University of Münster

Department of Information Systems

Salient Object Detection for Social Media Images

Seminar Thesis

in the context of the seminar

More than meets the A-eye: Reflecting Human Vision in Artificial Intelligence

submitted by

Annika Terhörst, Dominik Eitner, and Simon Luttmann

Principal Supervisor	Prof. Dr. Christian Grimme
Supervisor	Lucas Stampe, M.Sc. Chair for Data Science: Computational Social Science & Systems Analysis
Student Candidate	Annika Terhörst, Dominik Eitner, and Simon Luttmann
Matriculation Number	527050 503468 513676
Field of Study	Information Systems
Contact Details	annika.terhoerst@uni-muenster.de deitner@uni-muenster.de sluttman@uni-muenster.de
Submission Date	5.12.2025

Contents

0.1 Introduction

The topic of image segmentation is a field of interest in computer vision since the 1970s and remains highly relevant to this day. Recent advances in segmentation models enable point-based, zero-shot segmentation. By combining real eye-tracking data with such a model, we aim to create a practical approach for salient object detection for social media images that reflects actual human visual behaviour. Using gaze data directly as prompts, we aim to derive data-driven object masks that represent what was looked at most. Beyond developing a segmentation pipeline, we are motivated to illustrate what kinds of insights become possible once salient objects in those social media images can be reliably extracted.

In existing literature, salient object detection has been implemented for specific use cases, but not within the social media context with its highly diverse and complex content. Therefore, research has not yet investigated what could be done with extracted salient objects in the social media context. Our work aims to provide an initial exploration of the research gap between social media, real human gaze behaviour and salient object detection.

0.2 Theoretical Background

0.2.1 Image Segmentation

Image segmentation is the process of dividing an image into different regions by grouping pixels and assigning each pixel a label. This step is an important part of many computer vision applications, such as detecting tumors in medical images or identifying pedestrians in autonomous driving. According to human visual perception, the identified regions are non-overlapping and meaningful - however, defining what exactly counts as a "meaningful" region can be difficult, as human perception is subjective and object boundaries are not always clear (Yu et al., 2023).

There are three common types of segmentation: *Semantic segmentation* assigns every pixel in an image a semantic label, such as "car" or "sky". *Instance segmentation* separates individual objects within the same class, for example distinguishing several people in one image. *Panoptic segmentation* combines both approaches by providing pixel-wise class labels and also identifying individual object instances.

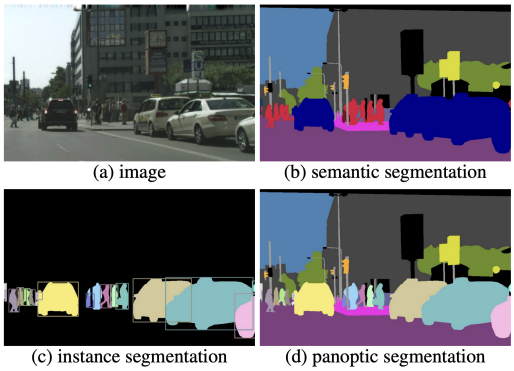


Figure 1 Types of image segmentation by Kirillov et al., 2019

Earlier approaches to image segmentation include algorithms such as k-means-clustering (Dhanachandra et al., 2015). Yet in recent years, deep learning models have significantly improved the segmentation effect and performance, therefore becoming the dominant method for solving segmentation tasks in complex environments (Minaee et al., 2022).

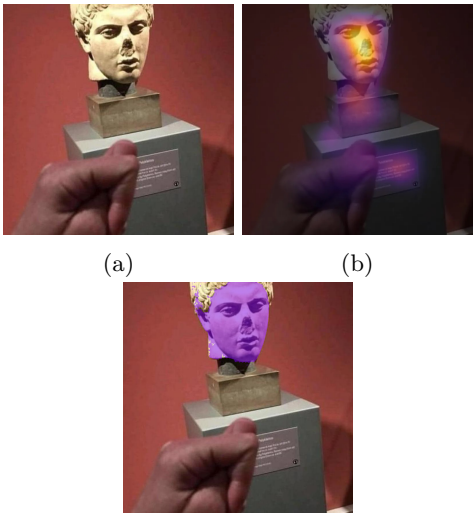
According to Zhou et al., 2024, the above-described image segmentation methods fall into the category of generic image segmentation (GIS). The category of promptable image segmentation (PIS) extends GIS by specifying the target to segment through a prompt. This prompt can have various forms such as text, box or points.

