Open-set Semantic Search in Unstructured Environments

Paulis Barzdins, Ingus Pretkalnins, Guntis Barzdins
Institute of Mathematics and Computer Science, University of Latvia, Riga, Latvia
{paulis.barzdins, ingus.pretkalnins, guntis.barzdins}@lumii.lv

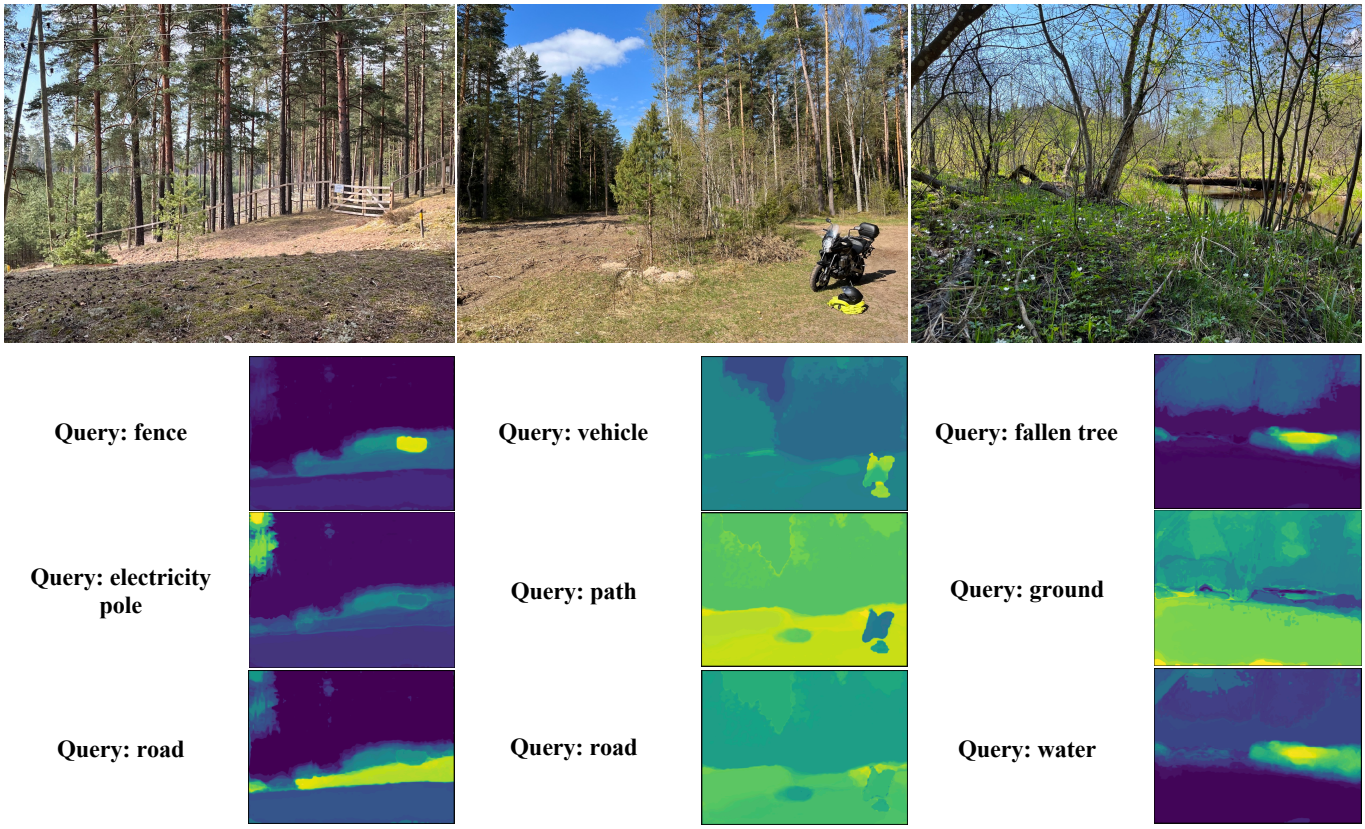

Figure 1. Three example pictures of unstructured Latvian environments, each queried with three textual prompts, results shown in yellow. The queries can be any text, long sentences as well, as long as CLIP can generate a text embedding.

ABSTRACT

As part of a general perception system for robotics in unstructured environments, the component tackled here is generating pixel level semantic embeddings for 2D images. These are pixel level CLIP embeddings which allow open-set text query correlated heatmaps (Fig. 1). In adapting SOTA technologies and improving them for our unstructured environment task we found that no single approach yet reliably outperforms others, thus in the longer term we will be working towards an ensemble of methods and models for 2D open-set semantic segmentation data, which will then be projected to a 3D data structure for the ability to map, navigate, and plan.

KEYWORDS

Open-set semantic segmentation, unstructured environments, CLIP, robotics, computer vision.

1. Perception System

The results reported on here are part of a larger project working towards a perception system for robotics, with the hope to create a general solution for perception in unstructured environments. The focus is on unstructured environments, where you need computer vision to understand the surroundings and make decisions (like GANav Unstructured Environment Terrain Segmentation [1]), as opposed to structured, where the surroundings follow strict rules, like a factory floor. The eventual goal would be a semantic 3D map that allows some textual reasoning, but an intermediary to that is generating 2D semantic information which can later be projected to 3D. That is the part we are investigating here. We show some example results of our model in Fig. 1 on images of environments which are relevant to potential end-users (like forestry and defence).

