The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

Unleashing the Reasoning Capabilities of Vision Language Models for Effective Image-based Roadside Tree 3D Measurement

Chen Long[1], Zhen Dong[1], Bisheng Yang[1]

[1] LIESMARS, Wuhan University, Wuhan 430079, China - (chenlong107, dongzhenwhu, bshyang) @whu.edu.cn

Keywords: Vision Language Model, Deep Learning, Street View Images, Roadside Tree Measurement

Abstract

The Diameter at Breast Height (DBH) and Tree Height (TH) are key morphological parameters of roadside tree, their accurate measurement is conducive to quantifying various ecological benefits of trees. Compared with traditional field survey or laser scanning methods, using low-cost street view images as an alternative data source is a promising measurement method. However, existing methods rely on preset reference systems or manual interpretation, which results in poor generalization and low efficiency. Recently, Visual Language Models (VLMs) have shown potential in mimicking human visual reasoning, but their direct application fails to address 3D measurement tasks. To tackle this, we propose a VLM-based Tree 3D Measurement Network, named VLM-TMN. Our key idea is to adjust VLMs to focus on the semantic and geometric information of trees to achieve effective measurement. Specifically, it contains several designs: 1) A Depth Projector Module integrating explicit depth supervision and implicit depth encoding to enhance geometric understanding. 2) A Magnifying Glass Strategy that amplifies visual perception by dynamically focusing on critical tree regions. Built upon LLaVA-7B, our method reduces DBH measurement errors from 24.39 cm to 7.08 cm RMSE, achieving a 7.57% improvement over standard supervised fine-tuning approaches and significantly outperforming existing methods (7.08 cm $<$ 15 cm). The results demonstrate how VLM-TMN can be effectively repurposed for urban ecological parameter quantification, providing a cost-effective solution for sustainable city planning.

1. Introduction

Roadside trees are an important part of the urban ecosystem. Diameter at Breast Height (DBH) and Tree Height (TH) are the most important morphological parameters of trees, and their accurate measurement can be used to quantify various ecological benefits of vegetation, serve in urban planning, and provide decision-making support for sustainable urban construction.

In the past, field surveys using mobile devices (e.g., altimeters, LiDARs) have provided accurate roadside tree measurements, but their high costs and labor demands hindered widespread application. Recently, with the rapid expansion of geographic big data, some scholars have begun to use low-cost, high-coverage street view images as an alternative data source. As a pioneer, Wang et al. (Wang et al., 2018) used absolute size reference objects in the image to measure trees using the proportional relationship between pixels. Although this approach is accurate, the measurement process still relies on human interpretation of the images. After that, Choi et al. (Choi et al., 2022) and Liu et al. (Liu et al., 2023a) proposed automated algorithms that provide absolute scale by fixing the camera's pose and height, or by estimating depth, and predicted tree masks to replace the manual interpretation process. However, these methods are based on pre-set reference systems, and this measurement benchmark presents significant vulnerabilities in complex urban scenes. To this day, efficient and accurate measurement of trees from images remains a challenge.

When human experts make visual measurements, they will actively reason and find reference clues (e.g., the sign height) rather than relying on predetermined standardized objects. Recently, Visual Language Models (VLMs) (Liu et al., 2023b, Yang et al., 2023) have been designed to simulate this active human thinking process. Benefiting from this technology, our goal is to unleash the reasoning capabilities of VLMs to achieve efficient image-based tree 3D measurements. However, directly applying these vision paradigms to 3D measurement tasks is not effective. *The geometric semantic association of objects is the key information for visual measurement*, but VLMs lack depth information guidance during training, resulting in their inability to understand spatial geometry. In addition, due to the limit of fine-grained semantic expression of visual encoders, VLMs are difficult to focus on key visual areas based on text input.

To solve this problem, we propose a VLM-based Tree 3D Measurement Network, named VLM-TMN. *Our key idea is to adjust the VLMs to focus on the semantic and geometric information of the object to achieve effective measurement*. Specifically, it contains several designs. 1) We introduced the Depth Projector Module to extract geometric information through explicit depth supervision and implicit depth coding. 2) To force VLMs to pay attention to the visual details of trees, we designed a magnifying glass strategy that enhances the visual perception of the model by magnifying key areas and expanding the visual features of the image. We use Llava-7B (Liu et al., 2023b) as our baseline, and we improved its DBH measurement root mean squared error (RMSE), from 24.39 cm to 7.08 cm. In addition, compared to using Supervised Fine-Tuning (SFT) for VLMs, our design brings +7.57% improvements in RMSE, and makes VLM-TMN far exceed existing methods (7.08 cm $<$ 15 cm).

