in vineyards.**

The models were trained using a dataset composed of vineyard images, annotated to indicate the location of paths. The evaluation was based on accuracy and processing time to determine the models' ability to correctly segment the paths within the images.

The research is driven by the desire to provide effective tools for optimal vineyard management, by enhancing the planning of robotic routes for precision agriculture and sustainable resource management.

I. INTRODUCTION

In the context of this study, the term "path" refers to lanes that are navigable by robots within vineyard fields. These paths are crucial for robotic navigation and allow for optimized route management within vineyards. Semantic segmentation, a key task in computer vision, finds diverse applications in fields such as medicine, automotive, and agriculture. In agriculture, particularly in vineyard management, the ability to precisely segment elements within an image is essential for optimizing routes and agricultural operations, thereby enhancing efficiency and productivity. Semantic segmentation involves grouping together parts of images that belong to the same object class [1].

The implementation of semantic segmentation in agriculture, and more specifically in vineyard management, is of paramount importance. It allows for the optimization of routes and agricultural operations, thus contributing to increased efficiency and productivity. The need for precise detection of paths in existing vineyards is imperative for the adoption of precise agricultural practices and sustainable agricultural resource management.

Identifying paths in vineyards is a complex task, exacerbated by the unstructured nature of the soils, the presence of grasses, and other vegetations. Conducting experiments in real-world conditions, on existing vineyard fields, rather than in simulations, increases complexity but is paramount to develop concrete and effective solutions suited to real-world situations.

This study aims to explore and assess the effectiveness of advanced deep learning models in semantic segmentation of paths in vineyards, under real-world conditions. The paper is structured as follows: Section 2 reviews the existing literature, Section 3 details the materials and methods employed and Section 4 presents the obtained results.

[1]Université de Picardie Jules Verne, 48 Rue d'Ostende, 02100 Saint Quentin, France fama.ngom, yulin.zhang@u-picardie.fr

Fig. 1. U-Net semantic segmentation inference on vineyard image

II. RELATED WORK

Numerous deep learning models have been developed to enhance the accuracy and efficiency of semantic segmentation. These models are diverse and can be categorized based on their architecture and segmentation approach.

A. Fully Convolutional Models

Fully Convolutional Network (FCN) [2] utilizes end-to-end, pixel-to-pixel training to yield impressive segmentation results, *ParseNet* [3] augments this approach by incorporating global context to the convolutional networks, enabling enhanced segmentation performance on multiple benchmarks with minimal additional computational overhead.

B. Encoder-Decoder Based Models

Models such as *SegNet* [4], *HRNet* [5], *UNet* [6], and *VNet* [7] are classified as Encoder-Decoder-based models. In these architectures, an encoder captures the global context of the image by reducing its spatial resolution, and a decoder reconstructs the original resolution of the image using this context, allowing for detailed and precise segmentation of the objects within the image.

C. Multiscale and Pyramidal Models

Multiscale and Pyramidal Models, like the *Feature Pyramid Network* [8](FPN), have been effectively utilized for semantic segmentation in complex environments such as satellite imagery, where FPN, with its fully convolutional

