

Leveraging Segment-Anything model for automated zero-shot road width extraction from aerial imagery

Nan Xu¹, Kerry Nice², Sachith Seneviratne^{2,3}, Mark Stevenson^{1,2}

¹*Department of Infrastructure Engineering, Faculty of Engineering and IT*

²*Transport, Health and Urban Systems Research Lab, Melbourne School of Design*

³*Optimization and Pattern Recognition Group, Faculty of Engineering and IT*

The University of Melbourne

Parkville, Victoria, Australia

naxu@student.unimelb.edu.au, {kerry.nice, sachith.seneviratne, mark.stevenson}@unimelb.edu.au

Abstract—Segment-Anything model (SAM) is a foundation segmentation model published in April 2023. Trained on an unprecedented 11 million annotated images, the model can generate segmented masks bearing clear-cut contours by integrating user-provided prompts. It is zero-shot transferable, requiring no task-specific training. Nevertheless, its applicability for geographic vision tasks has not been fully evaluated. There is no automated prompt-feeding method incorporating with SAM that can work efficiently for purposeful batch processing as well. To fill these gaps, we developed a process that can be executed automatically from visual-prompts extraction to road width measurement, utilizing OpenStreetMap (OSM) and SAM. By examining the quality of segmentation in various image contexts, we evaluated the capacity and limitations of SAM working on aerial imagery. Through comparing measured widths to VicRoads records, we validated the specially designed width-measuring algorithms for high precision and accuracy. After this process, prompt-indicated zero-shot approach in solving basic geographic vision tasks is to be shaped synchronously on both theory and application ends.

Index Terms—prompt, Segment Anything, zero-shot, without training, OpenStreetMap, road extraction, road width, remote sensing, aerial imagery

I. INTRODUCTION

The Coupling between remote sensing and vision-based technology has been long established. It is consolidated with the wide use of deep learning (DL). DL processes large-scale data using deep neural networks to extract substantial Earth information from remote sensing images. This information, in turn, plays an indispensable role in finding better solutions to geographic problems in any spatial or temporal coverage. [1]

As high-resolution remote sensing images and advanced DL models become more accessible, many previously challenging tasks become less difficult. In the light of a newly published pre-trained DL model Segment-Anything, or SAM, this paper attempts to re-evaluate a long-standing geographic task: road extraction from remote sensing images and find an efficient method for accurate width measurement based on road shapes extracted.

SAM [2] is a prompt-indicated segmentation foundation model which is trained on a SA-1B dataset composed of 1 billion masks and 11 million images, the largest of its kind to date. By being given simple prompts, e.g., points, boxes etc., the model can generate segmented objects with

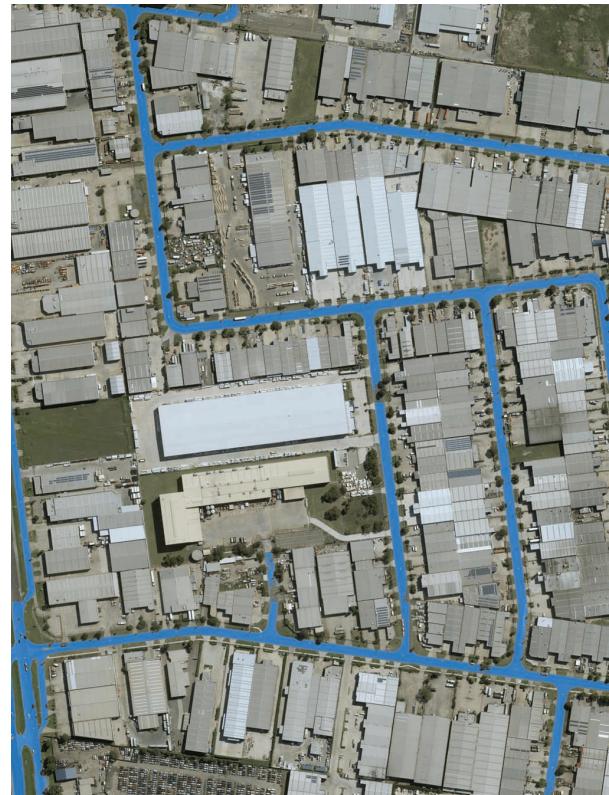


Fig. 1: Sample result from automated-SAM segmentation on an area of $\sim 702,768m^2$ in Melbourne, Australia

clear boundaries. Prompt-approach is seldom found in the vision-based artificial intelligence (AI) field, even if it was largely used and achieved fair success in other sub-fields like natural language processing. SAM is claimed as zero-shot applicable on new datasets or tasks. No specific training is required. Considering the scale of data required to train SAM, the investment of time and resources maybe unaffordable for many AI researchers. Transferability is a compelling method to provide an alternative resource efficient approach to quick

use, limited scope extension, or modification of foundation models to achieve comparable results to training. However, this is impossible without a better understanding of the model's inherent capacity and limitations. An assessment of SAM is therefore required.

