Baltic J. Modern Computing, Vol. 12 (2024), No. 1, 97-109
https://doi.org/10.22364/bjmc.2024.12.1.06

From Captions to Pixels: Open-Set Semantic Segmentation without Masks

Paulis BARZDINS, Ingus PRETKALNINS, Guntis BARZDINS

Institute of Mathematics and Computer Science,University of Latvia, Raiņa bulvaris 29, Riga, LV-1459, Latvia

{paulis.barzdins, ingus.pretkalnins, guntis.barzdins}@lumii.lv

ORCID 0009-0006-7186-9776, ORCID 0009-0007-6341-1295, ORCID 0000-0002-3804-2498

Abstract. This paper presents a novel approach to open-set semantic segmentation in unstructured environments where there are no meaningful prior mask proposals. Our method leverages pre-trained encoders from foundation models and uses image-caption datasets for training, reducing the need for annotated masks and extensive computational resources. We introduce a novel contrastive loss function, named CLIC (Contrastive Loss function on Image-Caption data), which enables training a semantic segmentation model directly on an image-caption dataset. By utilising image-caption datasets, our method provides a practical solution for semantic segmentation in scenarios where large-scale segmented mask datasets are not readily available, as is the case for unstructured environments where full segmentation is unfeasible. Our approach is adaptable to evolving foundation models, as the encoders are used as black-boxes. The proposed method has been designed with robotics applications in mind to enhance their autonomy and decision-making capabilities in real-world scenarios.

Keywords: semantic segmentation, robotics, robot vision, computer vision, deep learning.

1. Introduction

Semantic segmentation, a core task in computer vision, assigns a semantic label to each pixel in an image, enabling machines to understand and interpret visual scenes (see Fig. 1). Despite significant progress in this field, perception remains a difficult and unsolved problem (Majumdar, 2023). Existing approaches to semantic segmentation

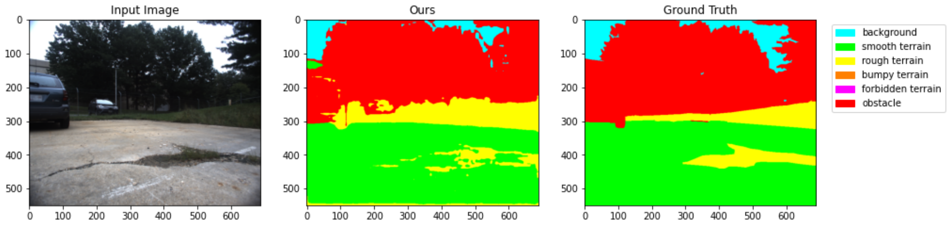

Figure 1. Image from RUGD terrain dataset with the six labels of RUGD-6. This is one application of our open-set semantic segmentation model for robot vision.

heavily rely on large-scale annotated datasets composed of segmented masks, which are labour-intensive and expensive to generate. Furthermore, these methods necessitate extensive training on powerful computational resources, rendering them less accessible for researchers and practitioners in more resource-constrained settings.

In this paper, we propose a novel approach to open-set semantic segmentation in unstructured environments, without the need for annotated masks, as fully segmenting unstructured environments is unfeasible – consider using Segment Anything Model (SAM) (Kirillov et al., 2023) to segment everything in a forest. Our method leverages pre-trained encoders from foundation models, allowing us to focus on training only a small linear mapping between text and image modalities. This significantly reduces the computational burden and time for training, while still achieving competitive results in semantic segmentation.

A key aspect of our approach is the use of image-caption datasets for training the semantic segmentation model. We introduce CLIC (Contrastive Loss function on Image-Caption data), a novel contrastive loss function that enables training a semantic segmentation model directly on an image-caption dataset. The training method has similarities to that of Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021), but unlike CLIP we don't throw away all spatial information within an image. This unique combination of leveraging foundation models and using image-caption datasets for semantic segmentation represents our novel scientific contribution in the field. The results in this paper are on par with the state-of-the-art by the metric of pixel accuracy but achieved with a simpler method.