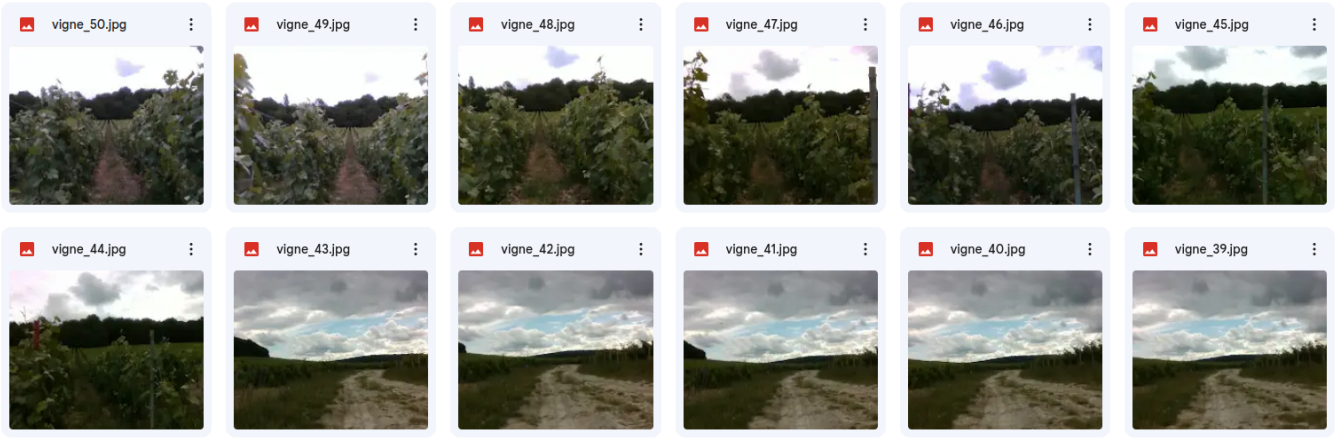

Fig. 2. Dataset: Capturing Images in the Vineyard

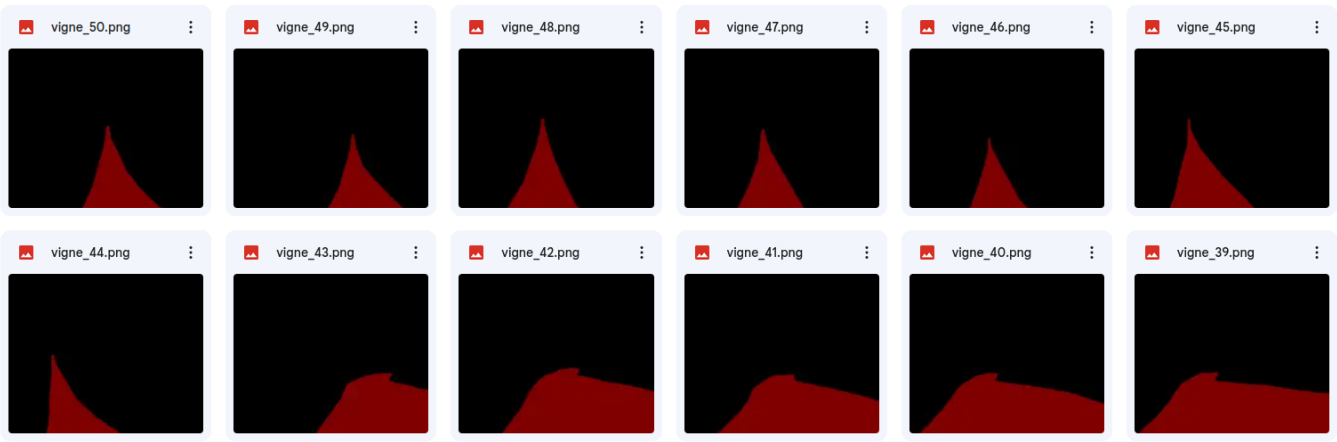

Fig. 3. Dataset: Images Mask generated by label me for Capturing Images in the Vineyard

architecture, enables automatic multi-class land segmentation, exhibiting reliable results and efficient memory usage. The *Pyramid Scene Parsing Network* [9] (PSPNet)leverages global context information through pyramid pooling, achieving state-of-the-art performance in scene parsing tasks across diverse and unrestricted scenes, and setting new records in accuracy on several benchmarks.

D. Attention Models

Attention models like RAN, PANet, CCNet, DANet, PSANet, and EMANet have modernized semantic segmentation by allowing neural networks to sequentially focus on specific parts of a complex input, a technique that emulates human cognitive problem-solving processes. This methodology, which involves breaking down intricate tasks into more manageable segments, has proven to be instrumental in enhancing the precision and effectiveness of segmentation tasks in various complex scenarios.

E. Other Models

Other models like *CNN+CRF*, *DeepLabv3*, and *ReSeg* have also been explored for various semantic segmentation applications.

III. Methodology

Before proceeding with the training on our own dataset, we undertook a series of evaluations using models pre-trained on the ADE20K [10], [11] and Cityscapes datasets. The ADE20K dataset is comprised of over 20,000 scene-centric images exhaustively annotated with pixel-level objects and parts labels, spanning 150 semantic categories, including both stuffs like sky, road, grass, and discrete objects like person, car, bed. Conversely, the Cityscapes dataset focuses on urban street scenes. The objective was to understand the ability of these models to generalize and to observe their performances in terms of segmentation accuracy and inference delay. To do this, we used several implementations, including MMSEG, to ensure a comprehensive and diversified evaluation of each model. The preliminary tests were crucial in enabling the selection of models, specifically UNet, PSPNet, and HRNet, that exhibited superior accuracy in segmenting paths in vineyard images and also demonstrated optimized processing delay. This consideration is pivotal, given the intended application of the model within a robot, where processing delay significantly impacts the feasibility of real-time operations.

The initial attempts at segmentation were not entirely suc-

cessful in obtaining distinct path delineations. The datasets initially used harbored extensive class diversity and complexity and were not vineyard-specific, making the segmentation task more challenging. Given these complexities, there was a clear need to curate and train a new dataset, consisting of images explicitly captured from vineyards, to tailor the model to the specific nuances and requirements of vineyard images.

A. Dataset Vigne Image

We utilized the AgileX Scout Mini robot 4, equipped with a RealSense D455 camera, to capture images in a vineyard. To create a dataset tailored to our specific needs, we used the LabelMe [12] software to annotate the images. We defined two distinct classes: the path and the rest of the image, to maintain a simple structure focused on our main objective. A total of 75 images were meticulously annotated, following this methodology, to form our custom dataset 3. We divide the dataset into three parts; Training Dataset Used to adjust the model's weight, Validation Dataset Used to minimize overfitting. The network weight is not adjusted, Testing Dataset: Used to assess the learning model to confirm the actual predictive power of the network.