0.2.2 Salient Object Detection

The human visual system pays more attention to certain parts in an image, a property known as saliency. Inspired by this mechanism, *saliency detection* models aim to predict which regions in an image are most likely to attract human vi-

sual attention. These models typically provide saliency maps in form of heat maps, in which higher intensity values indicate regions detected to be more important (Ahmadi et al., 2018).

Salient Object Detection (SOD) – also referred to as salient object segmentation (Borji et al., 2019) or saliency segmentation (Kakanopas & Woraratpanya, 2021) – goes one step further by segmenting the most salient object(s) of an image. SOD can be interpreted as a two-stage process: 1) Detection of the most salient object and 2) Accurate segmentation of the region of that object. In contrast to general image segmentation, SOD focuses on segmenting only those objects that are (or that are predicted to be) most salient (Borji et al., 2019; T. Liu et al., 2011). Figure 2 illustrates the difference between saliency detection and salient object detection.



(a) (b)

(c)

Figure 2 (a) the original image, (b) saliency map (Alexander Kroner, 2025) and (c) salient object detection mask generated using SAM3 guided by the eye-tracking data

0.2.3 Image Segmentation Models

Table 1 offers a structured overview of prominent and state-of-the-art image segmentation models, organized according to the segmentation tasks for which they are most suitable.

For our use case, only promptable segmentation models are suitable, as the eye-tracking data provides point inputs that will be used to implement SOD. Among the identified promptable models, SAM and its successors are the only widely adopted models that natively support point-based prompts (Kirillov et al., 2023). In contrast, Grounded-SAM (Ren et al., 2024) and Florence-2 (Xiao et al., 2023) are limited to text prompts, and SEEM, whose last update dates back to 2023, is less actively maintained and less commonly used than SAM (Zou et al., 2023).

The Segment Anything Model (SAM) was developed by Meta AI and first introduced in mid-2023 (Kirillov et al., 2023). SAM performs object segmentation based on prompts, including points and bounding boxes. With more than 15,000 citations, SAM has become one of the de facto standards for domain-specific applications and is already employed in several specialized salient object detection settings, such as camouflage object segmentation and medical image segmentation (T. Chen et al., 2025), RGB-T SOD (Z. Liu et al., 2025), and text-driven SOD (Yuan et al., 2026). The most recent version, SAM 3, was released on 19 November 2025. Its rapid adoption - reaching over 5.3k GitHub stars within two weeks - indicates strong community interest (Meta Research, 2025) . Compared to previous versions, SAM 3 introduces the ability to detect and segment instances that match a given text description, and to further refine detections using visual examples (Carion et al., 2025).

Generic Image Segmentation		
Instance segmentation	**Semantic segmentation**	**Panoptic seg tation**
Mask R-CNN He et al., 2018	DeepLabV3 L.-C. Chen et al., 2017	Mask2Former Cheng et al., 2022
YOLOv11-seg Ultralytics, 2025	FCN Long et al., 2015	Panoptic FPN Kirillov et al., 2019
YOLACT Bolya et al., 2019	U-Net	Mask DINO Li et al., 2022
Mask2Former Cheng et al., 2022	SegFormer Xie et al., 2021	...
...	...	

Table 1 Overview of prominent image segmentation models categorized by segmentation task.

0.2.4 Evaluation of Image Segmentation Models

When evaluating image segmentation models, a distinction can be made between subjective and objective methods. Subjective evaluation involves a human assessing the quality of the segmentation results. Although this approach is convenient, the judgement may vary significantly between evaluators (Wang et al., 2020). Objective evaluation methods typically rely on comparing ground truth masks with the masks generated by the model on a pixel-based level. A commonly used metric is Intersection over Union (IoU), which measures the overlap between the prediction and the ground truth (e.g. Kirillov et al., 2023). Metrics commonly used in salient object detection include the F-measure, Precision-Recall and the Mean Absolute Error (MAE) (Borji et al., 2019).

0.3 Methodology

0.3.1 Experimental Design

0.3.2 Planned Practical Steps

The planned next steps can be divided into two main parts. The first part focuses on the optimization of the saliency object detection process, while the second part focuses on the extraction of insights from the segmented objects, based on an exploratory data analysis (Tukey, 1977).