Our approach is adaptable to evolving foundation models. As more advanced and powerful foundation models become available, our method can be easily re-trained by updating only the small mapping neural network. This ensures that our method improves alongside the development of novel big models. Additionally, our method addresses the challenges faced in niche tasks, like navigating unstructured environments, where large-scale segmented mask datasets are not readily available; attaining pixel level knowledge from image-caption datasets resolves the issue.

Our primary focus is semantic segmentation in unstructured environments, with the larger purpose being robot vision. By enabling robots to accurately perceive and understand their surroundings, our approach has the potential to enhance their autonomy and decision-making capabilities in real-world scenarios. Specific sectors such as forestry and the military have interest in perception for unstructured environments, suggesting the value of adding automation to these fields.

Review our code and implementation at our GitHub repository https://github.com/paulispaulis/CLIC-semseg for further details.

2. Related Work

To standardise our definition of "unstructured environment", we will use the Robot Unstructured Ground Driving (RUGD) dataset (Wigness et al., 2019). The RUGD dataset, designed for off-road autonomous navigation applications, captures environments that lack the structural cues commonly found in urban city autonomous navigation datasets. The data was collected with a remote-controlled vehicle that has a video camera mounted. It is small enough to manoeuvre in cluttered environments and rugged enough to traverse challenging terrain, thus exploring more unstructured areas of an environment.

In this context, we do not require a specialist model trained exclusively on RUGD, such as GA-Nav (Guan et al., 2022), a specialist terrain segmentation model. GA-Nav, while adept at applying the six labels of RUGD-6 introduced by Guan et al. (2022) (four terrain labels: smooth, rough, bumpy, forbidden; obstacle; background), lacks higher-level semantic understanding. Instead, we sought an open vocabulary model that performs well on RUGD without specific training. Such a model would assist a robot not only in navigation but in finding any arbitrary human-named object of interest. Thus, open-set understanding can later enable more intelligent robot action upon human commands.

The LSeg model (Li et al., 2022) emerged as a suitable open-set model that performs adequately on RUGD. It introduces a novel approach to semantic image segmentation using descriptive input labels and is capable of generalising to previously unseen categories at test time without requiring re-training or additional training samples. This open-set model has demonstrated competitive zero-shot performance compared to existing zero- and few-shot semantic segmentation models, and even matches the accuracy of traditional segmentation algorithms when a fixed label set is provided. Since its introduction, LSeg has served as a reference standard in open-set semantic segmentation, inspiring other models and serving as a benchmark for comparison (Ghiasi et al., 2022; Jatavallabhula et al., 2023). However, we noticed that many models that have emerged after LSeg and supposedly outperform it, perform worse on the RUGD dataset. They all, unlike LSeg, rely on prior mask proposals. Prior mask proposals are class-agnostic mask proposals generated before the prompt is considered.

Mask2former (Cheng et al., 2022) is a popular choice for generating these mask proposals (Ghiasi et al., 2022; Jatavallabhula et al., 2023). However, its proposals are ill-suited for unstructured environments as they are primarily designed for objects. Although the panoptic option allows it to generate some proposals for what could be termed "background", it does not provide sufficient granularity for terrain segmentation. Consequently, models that utilise prior mask proposals and supposedly outperform LSeg actually perform worse in unstructured environments. This observation is supported by a comparison of LSeg and ConceptFusion (Jatavallabhula et al., 2023), a model using prior mask proposals, on the same unstructured environment image (see Fig. 2), where LSeg clearly outperforms ConceptFusion.

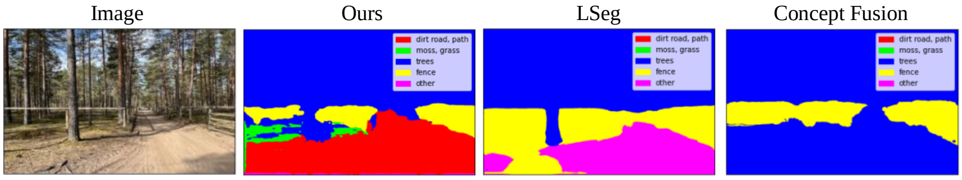

Figure 2. Semantic segmentation of an unstructured environment by our model, LSeg, and Concept Fusion. Labels in order are “dirt road, path”, “moss, grass”, “trees”, “fence”, “other”.