In summary, our contributions are as follows.

- We propose VLM-TMN, which releases the visual reasoning capabilities of VLMs to estimate morphological parameters of trees from images.
- We designed a Depth Projector Module and a Magnifying Glass Strategy to adjust the VLMs to focus on semantic and geometric cues, respectively.

- Extensive experiments demonstrate the strength of our methods. Compared with existing methods, VLM-TMN achieves accurate and robust parameter computation.

2. Related Work

2.1 Tree Measurment Methods

Diameter at breast height (DBH) and tree height (TH) are the key morphological parameters of trees (Yang et al., 2020). Depending on the datasource, existing methods can be categorized as follows.

Conventional Measurement Techniques. Traditional methods rely on manual field measurements by means of a tape measure or optical instruments. DBH is usually measured using a circumference ruler or a wheel rule (West, 2009), while TH measurements are mostly carried out using a Blume-Leiss altimeter (Villasante and Fernandez, 2014) based on the principle of triangulation. These methods require a lot of manpower for data recording, which is time-costly and severely limits the efficiency of the measurements.

LiDAR-driven Approaches. With the development of laser scanning equipment (Yang et al., 2024), researchers have begun to utilize these efficient and accurate digital measurement tools to enhance the efficiency of tree measurement. For example, Airborne laser scanning systems (ALS), with their large acquisition range and high acquisition efficiency, are frequently used for tree height measurements in forestry (Giannetti et al., 2018). However, due to canopy shading, this top-down scanning method of airborne laser scanning systems makes it difficult to achieve accurate DBH measurement (Mielcarek et al., 2020). In contrast, the use of top-down ground station laser scanning as well as mobile and backpack laser scanning systems (Estornell et al., 2021, Campbell et al., 2023) are more often used to measure precise geometric features of individual trees. They use Simultaneous Localization And Mapping (SLAM) or depth completion technology (Long et al., 2024) to achieve real-time mapping and provide accurate measurements of tree parameters (Pierzchała et al., 2018). Despite the excellent accuracy, the high equipment cost limits its wide applications (Wu et al., 2023).

Vision-based Methods. Street-view image is a cost-effective data source with significant scalability potential. However, the inherent scale ambiguity in monocular images poses fundamental challenges for vision-based measurement applications. (Wang et al., 2018) pioneered an approach employing standardized urban objects as reference scales, assuming coplanar alignment between targets and references for pixel-to-metric conversion via ImageJ software (Schneider et al., 2012, Rueden et al., 2017). Subsequent work by (Hu et al., 2023) expanded the reference database and applied the methodology to analyze roadside tree distributions in Hangzhou. While effective, these techniques remain constrained by manual intervention requirements. Recent automated implementations (Choi et al., 2022, Liu et al., 2023a) leverage deep learning for tree instance segmentation, deriving dimensional estimates from pixel counts. These methods necessitate stringent camera calibration parameters and level imaging conditions to establish pixel-ground truth relationships. Such engineered constraints limit methodological generalizability across diverse scenarios. Moreover, performance is intrinsically tied to segmentation quality, with complex urban environments exacerbating error propagation in purely appearance-based measurement systems.

In summary, efficient and accurate measurement of trees from images remains a challenge. To solve this, we mimic the human thought process by unleashing the visual reasoning capabilities of VLMs to achieve efficient image-based tree 3D measurements.

2.2 Vision Language Model

The Large Language Model (LLM) is a text processing model designed to understand and generate human language (Devlin et al., 2019). Benefiting from billions of learnable parameters and vast amounts of text data, LLM can understand complex patterns in linguistic data and perform a wide range of tasks, including text summarization, translation, sentiment analysis, and more (Naveed et al., 2023).

