GENERATIVE OBJECT DETECTION MODELS FOR MOBILE ROBOTICS APPLICATIONS

van Genuchten, H.J.M,
ID: 1297333
Artificial Intelligence and Data Engineering Lab
Eindhoven University of Technology
h.j.m.v.genuchten@student.tue.nl

Dr. Tomczak, J. M.
Artificial Intelligence and Data Engineering Lab
Eindhoven University of Technology
j.m.tomczak@tue.nl

Keulen, B.
Avular
b.keulen@avular.com

February 5, 2024

ABSTRACT

This research proposes a Generative Segmentation Model for mobile robotics, prioritizing novel object recognition and robust handling of Out-of-Distribution (OoD) data. The objective is to enhance mobile robotic systems' navigation in unknown terrains by creating a model capable of resilient performance against OoD or, when necessary, signaling and triggering a fail-safe mode. Motivated by AI challenges in dynamic scenarios, the research integrates Uncertainty Quantification (UQ) to address Neural Networks' limitations with OoD data. UQ plays a crucial role in distinguishing certain and uncertain predictions, allowing for preemptive reactions or human intervention. Furthermore, UQ can be used for active learning, which reduces training costs.

The research will compare Generative Segmentation Models with state-of-the-art (SOTA) models, by assessing robustness against OoD and novel object recognition. Furthermore, the effectiveness of UQ is assessed. Finally, if time permits, this research will investigate the possibility of efficiently incorporate temporal data into Denoising Diffusion Segmentation. In summary, this research outlines a concise plan for a generative Object Detection model, integrating Uncertainty Quantification to address challenges in mobile robotics and contributing to the advancement of AI applications in real-world scenarios. **Wordcount: 183**

1 Overall aim and goals

The main goal of this thesis will be the implementation of a (generative) Object Detection model for mobile robotics. An important part of this is the ability to recognize novel objects. Furthermore, it is important to be able to have a good understanding of the quality of the prediction. As it is well known that Neural Networks often perform bad when presented with Out-of-Distribution (OoD) data. As mobile robotics are required to navigate unknown terrain, it is important that the model is capable of handling OoD data. This can be achieved either by ensuring the model is robust against OoD data or enabling the model to warn the user when OoD data is detected, and fallback to a fail-safe mode.

1.1 Motivation and Challenges

Artificial Intelligence solutions are increasingly put in new and challenging scenarios. They are known to work well when the training and inference set are from the same distribution [25]. However, it has been shown that models will predict with high scores for inputs that are not relevant [30, 36]. It has also been shown that this can be used to attack these networks[20, 11]. Uncertainty Quantification (UQ) can enable a system to detect when a prediction might be of lesser quality [10], and allow it to preemptively react to that[34]. Either by stopping or requesting human intervention.

Furthermore, understanding when models are uncertain, allows for more effective data sampling and labeling by making use of active learning schemes [33, 2, 44]. The latter is especially useful in industries where labeling is expensive or time-consuming.

Object Detection is an especially difficult field as the model has many outputs of different types. It requires both regression (for localization) and classification [16]. Moreover, there are multiple uncertainties present:

- Objectness: Is a bounding box an object, instead of background?
- Classification: Is a bounding box of a certain class?
- Size: Is the size of the bounding box correct?
- Location: Is the location of the bounding box correct?

A part of these problems are less prominent in segmentation tasks. However, segmentation requires more laborious labels which increases the cost of labeling. Therefore, it is important that segmentation is either very data efficient or is robust against OoD, so that it can be (pre)trained on public segmentation datasets.

1.2 Research questions

To achieve the above-mentioned goal I will try to answer the following research questions:

- Is it possible to detect novel objects using Generative Segmentation Models?
- What is the performance difference between SOTA Segmentation Models and a Generative Segmentation Model?
- Is the Generative Segmentation Model more robust against OoD?
- Is the uncertainty quantification given by the Generative model an useful indicator ...
  detecting wrongly labeled items?
  for active learning?
- Can temporal data efficiently be used in Denoising Diffusion Segmentation?

1.3 Broad Literature Analysis

This project covers broader research areas, each will be covered separately in subsections 1.3.1 and 1.3.2. Related work in the combination will be described in subsection 1.3.3

1.3.1 Uncertainty Quantification

The ability to distinguish certain and uncertain outputs from machine learning models is useful for various reasons. It can be used for active learning [44, 33, 23, 2], which makes better use of limited labeling capacity.

[14] distinguishes two kinds of uncertainty. Aleatoric uncertainty, the uncertainty that is caused by imprecise input data, and epistemic uncertainty, the uncertainty that is caused by the model. Aleatoric uncertainty is often caused by imprecise measurements and can be reduced by improving the quality of our dataset and inputs. The latter, epistemic uncertainty, can be reduced by improving either the training procedure, increasing the amount of data or improving the model architecture.