An interesting model in this context is OpenSeg (Ghiasi et al., 2022), which, while still reliant on prior mask proposals, recognises the value of image-caption datasets over segmented mask datasets. OpenSeg utilises a mixture of segmented mask datasets to train its teacher model, which is then used to auto-label the image-caption data. This approach demonstrates an appreciation for the scalability of image-caption datasets, which is a promising direction for our research. However, the authors of OpenSeg have not yet released their code. This makes it challenging to fully test and verify their work. Despite this limitation, their acknowledgement of the potential of image-caption datasets over segmented mask datasets is encouraging and aligns with our approach of improving upon the mechanisms of LSeg by using image-caption datasets instead of segmented masks, while not introducing the use of prior mask proposals.

Perhaps the first work to do away with masks completely is Xu, J. et al. (2023), but their architecture still uses prior masks which, as they themselves say in their paper, limits their model to foreground objects and thus is not applicable to unstructured environments. Work done by Xu, M. et al. (2023) addresses unstructured environments similar to RUGD-6 but is trained on the mask-annotated COCO-Stuff dataset.

There is alternative work being done on connecting words with images from the side of Large Language Models (LLMs). First to note is PaLM-E (Driess et al., 2023), a large embodied multimodal model, trained end-to-end with custom collected multi-modal data. It directly incorporates real-world continuous sensor modalities into language models to establish a link between words and percepts. The model takes multi-modal sentences as input, which interleave visual, continuous state estimation, and textual input encodings. These encodings are trained end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks. Three other notable next-generation models that have emerged as this paper was being written are Bard (Manyika, 2023), Language Instructed Segmentation Assistant (LISA) (Lai et al., 2023), and GPT-4Vision (OpenAI, 2023). These models combine the power of LLMs with image understanding and reasoning capabilities, and in the case of LISA, even segmentation capabilities. All these models are very promising, although their size and computational requirements would likely preclude their deployment on portable robots operating in unstructured environments.

3. Our Approach: CLIC for Semantic Segmentation

In our pursuit of open-set semantic segmentation in unstructured environments, we adopted a new approach. We leveraged the strengths of pre-trained foundation models and image-caption datasets and combined them with CLIC, the novel contrastive loss function, to overcome the limitations of existing methods. Our philosophy is centred around three core principles: scalability, adaptability, and practicality.

Firstly, scalability is a key consideration in our approach. We believe that the ability to learn from weak labels is crucial for scaling training data. More data means that the model is directly exposed to greater vocabulary during training. We learn visual-semantic alignment from image-caption datasets. This contrasts with other methods, such as LSeg using only semantic mask datasets (thus limiting vocabulary exposure while hoping that the CLIP text encoder would give meaningful correlations to new vocabulary), or ConceptFusion which relies solely on CLIP for visual-semantic alignment by cropping class-agnostic mask proposals and running CLIP on them. Using image-caption datasets allows us to expand the range of vocabulary where our model has had direct experience and enhance its learning capabilities.

Secondly, our approach is designed to be adaptable. We acknowledge the dynamic nature of unstructured environments, where it is often impossible to generate masks before knowing the prompt. As such, we generate the masks only after receiving the prompt, allowing our model to adapt to diverse and unpredictable terrains. All this while maintaining the open-set query ability, since adaptability is particularly important in the context of robotics, where both navigation and open-set language understanding is crucial.

Thirdly, our approach is grounded in practicality. We aim to develop a robust, generalist model with zero-shot abilities, rather than a terrain specialist model. This enhances the model's versatility and resilience, ensuring it performs well even in terrains not seen in the training dataset.