Building on the success of LLMs, many studies have begun to explore how LLMs can jointly process visual and text information. In 2023, OpenAI released GPT-4V (Achiam et al., 2023), which incorporated image inputs into LLM to build the first Visual Language Model (VLM). Subsequently, LLaVA (Liu et al., 2023b), a series of visual language models were successively proposed. They align 2D images to the language model through image encoders and projection layers. Other models such as BLIP2 (Li et al., 2023), Qwen (Bai et al., 2023) use a more complex QFormer architecture to guide the compression of visual features using textual cues. These models build massive multimodal linguistic image datasets to fine-tune instructions to existing LLM models by means of GPT-4 generation or manual collection. Numerous studies have shown that VLM models are capable of constructing worldviews from massive amounts of data, which permits the models to emulate the human mindset for visual understanding. As a result, VLM has shown excellent performance in segmentation, classification, and caption of images (Zhang et al., 2024).

However, VLMs performs poorly in some spatial comprehension-based tasks (Cheng et al., 2024). Due to the inherent differences between modalities, it is difficult for VLM to capture accurate spatial information directly from images, such as determining the absolute distance between objects. In addition, in order to align the image modalities with the text modalities, most of the existing methods choose the CLIP (Radford et al., 2021) as the encoder of the image. However, the globally aligned training method of CLIP, limits the fine-grained semantic understanding ability of VLMs (Kaul et al., 2024). To address this problem, we propose a Depth Projector Module and a Magnifying Glass Strategy to adjust the VLMs to focus on the semantic and geometric information of the object for effective measurement.

3. Method

Our goal is to unleash the reasoning capabilities of VLMs to enable efficient image-based tree measurements. Fig.1 shows the overall process of VLM-TMN. Specifically, it takes an image and the corresponding bounding box of the measuring tree as input, the bounding box can come from the output of any object detection network. Our network consists of three parts: Visual Encoder, Depth Projector, and Text Encoder. These modules map the input to the unified text feature space. Then, we use Llava-7B as the backbone to process these multimodal features and output the morphological parameters of each tree.

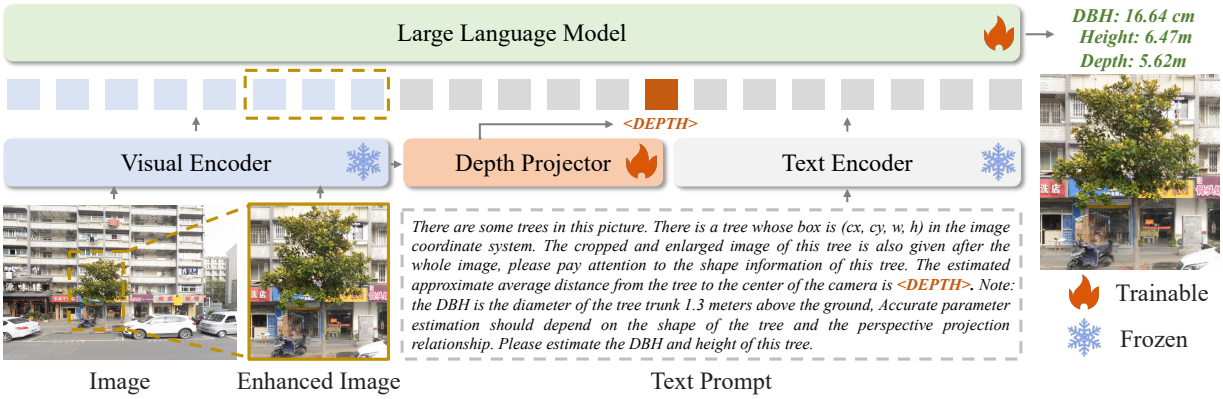

Figure 1. Overview of VLM-TMN architecture.

3.1 Visual Encoder

Native VLMs usually use the pre-trained CLIP (Radford et al., 2021) model as the visual encoder to extract image features. CLIP can map text and visual features to the same space, but the fine-grained visual semantics of CLIP are poorly expressed, making it difficult for VLMs to focus on the key areas of the image from text input, thus causing illusions. To solve this problem, we propose a magnifying glass strategy. First, we use the same image encoder to encode the image feature $\boldsymbol{F}_v$, it contains contextual cues to the overall scene. Then, to force the VLM to focus on the measuring object, we apply explicit cropping. Specifically, given a bounding box, we first expand the bounding box range by 10 pixels as the cropping range, and then crop and enlarge this area to the original image size to obtain an enhanced image. After that, we input the enhanced image into the same visual encoder to get the visual enhancement features $\boldsymbol{F}_e$. To reduce computing consumption, we compress the $\boldsymbol{F}_e$ into 16 tokens and concatenate them with the $\boldsymbol{F}_v$ to perform visual feature extraction. This strategy allows us to improve the performance of the model without introducing any learnable parameters.