Uncertainty Quantification (UQ) is an important factor to increase the trust in automated processes based on machine learning. Current methods often sample from existing networks [15, 28, 29] or predict parameters of a distribution [8, 35].

[37] apply uncertainty quantification on a diffusion model for medical imagery. They show it is capable of predicting both the aleatoric and epistemic uncertainties. Which allows them to catch up to 90% of the unreliable predictions.

1.3.2 Object Detection

There are two main paradigms within Object Detection, One-Stage detectors[46, 3, 41, 27, 13], which directly predict both the bounding box and class in a single forward pass, and Multi-Stage detectors [19, 18], which first detect regions of interests to then subsequently classify these regions. A more recent development is DiffusionDet [7], which iteratively 'de-noises' a set of random bounding boxes towards the ground truth bounding boxes. Furthermore, there are some general tricks that have shown to provide improved performance at the cost of an increased training time for most object detection algorithms. These have been dubbed "Bag of Freebies" by [45], as they do not increase inference time.

One-Stage Detectors The Single Shot Detector proposed by [27] makes use of many anchor boxes. For each anchor box an offset and class is predicted, which are subsequently merged using Non-Max-Suppression (NMS). During training all bounding boxes with an Intersection over Union (IoU) greater than a threshold are matched and should be predicted with that class. All unmatched boxes should be classified as a 'background' class. This leads to a huge class imbalance. To tackle this, hard negative mining is used. For every positive match, at most n (in the paper 3) negative boxes are included during the loss calculation.

A more recent development is the DEtection TRansformer (DETR) [5]. Which uses transformers ([40, 12]) to generate the set of bounding boxes in parallel. It is one of the few models which does not make use of a NMS function to remove duplicate bounding boxes. Furthermore, it is easily extendable to panoptic[1] segmentation.

Multi-Stage Detectors Multi-stage Detectors are usually made up of two parts. The first part produces proposals, i.e. patches of objects. A second network then predicts the class. The most notable family of models are the Region-CCN (R-CCN) models [18, 31]. These methods are generally slower as they require multiple inferences for the same image.

Zero-shot Detection Generalizing to unseen objects is a difficult problem, as it is hard to distinguish between a "background" and an unseen class. In [1] the problems of zero-shot object detection are explained. An often used method to tackle the problem of unseen class labels is the use of multimodel models that have a shared latent space containing both the image and (usually) text embeddings. These models can then be queried to return unseen classes. However, this does not tackle the problem of detecting unknown[2]

Recent developments in Vision Transformers (ViT) [12] have been used to pre-train backends for object detection tasks [4]. The benefit of ViT is the ability to use unlabeled data to pre-train the transformer part of the model. A relatively small fine-tuning dataset can then be used to train the model for a specific task. They have also been shown to segment objects in a fully self-supervised way [6].

Segmentation Closely related to object detection is image segmentation. Within image segmentation there are 3 main subfields: semantic, instance and lastly panoptic segmentation. Semantic segmentation assigns a semantic value to each pixel of the image [42, 43]. Instance segmentation segments the various instances within an image, assigning a different ID to each instance in the same image. Finally, panoptic segmentation combines the two, assigning both an ID and a semantic value to each pixel.

1.3.3 In combination

The combination of Uncertainty Quantification and Object Detection is especially difficult as Object Detection is both a regression and a classification task. A recent advancement in this is CertainNet by [16]. Which is capable of providing sample free certainty estimations on the various aspects of the prediction. A followup paper also by [17] expands the methods into segmentation. They furthermore propose the U3HS framework to enhance the capabilities of the model to detect unknown unseen instances. These unknown instances can still be used to do tracking, trajectory prediction and path planning during inference. Furthermore, within an online training scheme it allows for active learning.

2 Research approach

2.1 Overall methodology and decomposition

The project consist of 2 (+ 1 possible extension) parts, the first is the reproduction of LDMSeg [39]. The second part will be the extension of that work by adding UQ to it. Possibly adapting ideas from [17] to the generative approach of LDMSeg. Finally, if time permits an extension to temporal data will be made. This would both improve the inference efficiency and hopefully the quality of the model. Moreover, it will show the benefits of using generative models versus discriminative as the extension should not require retraining.

2.2 Methods and techniques

Model formulation The Latent Diffusion Model for Segmentation requires a dataset consisting of images $x \in \mathbb{R}^{H \times W \times C}$ and corresponding panoptic segmentation masks $y \in \mathbb{R}^{H \times W \times N}$. An auto-encoder is trained to get the latent

[1] Panoptic segmentation is the combination of instance and semantic segmentation.