Using the CLIP text encoder enables open-set prompts; CLIC then enables open-set segmentation; the open-set segmentation together with being able to use large scale image-caption datasets enables zero-shot performance. Thus, our approach builds upon the strengths of existing models such as LSeg and OpenSeg, while addressing their limitations. We use CLIP for text encodings to harness the generalisation capabilities demonstrated by LSeg, and train on image-caption data to expose our model to a wide range of vocabulary. Our model currently uses the Mask2Former image encoder and CLIP text encoder, but these can be replaced with other models in future iterations (see Section 5).

In essence, our approach offers a novel framework for semantic segmentation in unstructured environments. By integrating image-caption datasets and foundation models, we train a model that is scalable, adaptable, and practical. This approach not only addresses the challenges of existing methods, but also opens new possibilities for semantic segmentation in real-world scenarios.

3.1. Components Used

Our approach to open-set semantic segmentation incorporates several key components, including the Mask2Former image encoder (Cheng et al., 2021), the CLIP text encoder, and the Common Objects in Context (COCO) image-caption dataset (Lin et al., 2014).

Mask2Former, a versatile architecture for image segmentation, has been a vital component of our research. Mask2Former is a ground-breaking model that can address any image segmentation task, including panoptic, instance, or semantic segmentation. It employs masked attention to extract localized features by limiting cross-attention within predicted mask regions. The model has shown remarkable performance, surpassing specialized architectures on several popular datasets. In our work, we specifically use the image encoder from Mask2Former to extract features from images.

In parallel to the image encoder, we utilize the text encoder from the CLIP model. CLIP is a transformer-based model trained to understand and generate meaningful representations from both images and text. It has been designed to learn visual concepts from natural language supervision. The integration of CLIP's text encoder allows us to leverage the model's ability to understand the semantics of the scene.

Lastly, the COCO image-caption dataset is used for training our model. The COCO dataset is a large-scale object detection, segmentation, and captioning dataset that includes images of complex everyday scenes with detailed annotations. By using only image-caption data from the COCO dataset, we aim to enhance the model's ability to understand components of unstructured environments and other niche tasks and segment the semantics of the scene. This, combined with the other components, forms the backbone of our research into open-set semantic segmentation.

While we used the COCO dataset for training our model, other image-caption datasets are abundant. MS-COCO contains 200,000 images, but there are other datasets, like Google's Conceptual Captions (Sharma et al., 2018) or Open Images Dataset (Kuznetsova et al., 2020), containing 3.3 million and 9 million images respectively.

Future research could explore the integration of these datasets to further improve the performance of open-set semantic segmentation models in unstructured environments.

3.2. Model Architecture and CLIC

Our model stands on a two-pronged architecture of pre-trained foundation model image and text encoders (see Fig. 3); their output embeddings are aligned by a linear transform which we train ourselves using only image-caption datasets.

We use the Mask2Former encoder to transform the input images. The output of this process is an embedding of size $[256, 96, 96]$, where 96×96 is the image resolution and 256 is pixel embedding length. To fit the available memory during training, we downscale the output resolution 8 times, resulting in an embedding of size $[256, 12, 12]$. To augment embeddings for robustness and generalisation, we add noise with a normal distribution and a standard deviation of 0.07 to the image embeddings, a value obtained through random search.

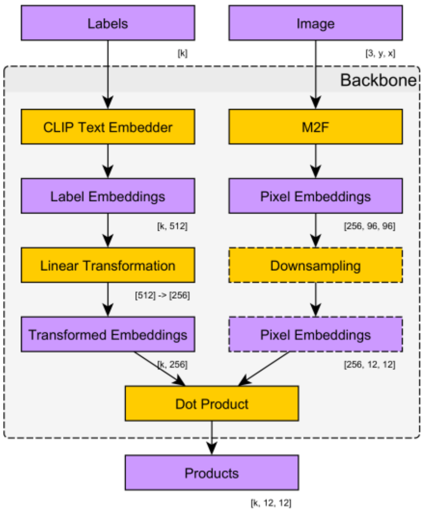

Figure 3. The backbone of the model architecture (the dotted down-sampling boxes apply only to training, for memory conservation when training on large datasets).