Road extraction is a typical geographic task. Despite numerous studies that have been undertaken in the last decade, this task remains non-trivial in terms of the difficulty in finding a simple, coherent, robust, and transferable method. The task, though challenging, is fundamental for many extended applications or topics, such as smart cities, digital twin, autonomous driving, road safety, healthy and sustainable transport etc. The task is also basic for other tasks that need to rely on its output, for example, road width measurement can only be possible on clear road shapes extracted. A re-evaluation of road extraction using SAM will support these advanced usages.

In this research, an automated process integrating OSM with SAM is used to extract road shapes from aerial images while SAM is assessed to understand its capacity and transferability. Fig. 1 displays the qualitative result automatically generated by the process for a 16384x12288pixels image. Road widths are subsequently measured using an efficient algorithm for accurate width values.

This paper will contribute to the following areas:

- Evaluation of the applicability of prompt-indicated segmentation foundation model without task-specific training
- Evaluation of zero-shot transferability in terms of prompt effectiveness, image context and their relationship
- Development of a simple, cost-efficient, explainable, transferable method for road width measurement

II. RELATED WORK

SAM was introduced only recently; hence online papers are mostly not peer-reviewed, and few can be found related to remote sensing except the following listed. The favourable boundary accuracy of SAM encouraged the generation of large, labelled datasets for training involved tasks [3]. Broadly evaluated real-world applications in [4] indicate substantial limitations of SAM but it is not clear whether the errors are derived from prompts or the model itself. SAM was found incapable in road segmentation regardless of prompt type in [5] and generally inappropriate for overhead imagery. However, it displayed encouraging performance in synthetic aperture radar (SAR) imagery for glaciology that may benefit climate or Earth science [6]. Tuned SAM was used in [7] to detect landforms on Mars. Adapter for prompt fine-tuning was applied in [8] for shadow or camouflage object detection. These papers embody a good start for SAM evaluation, especially the road segmentation assessed in [4] and [5], but considerable work is still needed towards comprehensive understanding.

Road extraction as a basic graphic task has been widely studied. Width measurement, however, has not been sufficiently examined. Here width refers to the shorter side-to-side distance of a two-dimensional road shape which needs to be extracted first from image. A survey [9] listed extensive previous work on road extraction from 2D or 3D remote

sensing data in the last decade. In the 2D field, applied semantic segmentation (SS) methods can be classified into 3 types: 1) morphological feature-based processing including opening, closing, derivative map etc. 2) machine learning methods such as histogram of oriented gradient, scale-invariant features, support vector machine etc. 3) deep learning models mostly based on FCN, CNN, U-NET or GAN. All methods are tailored to perform road extraction for a fixed dataset. Refinement or error-reducing techniques are required for post-processing since output masks normally contain noise or outline errors due to SS involved pixel-level prediction and classification. Consequently, over-complexity is induced in width measurement [10] [11] [12] [13]. SAM conversely has only mask-level prediction [14], which can generate masks with clear contours. No post-processing is needed. Given the manifest geometry of road, the algorithm of width measurement can be much simplified.

In the following content, the whole process is generally described in Section III-A and details are elaborated in Section III-B-III-G to ensure replicability. Results of road extraction are discussed in relation to image context (size, content) in Section IV, in which Section IV-A is an overview of the results showing the model's capacity and limitations; Section IV-B is about the correlation between image size or content complexity and prompt effectiveness. Width measurements are listed in Section IV-C where measured values are compared with government published values, if available, to check the correctness of the whole process. Conclusion and future work are discussed in Section V.