[2] Note the difference between unknown and unseen. Unknown being things we do not know exist, whereas unseen means it can be described by its attributes.

representations z_i and z_y for the image and target respectively. A diffusion model is then trained conditioned on the latent image z_i, resulting in Eq. 1.

$$p(y|x) = p(y|z_y, z_i) * p(z_y) \tag{1}$$

For the uncertainty quantification, similar to [37, 17, 38], instead of using the output of the model directly, the model produces 2 outputs per pixel which are then used to sample a Gaussian noise using $\mathcal{N}(\mu, \sigma)$. Where μ and σ are the outputs of the model. The loss can then be adapted as shown in Eq. 2.

$$L_u q(\mu, \sigma, y) = \frac{L(\mu, y)}{1 + \sigma} + \exp \sigma \tag{2}$$

The model can easily be extended with a temporal dimension by using the previous time-step latent target $z_{y,t-1}$ as seed for $z_{y,t}$.

Metric formulation Current SOTA models report on the following metrics:

- Panoptic Quality [24]
- Uncertainty Error [29]
- (AU)ROC [29]

To further investigate the generalization capabilities of the model, it will tested on various datasets. E.g. trained on CityScape [9] dataset and evaluated on MSCOCO [26]. For temporal segmentation only small datasets are available such as TSSB [32]. However, as the model does not need to be trained on temporal data, it should be good enough as a proof of concept.

2.3 Research plan and timeline

Week	Description	Expected Result	Deliverable
1	Setting up work environment	Have a working dev-env with proper CI integrations on git, so the quality of the repository is maintained.	Reproducible and testable github repository that can be easily cloned and worked on.
2–4	Reproduce [39]	Reproduction of important prior work	A training in (W&B) using the provided code by [39]
5–7	Add uncertainty quantification to the model	Added uncertainty quantification and it's respective metrics to the code base.	UQ results on small model
8–10	Set up reproducible experiments which can be compared to SOTA.	Results on small scale models and datasets.	
11–14	Investigate the feasibility of adding temporal data to the model	If feasible add temporal data to the model, else write about it it	
15–17	Writing report and running bigger experiments	Some results	nothing
18	Hand in draft-report to supervisors for final feedback	Implementing the received feedback	Final version of thesis
20	Defense	Successful defense	Master diploma

2.4 Knowledge utilisation/ valorisation / expected contributions and impact

The main contribution will be to industry, as one of the results of this research will be a model that can be used by Avular. Moreover, I hope the resulting code can be used to continuously improve the model using an active learning scheme based on feedback received from the models that will be interacting in the real world. Although the validation of the active learning is not in scope of this thesis.

3 Evidence that your research can succeed

The main contributions of this work rely on the combination of two prior works. Hence, the main evidence that the research is likely to succeed are found in literature. Primarily, the paper by [39] show the feasibility of panoptic segmentation with LDMs. Although they do not reach SOTA levels in Panoptic Quality (PQ) [24] on popular benchmarks. [39] do not show the OoD capabilities of the model. Adding UQ to the model is a known technique which has been shown to work in previous work in segmentation [34, 37].

Adding temporal information, by e.g. providing the previous segmentation mask, to the diffusion model makes intuitive sense. As the previous frame's segmentation mask is likely very similar to the current frame's segmentation mask. Furthermore, it possibly enables 'free' tracking as [39] show that the model is also capable of 'in-painting' the segmentation mask, in which each instance seems to keep the same panoptic ID. Finally, diffusion models have already been successfully used to generate videos [21], and are thus capable of temporal modeling.

4 Other Information

4.1 Data management

Will this project involve re-using existing research data? There are many open-source datasets containing millions of images. These are also widely used in reports to compare various approaches. I will also use these datasets. During development, I am mainly planning on using the Pascal VOC [22] datasets. However, in the final report I would like to report on more datasets if the time and compute allows for it.

I would like to also publicly provide any code required to rerun the experiments reported on in the paper, however as this research will be done in collaboration with a company it might not be possible to share all code. All code that can be made public, will be shared posted on GitHub in a public archived repository.

Will data be collected or generated that are suitable for reuse? Depending on if the model will be trained using company resources (data and/or compute), the model weights can be reused. If possible I would like to make the weights used for reporting the metrics on Pascal VOC publicly available such that the experiments can easily be repeated by third parties.

Possible data sharing restrictions As this project is in collaboration with a company, if company IP is used for (parts of) the research it might not be possible to publicize those parts of the code. However, as this is a research project and does not (necessarily) involve running the code on production units. It is likely that a large part of the code can be made public.

4.2 Motivation for choice of research group / supervisor / company

I chose for this project due to my interest and prior work experience with Object Detection. I have a preference of doing my thesis at a company as I feel like that I would have a more tangible end-goal versus a purely research based project. Furthermore, the robotic background of Avular interests me as I hope to be able to see the results of my work in the real world, applied on physical products.

References

[1] Ankan Bansal et al. *Zero-Shot Object Detection*. 2018. arXiv: 1804.04340 [cs.CV].