The text encoder utilises the CLIP text embedder to encode the captions associated with the images. The output is an embedding of length $[512]$. Like the image encoder, we add noise with a normal distribution and a standard deviation of 0.15 to the text embeddings, a value obtained through random search.

To align the dimensions and semantics of the image (256) and text (512) embeddings, we perform a linear transformation between the text and image embeddings. This transformation is the only model we are training.

We then perform a dot product operation between the transformed text embedding and each pixel embedding. The result is a $[12, 12]$ matrix, showing the relevance of each spatial pixel to the caption. We squash the values with the sigmoid function and then take the maximum from this matrix (see Fig. 4, left), to obtain a single value (max) representing the detection score of the caption being in the corresponding image.

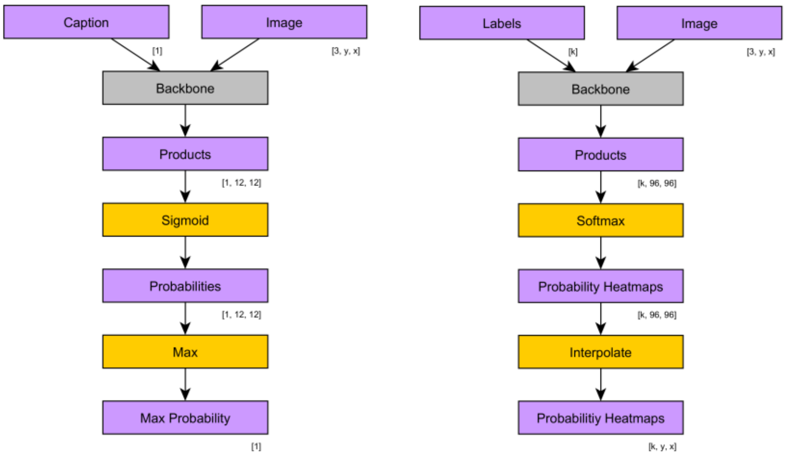

Figure 4. Model architecture at training (left) and inference/evaluation (right). Interpolation is upsampling to the original image resolution.

CLIC, the contrastive loss function on image-caption data, is a crucial novel component of our approach. It allows to train the linear transform above from image-caption data only. During training the combination of sigmoid and max returns a single value from 0 to 1 of whether the model sees the caption in any pixel within the image. If the image-caption pair is a match, the loss is calculated as $(max - 1)^2$, and if not, the loss is max^2. We use contrastive training. This loss is minimised by adjusting the parameters of the linear transformation. Before this paper it was thought that image-caption data is too coarse to directly be able to train a model for semantic segmentation (Xu, J. et al., 2023). CLIC does just that.

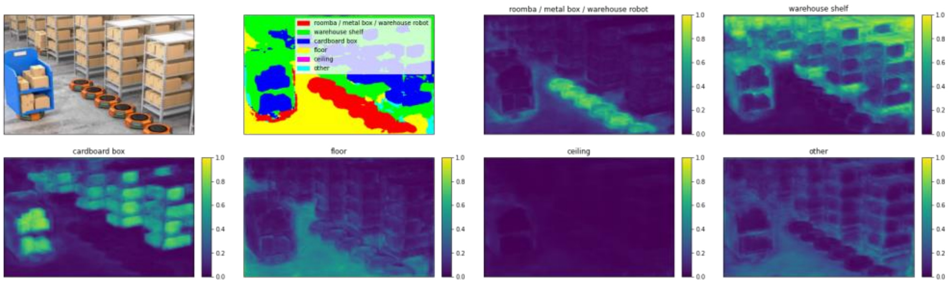

Figure 5. Output heatmaps from a query on a warehouse image, as well as the full segmentation output when the argmax for each pixel is taken from the heatmaps.