3.2 Depth Projector

As discussed before, accurate spatial perception is critical to 3D measurement tasks. However, the original VLMs lack depth guidance during the training process, which limits the measurement performance of the model. To solve this problem, a natural idea is to introduce a depth estimation model and use the estimated depth map as an additional input. However, additional input requires the design of additional depth feature extraction and feature fusion modules to ensure that the depth features can be mapped to the text modality. In contrast, we introduce a novel Depth Projector Module that directly extracts spatial information from image features. Specifically, we set $\boldsymbol{F}_e$ as query vectors, then use a Cross-Attention Block to search implicit depth clues from $\boldsymbol{F}_v$. After that, we map these depth clues into text feature space through an MLP layer to get $\boldsymbol{F}_d$. The overall process is shown in Eq.1.

$$\boldsymbol{F}_d = \mathrm{MLP}(\mathrm{CrossAttn}(\boldsymbol{F}_e, \boldsymbol{F}_v))) \tag{1}$$

Since $\boldsymbol{F}_v$ and $\boldsymbol{F}_e$ are aligned with the text features, it is easy to map the implicit depth features $\boldsymbol{F}_d$ to the text modality. In addition, to ensure the accuracy of depth clues, we also introduced explicit depth supervision to force $\boldsymbol{F}_d$ to predict the corresponding distance from each tree to the camera center.

3.3 Text Encoder

Research has shown that an appropriate prompt can guide the thought process of VLMs. So, we designed a prompt to describe the tree measurement task. First, we use the box to describe the measuring tree and inform the model about the concepts of DBH and TH (i.e., *Note: the DBH is the diameter of the tree trunk 1.3 meters above the ground.*). In addition, to help the model pay attention to the visual and geometric information, we set *"The cropped and enlarged image of this tree is also given after the whole image, please pay attention to the shape information of this tree."* in the prompt. We also embed the $\boldsymbol{F}_d$ as the word $\langle DEPTH \rangle$, and replace it into the text input, which further facilitates LLM's understanding of spatial distance information. After that, we input these prompts into the pre-trained CLIP text encoder to get $\boldsymbol{F}_t$. Finally, we connect the $\boldsymbol{F}_v, \boldsymbol{F}_e, \boldsymbol{F}_d, \boldsymbol{F}_t$ together and input it into the backbone to measure each tree.

3.4 Loss Function

The loss function of the whole model consists of two parts. First, to ensure the validity of the Depth Project Module, we explicitly supervise the Depth Token as shown in Eq 2.

$$L_1 = \|d - \hat{d}\| \tag{2}$$

Where, d is the depth from the camera to this tree decoded by the depth token and $\hat{d}$ is the corresponding ground truth. Then, we also supervised the estimated morphology parameters, as shown in Eq 3.

$$L_2 = \|dbh - \hat{dbh}\| + \|th - \hat{th}\| \tag{3}$$

The complete loss function can be expressed in Eq 4. α is the weight factor, which we set to 0.1 based on experience.

$$L = \alpha L_1 + L_2 \tag{4}$$

4. Experiment

4.1 Dataset

To evaluate our method, we constructed a benchmark dataset based on the WHU-RSTree dataset collected by Dong et al. (Zhen Dong, 2023). It is a multimodal urban tree instance segmentation dataset comprising approximately 68 km