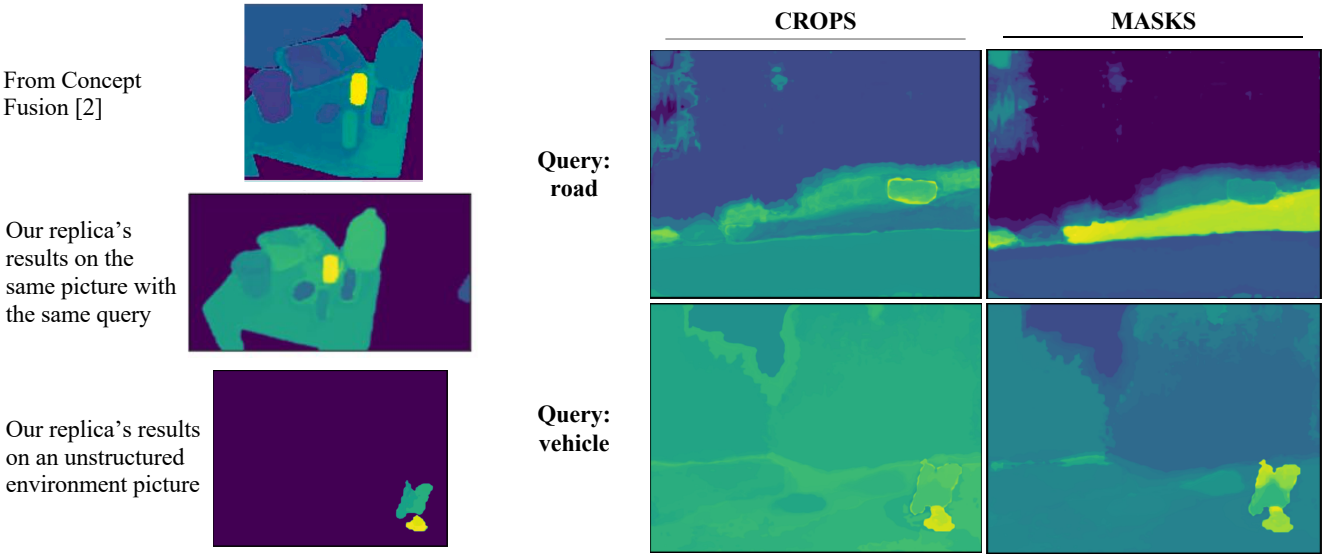

Figure 2. On the left: a mask from the Concept Fusion paper, then two masks generated by our replication of their work, using their settings; to show that we managed to replicate, but it doesn't directly work for unstructured Latvian environments. On the right: comparison between two ways of cropping object proposal masks – "crops" cuts a rectangle, "masks" paints black in the rectangle everything outside the mask.

2. Open-set Semantic Heat-maps

The method followed for generating open-set semantic heat-maps was from Concept Fusion [2]. It is to first use an object detection model (mask2former [3]) to detect potential objects; then run CLIP [4] on cropped areas of the proposed objects; followed by pixel level addition of CLIP embeddings that each pixel was part of. These pixel embeddings can then be dot product correlated with an open-set semantic query, by using CLIP to generate an embedding for the text query.

We managed to replicate the results of the Concept Fusion paper (that had no code released), but their exact settings performed poorly in unstructured environments (Fig. 2, on the left). The first change was to switch the type of object proposals from mask2former, from instance to panoptic. Thus, generating mask proposals not only for foreground objects, but larger background sections as well (important for unstructured scenes).

We then noticed that some queries lacked precision, which we hypothesised was from objects appearing in multiple crops of proposed objects (as when cropping large background objects, the foreground will be included as well). So, the second addition was to use masks in the crops, turning black all pixels outside the masked area when passing crop to CLIP. As seen in Fig. 2 on the right, this has significant improvements in multiple scenarios (although not all).

Thus, the SOTA method was adapted and improved for the unstructured environment task (compare Fig. 2 bottom left image to the heatmaps in Fig. 1).

3. Discussion

In trying multiple SOTA models and approaches for adding semantic data to images (CF [2], m2f [3], LSeg [5], ODISE [6]), we found that no single one is reliably best at our unstructured task dataset. Thus, some ensemble of tools and approaches is a likely candidate for achieving the best results.

As our current base we'll continue to use the Concept Fusion approach, as it allows modular improvements within the approach. For example, for object detection replacing mask2former with SAM [7] as we hypothesize that would give us more even object masks (mask2former was heavily biased to foreground objects) and allow us to control the granularity of our model. But there are more improvements to be made all along the pipeline, in cropping, addition of embeddings, and potentially even replacing CLIP.

With reliable 2D semantic data, it can then be projected to a 3D structure to take a step closer to a general perception system.

ACKNOWLEDGMENTS

This research is funded by the Latvian Council of Science project "Smart Materials, Photonics, Technologies and Engineering Ecosystem" project No VPP-EM-FOTONIKA-2022/1-0001 and by the Latvian Council of Science project lzp-2021/1-0479.

REFERENCES

[1] Guan, T., et al. (2022). GA-Nav: Efficient Terrain Segmentation for Robot Navigation in Unstructured Outdoor Environments. *IEEE Robotics and Automation Letters*, *7*(3), 8138–8145. https://doi.org/10.1109/lra.2022.3187278

[2] Jatavallabhula, K. M., et al. (2023). Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

[3] Cheng, B., et al. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1290-1299).

[4] Radford, A., et al. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

[5] Li, B., et al. (2022). Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*.

[6] Xu, J., et al. (2023). Open-vocabulary panoptic segmentation with text-to-image diffusion models. *arXiv preprint arXiv:2303.04803*.

[7] Kirillov, A., et al. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.