As discussed in the previous section, the current approach offers several options for improvement. These include image preprocessing, the adjustment of threshold values, the use of box prompts instead of point prompts, and the exploration of additional clustering methods such as k means. Moreover, it is possible to add further steps between the main stages of the workflow in order to create more stable and consistent results. As a last possibility, the model can be fine-tuned, using existing scientific datasets such as WXSOD or PASCAL-S (CCVL, 2018; Quan et al., 2025), which already include saliency masks. The overall goal of this first step is to produce the best possible saliency masks for social media images. Better masks allow for a more reliable extraction of insights in the next step.

After the optimization step, the quality of the segmented masks must be evaluated. This evaluation includes three main questions. First, it must be examined if the masks truly represent the salient parts of the image. Second, the mask quality must be assessed by measuring how well the masks fit the expected important regions. Third, it must be analyzed whether certain types of images work better or worse with saliency based segmentation. This helps identify strengths and limitations of the approach for different kinds of social media content.

Furthermore, the generated masks allow for additional exploratory fields of analysis that can be grouped into three areas: user insights, content insights, and accessibility support. The user related area includes predictive user profiling, where salient regions help identify which visual elements attract individual users. The content related area includes a deeper understanding of social media images, the detection of clickbait content, and guidance for creators who work with under optimized images. The accessibility related area focuses on focus aware alternative text generation, where the masks help identify the most important visual elements for users with visual impairments. Together, these groups show the wider potential of the approach beyond the core segmentation step.

0.4 Results

0.4.1 Current Results

Regarding the first iteration of our Saliency Object Detection Pipeline, we have successfully segmented some types of social media images. Figure TODO shows an example of a social media image, along with its generated saliency mask. The mask highlights the teacher, the computer screen and the specific area on the sceen, which

does represent the salient objects in this image. Also other images, showing small groups of people or single persons have been segmented quite well so far.

However, there are also several limitations, which have been observed in the current results. First, images with very complex scenes or too many objects tend to produce less accurate masks (see Figure TODO). Second, images with text overlays led to only some letters being grouped into the salient region, instead of the full text block (see Figure TODO). Third, images with landscapes or generally without clear focal points were not segmented effectively (see Figure TODO). These limitations indicate that while the current pipeline shows promise, there is still room for improvement in handling a wider variety of social media images.

0.4.2 Expected Results for the Planned Steps

Regarding the planned optimization steps, we expect to see significant improvements in the quality of the saliency masks. By implementing image preprocessing techniques, we anticipate that noise and irrelevant details in the images will be reduced, leading to clearer segmentation results. The other discussed optimization techniques aim to further refine the segmentation process, making it more robust across different types of images. Overall, we strive to achieve more consistently accurate saliency masks, which will enhance the reliability of subsequent analyses.

Further analyses based on the optimized masks are expected to yield valuable insights into user behavior and content characteristics on social media platforms. For instance, by examining which visual elements are most frequently highlighted in the masks, we can infer what types of content are more engaging to users. These insights can inform content creation and moderation strategies. As the second part of our planned work is mostly experimental, pivots and adjustments of the goal and scope might be necessary, which would lead to different expected results.

0.5 Discussion

0.5.1 Limitations

During our research we identified several limitations that shape how the results should be interpreted. These limitations reflect both technical challenges and aspects related to the nature of social media content. First, social media images show a high level of complexity. They include a wide range of content, contexts and formats. This diversity makes it difficult to define what should be considered salient across different situations. It also reduces the generalisability of models that are trained on more uniform data sources.

Second, we encountered technical constraints. Modern transformer based models that support tasks such as image captioning are often very large and require considerable computational resources. In many practical settings these models cannot be used directly or must be simplified, which can reduce performance. Third, the ground truth data available in existing datasets reflects subjective human judgement. What is perceived as salient can differ between individuals, which introduces uncertainty into training and evaluation. Finally, there are currently no datasets designed specifically for salient object detection in the context of social media. Most available datasets are collected in more controlled environments and do not capture the characteristics of social media content. This limits the ability to develop and evaluate models that address this specific setting.

0.5.2 Future Research

Future research can build on the findings of this seminar and explore areas that lie beyond its scope. One important direction is the development of datasets that focus specifically on social media content. Such datasets would give

models the opportunity to learn from examples that reflect the diversity, style and context that are characteristic of images shared online. Another promising direction is the investigation of lightweight models for social media image analysis. Since many existing approaches rely on large and resource intensive architectures, identifying smaller and more efficient models could improve accessibility and practical use.