III. METHOD

A. General description

The process is designed based on two main questions: a) How does the process generate prompts automatically and feed them to SAM? b) How does it measure width of road extracted from aerial imagery? As illustrated in Fig. 2, steps 1 through 4 is the solution to question a, while steps 5 and 6 give the answer to question b. OpenStreetMap (OSM) [15] is integrated in the process at the beginning to produce prompts for SAM and assists in width measurement. OSM is vital in arriving at correct width results while SAM is responsible for the quality

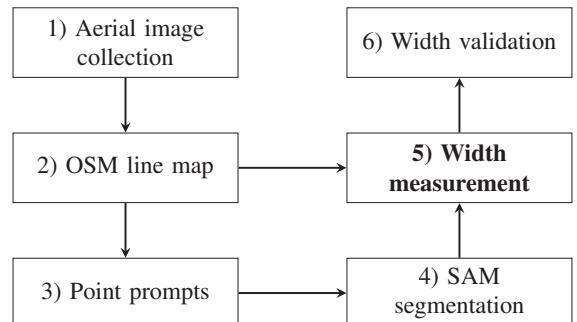


Fig. 2: Automated process for road width measurement

of road extraction. Yes/No steps could be added for scalable workflow to a larger dataset, but it is not the focus of this research. An example will be used to explain each step in detail in the following sections.

B. Aerial image collection

Aerial imagery of Melbourne are provided by MetroMap, a surveying services provider in Australia. The 5cm resolution imagery is captured by MetroCam, a self-developed high-flying camera system, in January 2020. Initially collected tiles (256x256) are concatenated into images (4096x4096) as in Fig. 4.

C. OSM line map

The bounding box of the geolocated image (Fig. 4) is used to capture road (line) map of interest from OSM as Fig. 5, which is a part of Melbourne’s drive map, captured by OSMnx [16] “graph from place” specifying network type as “drive” for all drivable roads in Melbourne as shown in Fig. 3.

D. Point prompts

The line map in Fig. 5 indicates the location of the road as indicated in OSM in the corresponding aerial image in Fig. 4 and can serve the purpose of prompts. Therefore, points can be selected from that line to form point prompts as input to SAM.

Points can be randomly or purposely selected. According to some initial experimental results, we propose to divide the square space of 4096x4096 image into 9 smaller almost-equal-sized blocks by adding virtual lines vertically and horizontally. The bounding boxes of formed blocks are the combinations of intervals ([0,1365], (1365,2731), [2731,4096]) on either dimension. By selecting the points located at the median position in each block, we can get maximum 9 points for the whole image. The purpose is to ensure evenly distributed efficient prompts while avoiding boundary positions.

The line in Fig. 5 is expected to cross 3 virtual blocks in the 1st column, so that 3 points will be picked as its point prompts.

E. SAM segmentation

Points attained from the last step are automatically fed into SAM as positive prompts to segment the desired object. SAM is set to select the best mask as the output from initially generated 3 masks for ambiguity tolerance. [2] The ground-truth level segmentation as shown in Fig. 6 indicates the effectiveness of the selected point prompts. Its clear-cut boundary can facilitate width measurement.

F. Width measurement

The segmented road mask from step 4 (III-E) is blended with the line map from step 2 (III-C) via interpolation alpha factor 0.8 in Fig. 7. Partial misalignment can be found in the blended image where OSM line is supposed to be the centreline of the road. The blended image is used to measure widths through the following sub-steps:

1) Finding the fastest path: If the OSM white line in Fig. 7 is written as an array, all positions of white pixels can be sorted and listed as [[0, 226]... [1, 228]... [2, 230] ...]. To get the fastest path the line resides, the array can be simplified by extracting the mid-point pixel from each column, as indicated in orange in Fig. 9. Mid-point pixels are calculated using (1), and the connection of them constitutes the fastest path.

$$stp + ceil((shp - stp)/2) \quad (1)$$

This approach ensures no minor direction change of the road will be missed in any fine increment analysis while computing can remain efficient by processing only about one fifth of the original pixels in the later calculations.

2) Adding perpendicular lines for width pixels: The OSM line was simplified into a list of discrete pixels in the last step and a vector formed by any two consecutive pixels from that list can indicate the position and orientation of a small portion of the road. V1, consisting of the first and second mid-point pixels, [0, 228] and [1, 230], represents the orientation of the road at point [1,230] with the length between points, [0, 228] and [1, 230] (Fig. 9, upper-right). The vector V2 that is perpendicular to V1 can be found through (2).

$$V1 \cdot V2 = 0 \quad (2)$$

The line extended from vector V2 represents the width of the road at point [1,230], which is drawn as zero-width red line in the blended image in Fig. 8. The number of pixels the red line crosses the road area (shadowed) is counted as 73 in this image as width pixels.