[2] Mélanie Bernhardt et al. "Active label cleaning for improved dataset quality under resource constraints". In: *Nature Communications* 13.1 (Mar. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-28818-3. URL: http://dx.doi.org/10.1038/s41467-022-28818-3.

[3] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. 2020. arXiv: 2004.10934 [cs.CV].

[4] Nicolas Carion et al. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV].

[5] Nicolas Carion et al. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[6] Mathilde Caron et al. "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[7] Shoufa Chen et al. *DiffusionDet: Diffusion Model for Object Detection*. 2023. arXiv: 2211.09788 [cs.CV].

[8] Jiwoong Choi et al. “Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving”. In: *Proceedings of the IEEE/CVF International conference on computer vision*. 2019, pp. 502–511.

[9] Marius Cordts et al. “The cityscapes dataset for semantic urban scene understanding”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3213–3223.

[10] Sina Däubener et al. *Detecting Adversarial Examples for Speech Recognition via Uncertainty Quantification*. 2020. arXiv: 2005.14611 [eess.AS].

[11] Yinpeng Dong et al. “Boosting adversarial attacks with momentum”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 9185–9193.

[12] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[13] Kaiwen Duan et al. “Centernet: Keypoint triplets for object detection”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6569–6578.

[14] Yarin Gal et al. “Uncertainty in deep learning”. In: (2016).

[15] Yarin Gal and Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML].

[16] Stefano Gasperini et al. “CertainNet: Sampling-Free Uncertainty Estimation for Object Detection”. In: *IEEE Robotics and Automation Letters* 7.2 (Apr. 2022), pp. 698–705. ISSN: 2377-3774. DOI: 10.1109/lra.2021.3130976. URL: http://dx.doi.org/10.1109/LRA.2021.3130976.

[17] Stefano Gasperini et al. “Segmenting known objects and unseen unknowns without prior knowledge”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 19321–19332.

[18] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[19] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[21] Jonathan Ho et al. “Video diffusion models”. In: *arXiv:2204.03458* (2022).

[22] Derek Hoiem, Santosh K Divvala, and James H Hays. “Pascal VOC 2008 challenge”. In: *World Literature Today* 24.1 (2009).

[23] Neil Houlsby et al. *Bayesian Active Learning for Classification and Preference Learning*. 2011. arXiv: 1112.5745 [stat.ML].

[24] Alexander Kirillov et al. “Panoptic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9404–9413.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[26] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: http://arxiv.org/abs/1405.0312.

[27] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.

[28] Wesley J Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/118921efba23fc329e6560b27861f0c2-Paper.pdf.

[29] Dimity Miller et al. “Evaluating merging strategies for sampling-based uncertainty techniques in object detection”. In: *2019 international conference on robotics and automation (icra)*. IEEE. 2019, pp. 2348–2354.

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. “Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[31] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[32] Patrick Schäfer, Arik Ermshaus, and Ulf Leser. “ClaSP - Time Series Segmentation”. In: *CIKM*. 2021.

[33] Burr Settles. “Active learning literature survey”. In: (2009).

[34] David John Stracuzzi et al. “Quantifying Uncertainty to Improve Decision Making in Machine Learning.” In: (Oct. 2018). DOI: 10.2172/1481629. URL: https://www.osti.gov/biblio/1481629.

[35] Jakub Swiatkowski et al. *The k-tied Normal Distribution: A Compact Parameterization of Gaussian Mean Field Posteriors in Bayesian Neural Networks*. 2020. arXiv: 2002.02655 [cs.LG].

[36] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[37] Ryutaro Tanno et al. "Uncertainty modelling in deep learning for safer neuroimage enhancement: Demonstration in diffusion MRI". In: *NeuroImage* 225 (2021), p. 117366.

[38] Joost Van Amersfoort et al. "Uncertainty estimation using a single deep deterministic neural network". In: *International conference on machine learning*. PMLR. 2020, pp. 9690–9700.

[39] Wouter Van Gansbeke and Bert De Brabandere. "a simple latent diffusion approach for panoptic segmentation and mask inpainting". In: *arxiv preprint arxiv:2401.10227* (2024).

[40] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[41] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors". In: *arXiv preprint arXiv:2207.02696* (2022).

[42] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. *High-performance Semantic Segmentation Using Very Deep Fully Convolutional Networks*. 2016. arXiv: 1604.04339 [cs.CV].

[43] Wei Xia et al. "Semantic segmentation without annotating segments". In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 2176–2183.

[44] Bishan Yang et al. "Effective multi-label active learning for text classification". In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009, pp. 917–926.

[45] Zhi Zhang et al. *Bag of Freebies for Training Object Detection Neural Networks*. 2019. arXiv: 1902.04103 [cs.CV].

[46] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. "Objects as Points". In: *arXiv preprint arXiv:1904.07850*. 2019.