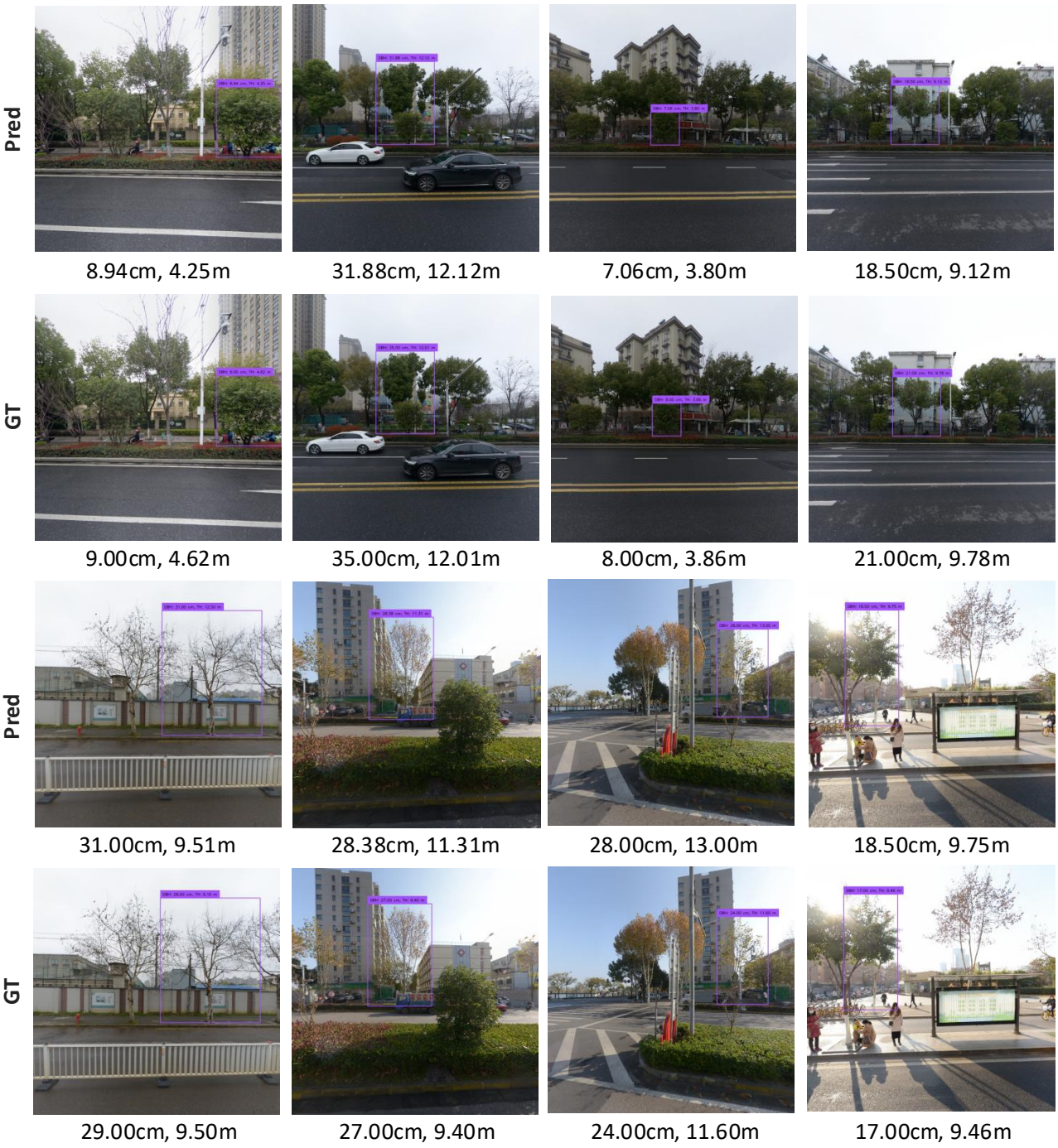

Figure 2. The quantitative results of VLM-TMN. For ease of reading, we marked the tree morphology computation results in the order of ”DBH, TH” at the bottom of the image.

Table 1. Results on benchmark dataset. Quantitative comparisons with existing image-based measurement methods, ↓ denotes this metric lower is better. The best results are in bold.

Method	Pipeline	RMSE ↓		NRMSE ↓		Time Cost ↓
		DBH	Height	DBH	Height	
Wang et al.	Manual interpretation	**1.9 cm**	**0.5 m**	/	/	0.67 h
Choi et al.	Detection + Segmentation	15 cm	3.28 m	0.44	0.24	/
Ours	Detection + VLM-TMN	7.1 cm	2.3 m	**0.23**	**0.17**	**0.27 s**

Table 2. Quantitative comparisons results, ↓ denotes this metric lower is better. The best results are in bold.