3) Converting width pixels to meters: We can convert width pixels to meters in the same ratio as image width to longitudinal coverage or height to latitudinal covered distance as (3).

$$\frac{\text{image width(height)}}{\text{lon(lat) coverage}} = \frac{\text{width pixels}}{\text{road width}} \quad (3)$$

The longitudinal and latitudinal coverages for the sample image are calculated using haversine formula to be 242.336 and 242.332 meters respectively. Therefore, 242 is considered safe to use as the conversion factor at any orientation in that image. The 73 width pixels is then converted to 5.6 meters via (3).

Measured widths are local at certain point with respect to its local orientation and subject to change with the location of that point. Therefore, the overall width needs to be determined by the most frequent value in multiple measurements. By selecting points from (1) at an interval, e.g., 50, a list of widths can be measured as follows,

[5.6, 9.6, 10.7, 10.7, 10.2, 10.7, 10.7, 11.0, 10.8, 10.6]

The most frequent value, 10.7m, is ultimately chosen as the final estimate for the width of the road in Fig. 4. The smaller interval set; the more measurements will be generated for a finer-grained width analysis.



Fig. 3: Melbourne's road network of drivable roads

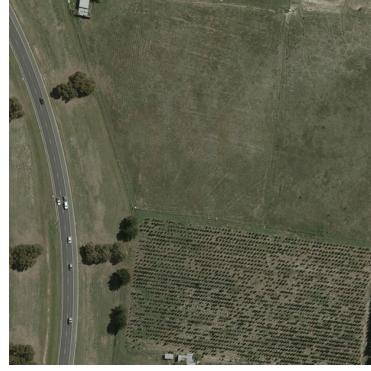


Fig. 4: Aerial image sample



Fig. 5: Cropped line map from OSM corresponding to the scene in Fig. 4



Fig. 6: SAM segmented road is masked in blue and prompt points are annotated as green stars

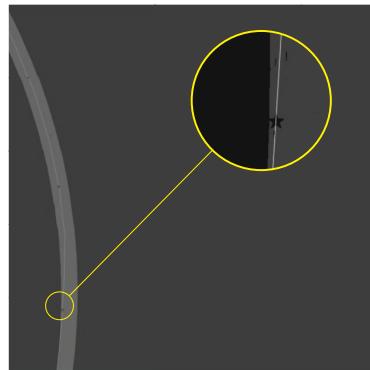


Fig. 7: Blended image of segmented mask and OSM line map (misalignment detected)

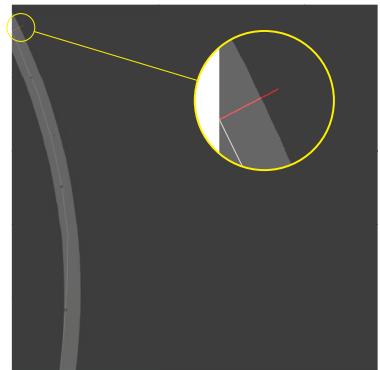


Fig. 8: Added perpendicular line to road orientation in red

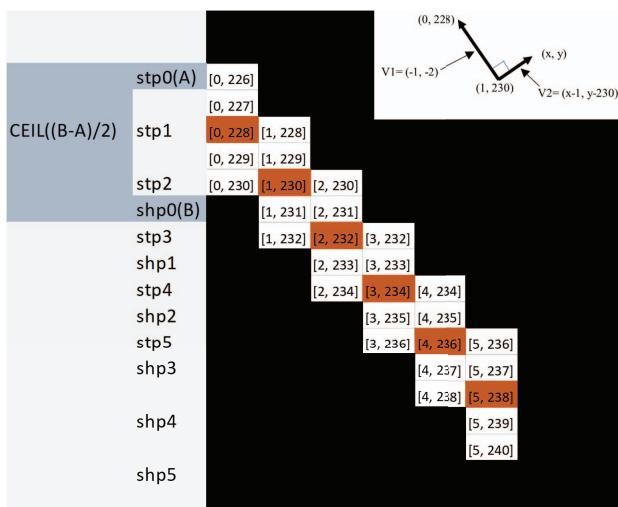


Fig. 9: Finding the fastest path that OSM line resides (white pixels are on OSM line, orange pixels are selected; stp: starting position; shp: shift position)

G. Width validation

“Road width and number of lanes” [17] is a dataset attained from VicRoads incorporated with the Department of Transport and Planning in the State government of Victoria, Australia. We capture each width record from this dataset using the same bounding box mentioned in III-C as a reference value. However, VicRoads dataset is not map-structured containing all geographic locations. Bounding box approach can fail when the road of interest was recorded at a location outside that box and width cannot be directly read. The sample image fell into this category. But if the road name “Hume” was queried in the VicRoads csv file, 2083 entries called “Hume Highway” can be found with width values around 11-12m, not far away from the measured value of 10.7m.