During evaluation (see Fig. 4, right), for the datasets with semantic masks, the model operates similarly as at training, but with a few key differences. The Mask2Former output is not downscaled, resulting in an image embedding of size $[256, 96, 96]$. The evaluation datasets come with a fixed set of k possible pixel class labels; all k class labels are passed through the CLIP text embedding, resulting in an output of size

$[k, 512]$, passed through the linear transformation to an output of $[k, 256]$, thus aligned and ready for the dot product operation. The result of the dot product operation is a $[k, 96, 96]$ matrix. SoftMax is applied to get a probability of each label at each of the $[96, 96]$ pixels. This is then interpolated to the original image dimensions to get a $[k, y, x]$ matrix, which can be interpreted as k comparable heat maps for each of the k labels (see Fig. 5). To get the full segmentation we simply take the argmax for each pixel. The full segmentation is compared with the correct masks from the dataset to evaluate the model's performance.

4. Empirical Evaluation and Results

In this section, we present the experimental setup and results that validate our approach for open-set semantic segmentation in unstructured environments. We evaluate performance on two metrics – pixel accuracy and mean intersection over union (mIoU). The first (Equation 1) measures what part of all pixels have been labelled correctly. The second (Equations 2 and 3) measures for each label the IoU (how well the predicted masks align with the ground truth) and then takes the mean over all labels.

$$\text{Pixel Accuracy} = \frac{\sum_{c=1}^{k} \text{True Positive}_c}{\text{Total Pixels}} \tag{1}$$

$$\text{IoU}_c = \frac{\text{True Positive}_c}{\text{True Positive}_c + \text{False Positive}_c + \text{False Negative}_c} \tag{2}$$

$$\text{mIoU} = \frac{1}{k}\sum_{c=1}^{k} \text{IoU}_c \tag{3}$$

We are considering only a full segmentation of the image, where every pixel is assigned a class label. If we want to find bicycles and pedestrians, we must also give a third class “background” for the pixels that are neither bicycles nor pedestrians. This limitation comes from the LSeg method of converting heatmaps to strict segmentation masks, also used in our approach. This also implies that pixel accuracy is a better metric for the model performance than mIoU because pixel accuracy operates over all pixels of the image, while mIoU is designed to operate over foreground objects of interest. mIoU can be used for a fully segmented image (with background classes) but this distorts the results. What mIoU catches, if used properly, is the model’s performance on classes that take up few pixels; this is good for catching the model ignoring small objects or seldom occurring classes.

We trained our model on the entire COCO dataset and then compared its performance with LSeg on the RUGD-6 dataset (see Table 1).

Table 1. Comparison of LSeg and our model on RUGD-6.

Model	Pixel Accuracy	Mean IoU
LSeg	83.6	60.0
Ours	**87.1**	54.0

On RUGD-6, our model outperformed LSeg in pixel accuracy, the metric we just argued is more appropriate for full segmentation. We also report mean IoU, where our model was slightly inferior. IoU scores and their mismatch are explored just above as well as at the introduction of Table 3. The results demonstrate that our model, which does not rely on prior masks, can achieve competitive results in unstructured environments. While the score was improved only slightly, it was achieved without the use of segmented masks, but only image-caption pairs. These are exponentially more available for training, meaning our method could still improve from a scaling of training data (also argued by Ghiasi et al., 2022, and Xu, J. et al., 2023).

At this point our goal of an open-set segmentation model, which works in unstructured environments and improves upon LSeg, has been reached. The following empirical evaluations illustrate our model's performance in early real-world use-cases.

To illustrate the applicability of our model in real-world scenarios and the benefit of open-set segmentation, we tested it on a completely new use case - a set of warehouse images from an industry partner. The model was able to accurately identify warehouse robots, shelves, boxes, floor, and ceiling, demonstrating its open-set capabilities and zero-shot performance. In this context, the improvement upon LSeg also becomes much clearer (see Fig. 6).