We retrained the selected models, UNet, HRNet, and PSPNet, using our specific dataset. These models were chosen due to their demonstrated effectiveness during preliminary tests on standard datasets. The training was carried out using PyTorch, a widely used deep learning framework, to facilitate the optimization and training process of the models.

We employed accuracy and delay as the key metrics to assess the performance of the models in terms of segmenting paths in vineyard images. The accuracy was pivotal in understanding the models' capability to correctly classify each pixel, and the delay was essential to evaluate the real-time applicability of the models, given the intended deployment in robotic navigation within vineyards.

Fig. 4. The AgileX Scout Mini Robot Used for Capturing Images in the Vineyard

B. Attention Models

IV. Experimentation and Result

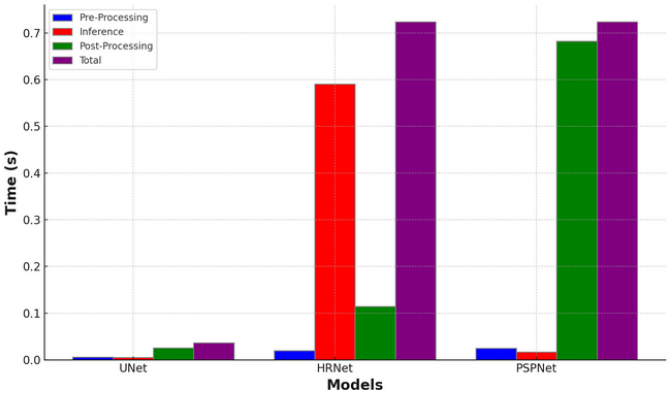

Fig. 5. Comparative Analysis of Processing Times and FPS: UNet, HRNet, and PSPNet

The models were trained and tested on a machine equipped with a Quadro P2200 graphics card and 5GB of RAM. This hardware configuration allowed for the execution of various experiments and assessments needed to compare the performances of the selected semantic segmentation models. This platform provided a stable and efficient environment for running tests and for the comparative evaluation of processing times and efficiency of each model through different phases, such as pre-processing, inference, and post-processing.

The models UNet, HRNet, and PSPNet have displayed varied performances in terms of FPS (frames per second) during the inference phase 5 , a crucial measure for real-time applications. UNet emerged as the fastest model, delivering approximately 27.55 FPS, highlighting its efficiency superiority compared to HRNet and PSPNet, which both secured around 1.38 FPS.

The analysis of UNet's results reveals remarkable performances, both in terms of overall execution time and at each specific stage of processing, including pre-processing, inference, and post-processing.

With an average overall execution time of 0.03580 seconds, UNet demonstrates exceptional efficiency for segmenting 640x480 vineyard images. To put this figure into perspective and for a more intuitive understanding, this time corresponds to processing approximately 27 images per second (frames per second, fps), which is pivotal for real-time applications.

In conclusion, UNet exhibits exceptional and stable performances in all aspects of image processing, positioning it as an ideal choice for applications where the speed and accuracy of inference are paramount.

V. Discussion

During our research, other significant works have emerged, notably the "Segment Anything Model" (SAM) by Meta AI. SAM is a promptable segmentation system, characterized by its zero-shot generalization to unfamiliar objects and images,

without the need for additional training. This model has the capability to "cut out" any object, in any image, with a single click, representing a notable advancement in the field of computer vision.

However, a major limitation of SAM is its extended processing delay, which is currently not suited for real-time applications. Thus, a potential area for future exploration could be the enhancement and adaptation of SAM to our specific context of vineyard segmentation. The goal would be to modify and optimize this model to reduce processing delay, thereby enabling real-time segmentation of paths in vineyards, which would be beneficial for robotic navigation applications in such environments.

VI. CONCLUSIONS

This study has assessed the performances of UNet, HRNet, and PSPNet for the semantic segmentation of paths in vineyards, utilizing a specific dataset. UNet stood out, offering fast and accurate inference, with 27 FPS, on a GPU Quadro P2200 with 5GB of RAM, making it optimal for real-time applications such as robot navigation in vineyards.

Future research could focus on utilizing the segmentation results to develop advanced navigation strategies and optimizing models to enhance precision and inference speed in diverse environments. The integration of additional learning techniques could also augment the autonomous navigation capabilities of robots.

REFERENCES

[1] M. Thoma, "A survey of semantic segmentation," *ArXiv*, vol. abs/1602.06541, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:9869210

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[5] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[8] S. Seferbekov, V. Iglovikov, A. Buslaev, and A. Shvets, "Feature pyramid network for multi-class land segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 272–275.

[9] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[11] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," *International Journal of Computer Vision*, vol. 127, pp. 302–321, 2019.

[12] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, pp. 157–173, 2008.