In the next section, width comparison will be conducted when reference values can be directly captured from the VicRoads dataset using bounding boxes. 47 out of 139 processed roads are compared to their reference values.

IV. RESULTS AND DISCUSSION

In the initial small batch test, we found that SAM works best in the single-road-no-intersection scenario, the error rate

increases with the complexity of roads in the image, such as higher number of roads and intersections, or higher variety in the classification of roads. However, limiting this complexity is not considered as an issue since image size for analysis can be customized. Therefore, we think the success rate on less-complex images, e.g., single-road-no-intersection, two-roads-one-intersection, etc., is a better indicator to reflect the capacity of SAM in road extraction, and it is relied on for complex images to expect good results when simpler sub-images are processed separately.

A. SAM segmentation results

Out of 2025 mainly single-road-no-intersection images (4096x4096), SAM generated 1873 ground-truth-level results through maximum 9 point-prompts. Sample results in Fig. 10 show robustness in various image contexts, compact building surroundings, or heavy forest etc. The model also made 152 wrong segmentations. Three types of reasons are identified for failures: errors in OSM, occlusion and SAM failures. Corresponding quantities and percentages are summarized in TABLE I. Wrong segmentation exclusively caused by SAM is less than 1% in the total batch. OSM error and occlusion occupied 7%, which is not higher compared with other 2D remote sensing published papers. Details are explained in Section IV-A1–IV-A3.

TABLE I: AERIAL IMAGERY SEGMENTATION

| SAM Segmentation results | | | |
|-------------------------------------|-----------|-----------|-------------|
| Total number | Good | Bad | |
| 2025 | 1873 | 152 | |
| Classified reasons for bad results: | OSM error | occlusion | SAM failure |
| | 92 | 54 | 6 |
| | 60.5% | 35.5% | 4% |

1) *OSM error*: The first possible reason is that OSM or aerial images are out of date. As in Fig. 11, there is no road shown in original image (OSM err-1o), but OSM indicates a road and results in the wrong segmentation (OSM err-1s). The second possible reason is road centerlines may have been misaligned in OSM. It can be partially misaligned as in Fig. 7 where the 3rd point is almost out in the enlarged view, or totally offset from the actual roads as in Fig. 11 (OSM err-2s).

2) *Occlusion*: Vertical occlusions can confuse SAM when prompts are placed on them. It is inevitable in 2D images when large greenery covers roads or there is vertical construction built above roads, e.g., trees and bridges are wrongly segmented in Fig. 11 (Occ-1s) and (Occ-2s) respectively, even though OSM line maps are correct.

3) *SAM failure*: Failure cases exclusively caused by SAM deserve more attention. Firstly, when roads look similar or identical to its adjacent objects, SAM showed difficulty in segmenting them. 4 out of 6 belong to this situation as sampled in Fig. 11 (SAM err-1o) in which 3 original images containing hardly visible but still recognizable roads by human

eyes cannot work in SAM. This may indicate segmentation is realized based on image morphological features, not the class of the object the model understands through given prompts. Secondly, the model tends to predict larger area as positive than ground-truth when it does not work, and no clear reason can be found as 2 pairs of sources (SAM err-2o) and results (SAM err-2s) in Fig. 11.

Another source of error is from the point prompts exactly located on shadows, which are excluded in most correctly segmented results. Pointing to shadows can confuse the model to generate wrong prediction which can be equally disagreed by successful cases.

In more complex content images, the segmentation success rate dropped significantly since the errors discussed above can be accumulated. Nevertheless, promising results are not uncommon as in Fig. 12. Stable good results can be expected when complex images are considered as compositions of separately processed simpler ones.