	Time Cost ↓	RMSE ↓		RRMSE ↓		REL ↓	
		DBH	Height	DBH	Height	DBH	Height
W/o All	**0.24s**	24.39cm	9.37m	0.80	0.73	1.12	0.82
W/o Depth & Visual	**0.24s**	7.66cm	2.41m	0.25	0.19	0.22	0.17
W/o Depth	0.26s	7.47cm	2.37m	0.24	0.18	0.22	0.18
Whole Pipeline	0.27s	**7.08cm**	**2.25m**	**0.23**	**0.17**	**0.21**	**0.16**

of point cloud data and 12,447 panoramic images collected from Nanjing, China. It includes annotations for more than 20,000 individual trees, covering instance segmentation, tree species, and morphological parameters. After our filtering and processing, we constructed a total of 19,613 datasets containing images, object detection boxes, and corresponding tree morphological parameters.

4.2 Implementation details

First, we split the dataset into training, validation, and test sets in a ratio of 8:1:1. Our model was trained on 1 NVIDIA RTX 4090 GPU using the PyTorch platform. The training batch size was 8. Each image we cropped to 224 x 224 as input, the initial learning rate was set to 0.0001, and a total of 100,000 iterations were performed to complete the training.

4.3 Evaluation metrics

As a measurement task, we use root mean squared error (**RMSE**), normalized root mean squared error (**NRMSE**), and mean relative error (**REL**) as evaluation metrics.

4.4 Results on benchmark dataset

We only used existing Image-based methods for comparison. Since none of the comparison methods have open source code, we directly report the metrics in their papers. The results are shown in Tab 1. As can be seen from the table, compared with the method proposed by Wang et al. (Wang et al., 2018), which relies on manual interpretation, VLM-TMN has higher efficiency. Compared with the automated method proposed by Choi et al. (Choi et al., 2022), VLM-TMN achieves better measurement accuracy due to its consideration of both visual and depth information of objects.

It is also worth mentioning that the comparison methods all choose ideal unobstructed photos for index evaluation, and the dataset we evaluate is derived from real street-acquired images with a large number of occlusions. Fig 2 illustrates the qualitative results of our method. In addition to ideal images, VLM-TMN achieves robust estimation in challenging scenarios, such as distant, small, or partially occluded trees. Overall, our method is more robust and accurate than existing methods.

4.5 Ablation study

We also conducted ablation experiments as shown in Tab 2. Specifically, we removed the depth projector module, the visual magnifying glass strategy, and the supervised fine-tuning, respectively. When all modules are removed, the model will degenerate into the original Llava-7B model. We can see from Tab.1, our design allows VLMs to have the 3D measurement capabilities. In addition, our proposed magnifying strategy and depth projector module effectively improve the performance (+ %7.57 RMSE) compared to simply fine-tuning the VLMs, and removing any components will cause performance degradation.

5. Conclusion and Future Work

In this paper, we propose VLM-TMN, unlike primitive measurements, our approach estimates roadside tree morphology parameters quickly and accurately from low-cost images by unleashing the inference potential of VLMs. To overcome the lack of spatial comprehension and limited fine-grained perception of the original model, we propose two novel modules. First, we design the magnifying glass strategy to force the network to focus on the fine-grained details of the objects through simple cropping and resize operations. Second, we introduce the Depth Projector Module, which gives the model spatial perception capability by means of implicit depth coding and explicit distance supervision. Experiments on benchmark datasets show that compared to using Supervised Fine-Tuning (SFT) for VLMs, our design brings +7.57% improvements in RMSE, and makes VLM-TMN achieve DBH with NRMSE of 0.23 and TH of 0.17, which significantly outperforms existing methods (0.44 and 0.24, respectively).

There are also some limitations in our work, for example, the performance may be degraded when facing some unseen scenarios. The emergence of Deepseek-R1 (Guo et al., 2025) gives us a new solution strategy. In our future work, we plan to use reinforcement learning and chain of thought to mine the worldview of VLMs, so that the model can think for itself according to different scenarios, and improve the model utility and generalization ability.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2023YFF0725200.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al., 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F. et al., 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Campbell, M. J., Eastburn, J. F., Mistick, K. A., Smith, A. M., Stovall, A. E., 2023. Mapping individual tree and plot-level biomass using airborne and mobile lidar in piñon-juniper woodlands. *International Journal of Applied Earth Observation and Geoinformation*, 118, 103232.