Further work may also examine how multimodal data can support salient object detection. Textual elements such as image captions can provide additional context and may enhance model performance when integrated with visual information. Moreover, researchers could explore new applications of salient object detection within social media analysis. This may include areas such as content understanding, user behaviour studies and automated moderation, where identifying salient elements may enable deeper insights and more effective tools.

0.6 Conclusion

- We explored the topic of Salient Object Detection, specifically in the context of Social Media Images.
- We reviewed existing methods and models for Salient Object Detection.
- We identified the unique challenges posed by Social Media Images, such as diverse content and formats.
- We implemented a first prototype to segment salient objects in social media images using a zero-shot approach.
- We discussed the limitations of our approach and proposed next steps for our research.

A Appendix

TODO: Add result pictures and/or our code here

Bibliography

Ahmadi, M., Hajabdollahi, M., Karimi, N., & Samavi, S. (2018, January). Context aware saliency map generation using semantic segmentation [arXiv:1801.00256 [cs]]. https://doi.org/10.48550/arXiv.1801.00256

Alexander Kroner. (2025, March). Visual Saliency Prediction - a Hugging Face Space by alexanderkroner. Retrieved December 2, 2025, from https://huggingface.co/spaces/alexanderkroner/saliency

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019, October). YOLACT: Real-time Instance Segmentation [arXiv:1904.02689 [cs]]. https://doi.org/10.48550/arXiv.1904.02689

Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. *Computational Visual Media*, *5*(2), 117–150. https://doi.org/10.1007/s41095-019-0149-9

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., ... Feichtenhofer, C. (2025, November). SAM 3: Segment Anything with Concepts [arXiv:2511.16719 [cs]]. https://doi.org/10.48550/arXiv.2511.16719

CCVL. (2018). Pascal-S. https://ccvl.jhu.edu/datasets/

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017, December). Rethinking Atrous Convolution for Semantic Image Segmentation [arXiv:1706.05587 [cs]]. https://doi.org/10.48550/arXiv.1706.05587

Chen, T., Cao, R., Yu, X., Zhu, L., Ding, C., Ji, D., Chen, C., Zhu, Q., Xu, C., Mao, P., & Zang, Y. (2025, November). SAM3-Adapter: Efficient Adaptation of Segment Anything 3 for Camouflage Object Segmentation, Shadow Detection, and Medical Image Segmentation [arXiv:2511.19425 [cs]]. https://doi.org/10.48550/arXiv.2511.19425

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R. (2022, June). Masked-attention Mask Transformer for Universal Image Segmentation [arXiv:2112.01527 [cs]]. https://doi.org/10.48550/arXiv.2112.01527

Dhanachandra, N., Manglem, K., & Chanu, Y. J. (2015). Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm. *Procedia Computer Science*, *54*, 764–771. https://doi.org/10.1016/j.procs.2015.06.090

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018, January). Mask R-CNN [arXiv:1703.06870 [cs]]. https://doi.org/10.48550/arXiv.1703.06870

Kakanopas, D., & Woraratpanya, K. (2021). A Unified Framework for Saliency Segmentation. In P. Meesad, D. S. Sodsee, W. Jitsakul, & S. Tangwannawit (Eds.), *Recent Advances in Information and Communication Technology 2021* (pp. 191–200). Springer International Publishing. https://doi.org/10.1007/978-3-030-79757-7_19

Kirillov, A., He, K., Girshick, R., Rother, C., & Dollár, P. (2019). Panoptic Segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9396–9405. https://doi.org/10.1109/CVPR.2019.00963

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023, April). Segment Anything [arXiv:2304.02643 [cs]]. https://doi.org/10.48550/arXiv.2304.02643

Li, F., Zhang, H., xu, H., Liu, S., Zhang, L., Ni, L. M., & Shum, H.-Y. (2022, December). Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation [arXiv:2206.02777 [cs]]. https://doi.org/10.48550/arXiv.2206.02777

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2011). Learning to Detect a Salient Object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(2), 353–367. https://doi.org/10.1109/TPAMI.2010.70

Liu, Z., Wang, X., Fang, X., Tu, Z., & Wang, L. (2025, October). SAMSOD: Rethinking SAM Optimization for RGB-T Salient Object Detection [arXiv:2510.03689 [cs]]. https://doi.org/10.48550/arXiv.2510.03689

Long, J., Shelhamer, E., & Darrell, T. (2015, March). Fully Convolutional Networks for Semantic Segmentation [arXiv:1411.4038 [cs]]. https://doi.org/10.48550/arXiv.1411.4038