B. Image-size effect

Image size, representing the complexity of roads in one image, plays a role in the success rate of SAM segmentation. To verify how it works, we tested one failed image (4096x4096) in its half (2048x2048), 1/4 (1024x1024) and 1/16 (256x256) size. The result is improved gradually with the reducing size and complexity of images as in Fig. 13. Some images can get decent segmentation after one-time size reducing as the top two samples in Fig. 14 while 50% of the bottom two samples needs further size-reducing. A decision is to be made on whether a new iteration is required. Here may exist a trade-off between the speed of processing and the stability of results.

When the image size was decreased to 256x256, one new phenomenon is found as in Fig. 15. The prompts help to generate a narrow white symbol line on the street. This indicates the model is good at finding differentiable edges even on extremely narrow or small objects once the prompts are all correctly placed. But on the other hand, the model has little understanding of the meaning of prompts.

C. Width measurement

Width is measured in the sampled 139 images out of the 1873 successfully segmented images, in which 47 can find their bounding-box-captured VicRoads records leading to a one-to-one comparison shown as curves in red and blue respectively in Fig. 16. The loess-smoothed curve in green is largely aligned with the reference curve. As discussed in the Section II, width measurement is not as actively studied as road extraction in recent years. The comparison with previous work is only conducted on available results as listed in TABLE II. The 47-image accumulated pixels in our work cover a larger geographic area. The higher error rate in 5-10m range can be caused by the limited number of sample points; only 12 out of 47 fall within this range and any large deviation in a small group can cause significant general error rate. The 10-20m range, however, performs much better.



Fig. 10: 1873 images with ground-truth-level segmentation

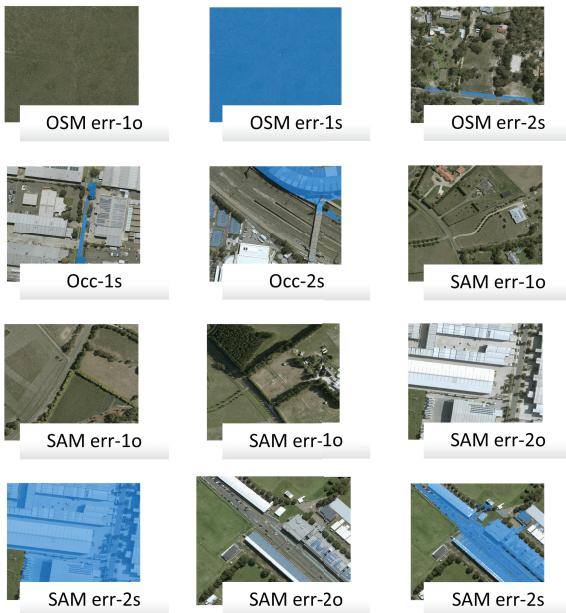


Fig. 11: 152 images with bad segmentation due to classified reasons (image title: error type -number of subtype (“o” for original, “s” for segmented))

TABLE II: COMPARISON WITH PREVIOUS WORK

| Width range (m) | Image Size (Accu.) , Resolution | | Error Rate | |
|--------------------------|---------------------------------|-------|------------|-------|
| | 5-10 | 10-20 | | |
| Guan (2010) [13] | | | 29.3% | 27.3% |
| Xia (2017) [12] | | | 32.2% | 39.4% |
| Luo (2018) [11] | | | 7.5% | 54.5% |
| Ours (2023) ^a | 192512x192512pixels, 5cm | | 36.8% | 14.6% |

^aError rates are calculated from median values.

V. CONCLUSION

SAM model has the ability for direct inference without training in road extraction application when suitable image size (content complexity) is chosen, and accurate prompts are given. Successful results are mostly ground-truth level with no pre, or post processing required. Therefore, downstream tasks like width measurement can be conducted efficiently to make reasonable estimates on real road width using a set of explicit and easy-replicable algorithms. Comparing with the previous training-processing-led methods, this workflow produced comparable results in the 5-10m width category and much improved ones on 10-20m.