Cheng, A.-C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S., 2024. SpatialRGPT: Grounded Spatial Reasoning in Vision Language Models. *arXiv preprint arXiv:2406.01584*.

Choi, K., Lim, W., Chang, B., Jeong, J., Kim, I., Park, C.-R., Ko, D. W., 2022. An automatic approach for tree species detection and profile estimation of urban street trees using deep learning and Google street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190, 165–180.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Estornell, J., Hadas, E., Martí, J., López-Cortés, I., 2021. Tree extraction and estimation of walnut structure parameters using airborne LiDAR data. *International Journal of Applied Earth Observation and Geoinformation*, 96, 102273.

Giannetti, F., Puletti, N., Quatrini, V., Travaglini, D., Bottalico, F., Corona, P., Chirici, G., 2018. Integrating terrestrial and airborne laser scanning for the assessment of single-tree attributes in Mediterranean forest stands. *European Journal of Remote Sensing*, 51(1), 795–807.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X. et al., 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, Y., Wang, H., Yan, H., Han, Q., Nan, X., Zhao, K., Bao, Z., 2023. Alternative scenarios for urban tree surveys: Investigating the species, structures, and diversities of street trees using street view imagery. *Science of The Total Environment*, 895, 165157.

Kaul, P., Li, Z., Yang, H., Dukler, Y., Swaminathan, A., Taylor, C., Soatto, S., 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27228–27238.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International conference on machine learning*, PMLR, 19730–19742.

Liu, D., Jiang, Y., Wang, R., Lu, Y., 2023a. Establishing a citywide street tree inventory with street view images and computer vision techniques. *Computers, Environment and Urban Systems*, 100, 101924.

Liu, H., Li, C., Wu, Q., Lee, Y. J., 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892–34916.

Long, C., Zhang, W., Chen, Z., Wang, H., Liu, Y., Tong, P., Cao, Z., Dong, Z., Yang, B., 2024. SparseDC: Depth Completion from sparse and non-uniform inputs. *Information Fusion*, 110, 102470.

Mielcarek, M., Kamińska, A., Stereńczak, K., 2020. Digital aerial photogrammetry (DAP) and airborne laser scanning (ALS) as sources of information about tree height: Comparisons of the accuracy of remote sensing methods for tree height estimation. *Remote Sensing*, 12(11), 1808.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A., 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Pierzchała, M., Giguère, P., Astrup, R., 2018. Mapping forests using an unmanned ground vehicle with 3D LiDAR and graph-SLAM. *Computers and Electronics in Agriculture*, 145, 217–225.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *International conference on machine learning*, PmLR, 8748–8763.

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T., Eliceiri, K. W., 2017. ImageJ2: ImageJ for the next generation of scientific image data. *BMC bioinformatics*, 18, 1–26.

Schneider, C. A., Rasband, W. S., Eliceiri, K. W., 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature methods*, 9(7), 671–675.

Villasante, A., Fernandez, C., 2014. Measurement errors in the use of smartphones as low-cost forestry hypsometers. *Silva Fennica*, 48(5).

Wang, W., Xiao, L., Zhang, J., Yang, Y., Tian, P., Wang, H., He, X., 2018. Potential of Internet street-view images for measuring tree sizes in roadside forests. *Urban Forestry & Urban Greening*, 35, 211–220.

West, P., 2009. Tree and forest measurement.

Wu, F., Wu, B., Zhao, D., 2023. Real-time measurement of individual tree structure parameters based on augmented reality in an urban environment. *Ecological Informatics*, 77, 102207.

Yang, B., Dong, Z., Liang, F., Mi, X., 2024. *Ubiquitous Point Cloud: Theory, Model, and Applications*. CRC Press.

Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J., 2023. An Improved Baseline for Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2312.17240*.

Yang, Z., Liu, Q., Luo, P., Ye, Q., Duan, G., Sharma, R. P., Zhang, H., Wang, G., Fu, L., 2020. Prediction of individual tree diameter and height to crown base using nonlinear simultaneous regression and airborne LiDAR data. *Remote Sensing*, 12(14), 2238.

Zhang, J., Huang, J., Jin, S., Lu, S., 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhen Dong, B. Y., 2023. Whu-usi3dv. `https://github.com/WHU-USI3DV`.