Meta Research. (2025, December). Facebookresearch/sam3 [original-date: 2025-07-17T16:15:40Z]. Retrieved December 5, 2025, from https://github.com/facebookresearch/sam3

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3523–3542. https://doi.org/10.1109/TPAMI.2021.3059968

Quan, C., Xiong, Y., Rongfeng, L., Qianyu, Z., Yu, L., Xiaofei, Z., & Bolun, Z. (2025). WXSOD: A Benchmark for Robust Salient Object Detection in Adverse Weather Conditions. https://arxiv.org/abs/2508.12250

Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., & Zhang, L. (2024, January). Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks [arXiv:2401.14159 [cs]]. https://doi.org/10.48550/arXiv.2401.14159

Tukey, J. W. (1977). Exploratory data analysis. *Reading/Addison-Wesley*.

Ultralytics. (2025). YOLO Docs Instance Segmentation. Retrieved December 6, 2025, from https://docs.ultralytics.com/tasks/segment/

Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: A survey of methods. *Artificial Intelligence Review*, *53*(8), 5637–5674. https://doi.org/10.1007/s10462-020-09830-9

Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., & Yuan, L. (2023, November). Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks [arXiv:2311.06242 [cs]]. https://doi.org/10.48550/arXiv.2311.06242

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021, October). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers [arXiv:2105.15203 [cs]]. https://doi.org/10.48550/arXiv.2105.15203

Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., Yang, T., & Gao, M. (2023). Techniques and Challenges of Image Segmentation: A Review [Publisher: Multidisciplinary Digital Publishing Institute]. *Electronics*, *12*(5), 1199. https://doi.org/10.3390/electronics12051199

Yuan, Y., Zhang, Y., Zhang, S., & Wang, H. (2026). SaliencyCLIP-SAM: Bridging Text and Image Towards Text-Driven Salient Object Detection. In Z. Lin, L. Wang, Y. Jiang, X. Wang, S. Liao, S. Shan, R. Liu, J. Dong, & X. Yu (Eds.), *Image and Graphics* (pp. 29–41). Springer Nature. https://doi.org/10.1007/978-981-95-3393-0_3

Zhou, T., Xia, W., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., & Cremers, D. (2024, November). Image Segmentation in Foundation Model Era: A Survey [arXiv:2408.12957 [cs]]. https://doi.org/10.48550/arXiv.2408.12957

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., & Lee, Y. J. (2023, July). Segment Everything Ev-

erywhere All at Once [arXiv:2304.06718 [cs]]. https://doi.org/10.48550/arXiv.2304.06718

Declaration of Authorship

We hereby declare that, to the best of our knowledge and belief, this thesis titled *Salient Object Detection for Social Media Images* is our own, independent work. We confirm that each significant contribution to and quotation in this thesis that originates from the work or works of others is indicated by proper use of citation and references; this also holds for tables and graphical works.

Münster, 5.12.2025

Signatures

Consent Form

Name: Annika Terhörst, Dominik Eitner, and Simon Luttmann
Title of Thesis: Salient Object Detection for Social Media Images

What is plagiarism? Plagiarism is defined as submitting someone else's work or ideas as your own without a complete indication of the source. It is hereby irrelevant whether the work of others is copied word by word without acknowledgment of the source, text structures (e.g. line of argumentation or outline) are borrowed or texts are translated from a foreign language.

Use of plagiarism detection software. The examination office uses plagiarism software to check each submitted bachelor and master thesis for plagiarism. For that purpose the thesis is electronically forwarded to a software service provider where the software checks for potential matches between the submitted work and work from other sources. For future comparisons with other theses, your thesis will be permanently stored in a database. Only the School of Business and Economics of the University of Münster is allowed to access your stored thesis. The student agrees that his or her thesis may be stored and reproduced only for the purpose of plagiarism assessment. The first examiner of the thesis will be advised on the outcome of the plagiarism assessment.

Sanctions Each case of plagiarism constitutes an attempt to deceive in terms of the examination regulations and will lead to the thesis being graded as "failed". This will be communicated to the examination office where your case will be documented. In the event of a serious case of deception the examinee can be generally excluded from any further examination. This can lead to the exmatriculation of the student. Even after completion of the examination procedure and graduation from university, plagiarism can result in a withdrawal of the awarded academic degree.

We confirm that we have read and understood the information in this document. We agree to the outlined procedure for plagiarism assessment and potential sanctioning.

Münster, 5.12.2025

Signatures