Although SAM showed good capacity on contour processing, it encountered difficulties handling complex or visually misleading images, which tells its capacity may rely more on morphological features instead of the understanding on



Fig. 12: Good segmentation results for complex content

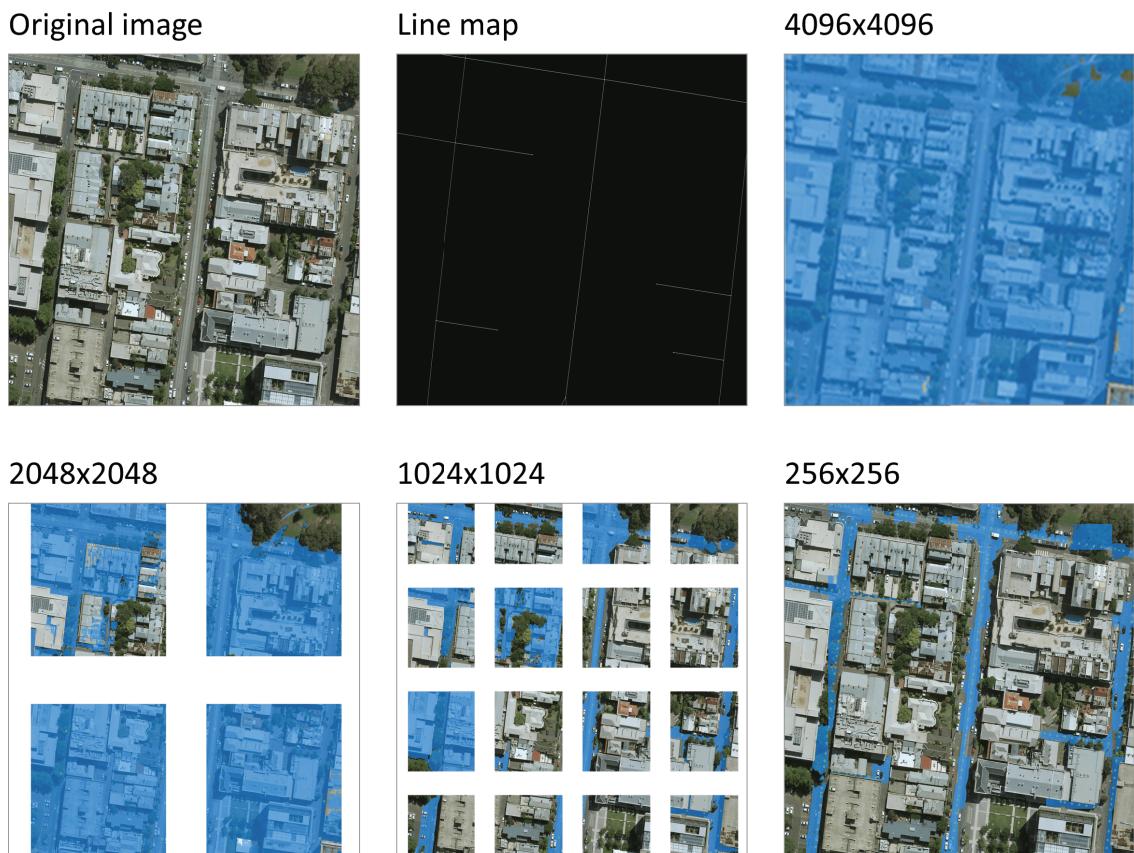


Fig. 13: Segmentation results affected by various image sizes

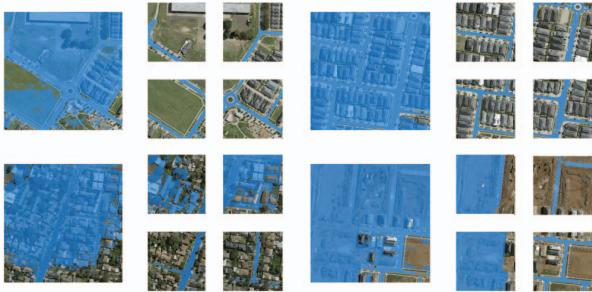


Fig. 14: 100% (top two) and 50% (bottom two) segmentation success due to one-time size-reducing



Fig. 15: 256x256 image with a white symbol line segmented

the object it processed. When conflicted prompts are given, it tends to predict all confusing areas as positive resulting in many over-masked results. Some changes on prompt layer for the model to better understand the meaning of prompts may prepare it to perform greater in any task.