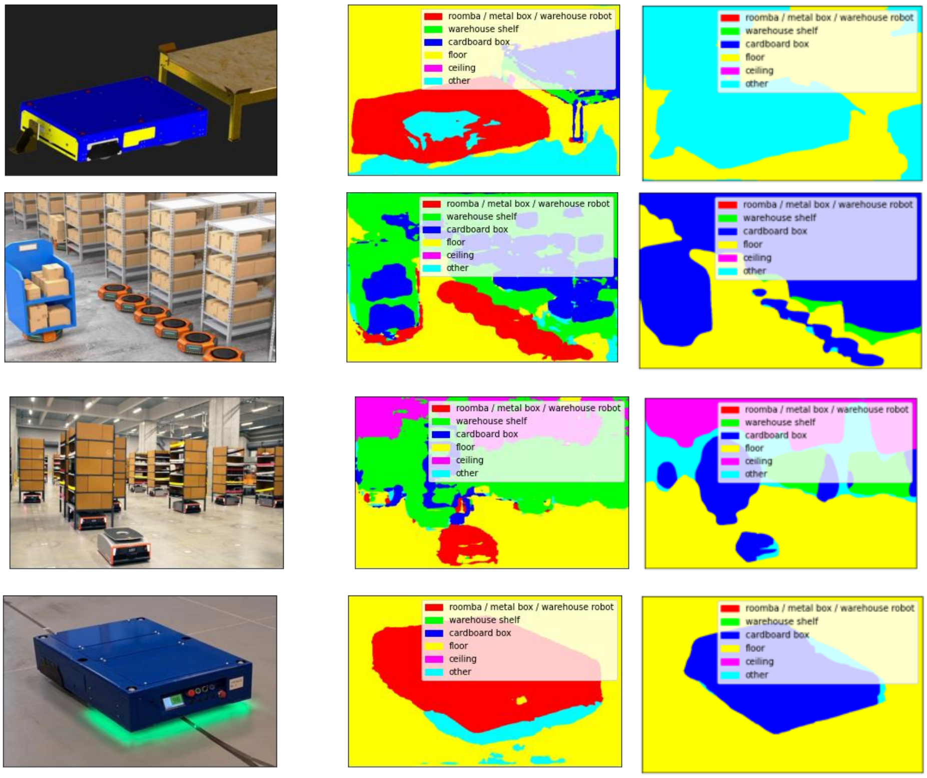

Figure 6. Warehouse images, zero-shot segmentation performed by our model trained only on COCO image-captions (column 2), then LSeg (column 3).

In addition to its zero-shot capabilities, our model can be fine-tuned for specific domains. As an illustration, we fine-tuned our model on the RUGD dataset and achieved results roughly on par with GA-Nav (see Table 2). This achievement is significant as it highlights the capacity of an open-set, non-task-specific model finetuned on one terrain dataset to compete with a specialist terrain segmentation model trained on multiple terrain datasets. This demonstrates the impressive depth of semantic information that the Mask2Former image encodings contain. However, some classes like "water" were underrepresented in the RUGD dataset, leading to lower mean IoU scores. This could be addressed by prompt engineering (adding “puddle, pond, river, stream” etc. as queries, not only “water”) or by adding more data from other terrain datasets. Despite these challenges, our model performed admirably in the more represented classes, also by the IoU metric (see Table 3), even outperforming GA-Nav for the Background class (containing void, sky, sign).

Table 2. Performance of our model trained only on COCO with no semantic masks (untuned) compared to our model finetuned on RUGD (finetuned) compared to GA-Nav (trained on RUGD and RELLIS (Jiang et al., 2021) masks). For these models we used the same validation split as GA-Nav for fair comparison.

Model	Pixel Accuracy	Mean IoU
GA-Nav	95.66	89.08
Untuned	86.6	48.9
Finetuned	92.5	77.4

Table 3. Per-class IoU comparison of our finetuned model and GA-Nav on the six RUGD-6 classes

Model (IoU)	Rough Region	Obstacle	Smooth Region	Background	Forbidden Region	Bumpy Region
Finetuned	87.40	86.75	82.85	**78.80**	65.60	64.59
GA-Nav	94.45	91.95	95.15	76.86	86.25	89.83

.

While fine-tuning can improve the model's performance in specific domains, it can also lead to forgetting if not done carefully. This was observed when the RUGD fine-tuned model performed poorly on the warehouse images (see Fig. 7). This issue can potentially be mitigated by alternating training batches between the COCO and the specific domain dataset.

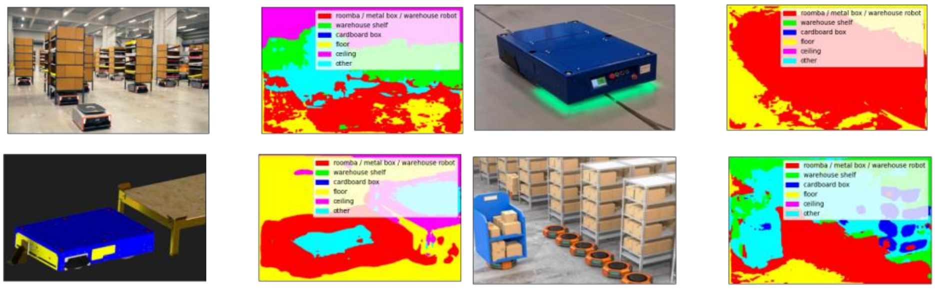

Figure 7. Display of forgetting when applying the RUGD finetuned model on the warehouse image task.

5. Conclusion and Future Work

This paper introduces an innovative method for open-set semantic segmentation in unstructured environments, which does not rely on annotated masks. By leveraging pre-trained encoders from foundation models and using image-caption datasets for training, we developed a model that requires fewer computational resources and less training time yet achieves competitive results in semantic segmentation.

Our approach is adaptable and scalable, and it provides a practical solution for scenarios where large-scale segmented mask datasets are not readily available. The model outperformed the LSeg model in pixel accuracy and demonstrated comparable results in mean intersection over union (mIoU) on the RUGD-6 dataset. Furthermore, our model showed promising zero-shot performance on a new use case involving warehouse images, as well as displaying potential for fine-tuning to specific domains.

Moving forward, we aim to refine our model and expand its capabilities. We plan to explore the integration of other image-caption datasets to further improve the model's performance. We also envision combining our framework with models like SAM and Generative Pre-trained Transformers (GPT) to create a powerful blend of reasoning and high-quality masks to advance semantic segmentation in unstructured environments and robot vision.

Acknowledgements

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

Cheng, B., Schwing, A., Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, **34**, 17864-17875.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R. (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).

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... Florence, P. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Ghiasi, G., Gu, X., Cui, Y., Lin, T. Y. (2022). Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision* (pp. 540-557). Cham: Springer Nature Switzerland.

Guan, T., Kothandaraman, D., Chandra, R., Sathyamoorthy, A. J., Weerakoon, K., Manocha, D. (2022). Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments. *IEEE Robotics and Automation Letters*, **7**(3), 8138-8145.

Jatavallabhula, K. M., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., Iyer, G., Saryazdi, S., Keetha, N., Tewari, A., Tenenbaum, J. B., de Melo, C. M., Krishna, M., Paull, L., Shkurti, F., Torralba, A. (2023). Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

Jiang, P., Osteen, P., Wigness, M., Saripalli, S. (2021). Rellis-3d dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 1110-1116). IEEE.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... Girshick, R. (2023). Segment anything. *arXiv preprint arXiv:2304.02643*.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., ... Ferrari, V. (2020). The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, **128**(7), 1956-1981.

Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J. (2023). Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., Ranftl, R. (2022). Language-driven Semantic Segmentation. *CoRR, abs/2201.03546.*

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (pp. 740-755). Springer International Publishing.

Majumdar, A. (2023). Robotics: An Idiosyncratic Snapshot in the Age of LLMs. *IRoM Lab* https://irom-lab.princeton.edu/wp-content/uploads/2023/08/Robotics_snapshot.pdf

Manyika, J. (2023). An overview of Bard: an early experiment with generative AI. *AI. Google Static Documents*.

OpenAI (2023). GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

Sharma, P., Ding, N., Goodman, S., Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume* ***1****: Long Papers)* (pp. 2556-2565).

Wigness, M., Eum, S., Rogers, J. G., Han, D., Kwon, H. (2019). A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 5000-5007). IEEE.

Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W. (2023). Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2935-2944).

Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X. (2023). Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2945-2954).

Received November 2, 2023, revised February 20, 2024, accepted March 6, 2024