On the other hand, small image size with less complex objects contained is a good choice not only for segmentation, as proved correlation between SAM performance and image complexity, but also for width measurement in terms of easier generalization of algorithm from one width to another. In addition, the use of 3D data could enable the removal of

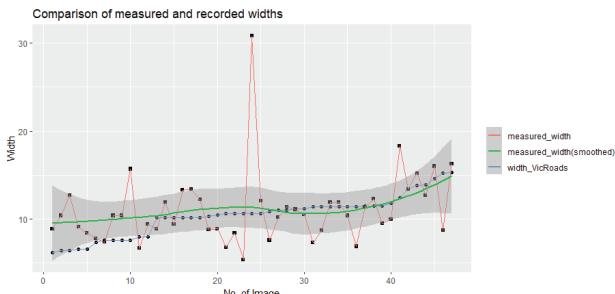


Fig. 16: Curves show measured values(red) versus recorded values(blue)

vertical occlusion errors. However, the efficient correction of OSM errors requires further investigation.

ACKNOWLEDGMENT

The first author is supported by Melbourne Research Scholarship. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

REFERENCES

- [1] M. Yasir, W. Jianhua, L. Shanwei, H. Sheng, X. Mingming, and M. Hossain, "Coupling of deep learning and remote sensing: a comprehensive systematic literature review," *International Journal of Remote Sensing*, vol. 44, no. 1, pp. 157–193, Jan. 2023, doi: <https://doi.org/10.1080/01431161.2022.2161856>.
- [2] A. Kirillov et al., "Segment Anything," arXiv:2304.02643 [cs], Apr. 2023, Available: <https://arxiv.org/abs/2304.02643>
- [3] D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang, "Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model," arXiv.org, May 03, 2023. <https://arxiv.org/abs/2305.02034>.
- [4] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng, "Segment Anything Is Not Always Perfect: An Investigation of SAM on Different Real-world Applications," arXiv (Cornell University), Apr. 2023, doi: <https://doi.org/10.48550/arxiv.2304.05750>.
- [5] S. Ren et al., "Segment anything, from space?," arXiv.org, May 15, 2023. <https://arxiv.org/abs/2304.13000>
- [6] S. Shankar, L. A. Stearns, and C. J. van der Veen, "Segment Anything in Glaciology: An initial study implementing the Segment Anything Model (SAM)," Jun. 2023, doi: <https://doi.org/10.21203/rs.3.rs-3011246/v1>.
- [7] S. Julka and M. Granitzer, "Knowledge distillation with Segment Anything (SAM) model for Planetary Geological Mapping," arXiv.org, May 15, 2023. <https://arxiv.org/abs/2305.07586>.
- [8] T. Chen et al., "SAM Fails to Segment Anything? – SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, Medical Image Segmentation, and More," arXiv (Cornell University), Apr. 2023, doi: <https://doi.org/10.48550/arxiv.2304.09148>.
- [9] Z. Chen et al., "Road extraction in remote sensing data: A survey," *International journal of applied earth observation and geoinformation*, vol. 112, pp. 102833–102833, Aug. 2022, doi: <https://doi.org/10.1016/j.jag.2022.102833>.
- [10] A. Grillo, V. A. Krylov, and G. Moser, "Road Extraction and Road Width Estimation Via Fusion of Aerial Optical Imagery, Geospatial Data, and Street-Level Images," Jul. 2021, doi: <https://doi.org/10.1109/igarss47720.2021.9554540>.
- [11] L. Luo et al., "Estimating Road Widths From Remote Sensing Images," *Remote Sensing Letters*, Jul. 2018, doi: <https://doi.org/10.1080/2150704x.2018.1484957>.
- [12] Z. Xia, Y. Zang, and J. Li, "Road width measurement from remote sensing images," Jul. 2017, doi: <https://doi.org/10.1109/igarss.2017.8127098>.
- [13] J. Guan, Z. Wang, and X. Yao, "A new approach for road centerlines extraction and width estimation," Oct. 2010, doi: <https://doi.org/10.1109/icosp.2010.5655728>.
- [14] C. Zhang et al., "A Survey on Segment Anything Model (SAM): Vision Foundation Model Meets Prompt Engineering," arXiv.org, Jul. 03, 2023. <https://arxiv.org/abs/2306.06211>.
- [15] OpenStreetMap, "OpenStreetMap," OpenStreetMap, 2023. <https://www.openstreetmap.org/>.
- [16] G. Boeing, "OSMNX: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks," *SSRN Electronic Journal*, 2016, doi: <https://doi.org/10.2139/ssrn.2865501>.
- [17] "Road Width and Number of Lanes - Victorian Government Data Directory," discover.data.vic.gov.au. <https://discover.data.vic.gov.au/dataset/road-width-and-number-of-lanes> (accessed Sep. 10, 2023).