Task-driven Sensing with Coarse-to-Fine Glimpse-based Active Perception

Oleh Kolner, Thomas Ortner, Stanisław Woźniak & Angeliki Pantazi
IBM Research Europe - Zurich
Zurich, Switzerland
olk@zurich.ibm.com

Abstract

Modern computer vision models commonly rely on passive sensing and process images in their entirety. Lacking the ability to zoom-in to task-relevant regions for detailed analysis, this approach becomes limited for high-resolution, cluttered scenes where only a small area is relevant for the task at hand. A particularly challenging problem in this context is instance detection that involves localizing specific object instances given a few visual examples. We introduce a novel active sensing model that uses a brain-inspired coarse-to-fine strategy to glimpse over the image by steering a retina-like sensor. The sensor uses a log-polar pixel layout that facilitates precise localization of task-relevant regions. Our model can be integrated with various state-of-the-art instance detectors. It improves their performance by up to 90%, making even small models developed for edge-devices, perform on par or even better than their large counterparts. In light of these performance gains, our model can become a complementary part in sensor hardware designs enabling active, task-driven sensing.

1 Introduction

Modern computer vision systems predominantly rely on passive sensing so that entire images are processed holistically and uniformly without the ability to selectively zoom-in to task-relevant regions for more detailed analysis. This poses significant challenges when dealing with high-resolution images, where only a small portion may be relevant to the task [1, 2]. Moreover, state-of-the-art foundational vision models, such as SAM [3], DINOv2 [4], and their extensions, operate on images of a fixed size, therefore requiring high-resolution images to be downscaled. However, downscaling often renders critical details nearly invisible, particularly limiting detection of small objects or distinguishing objects in cluttered scenes [5].

In contrast, the human visual system processes scenes actively and sequentially, using eye movements to focus on task-relevant regions [6]. Experimental evidence suggests that these movements, known as saccades, follow a coarse-to-fine strategy: macrosaccades are guided by low-resolution features to promising areas, followed by microsaccades for detailed inspection [7–9]. This strategy is supported by the structure of the human retina, that features a non-uniform visual resolution – high in its center and decreasing toward the periphery. This design balances the need for detailed visual information with a broad field of view [10, 11] and has inspired development of specialized hardware for visual sensors [12, 13]. However, such hardware is typically designed without software models to steer its high-resolution regions to task-relevant regions.

Several approaches took a high-level inspiration from the saccadic processing, and showed compelling results across various tasks such as image classification [14, 15], object detection [16], visual exploration [17] and visual reasoning [18]. Glimpse-based active perception (GAP) [19] took additional inspiration from the human vision. In particular, it leverages the concept of a saliency map

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning to Sense.

[20] to glimpse at the most important image parts, showing compelling results in out-of-distribution generalization and sample efficiency for visual reasoning. However, the aforementioned approaches have not explored the coarse-to-fine strategy of the saccadic processing, critical for dealing with high-resolution images and adapting the sensing to an externally provided task or goal.

Following the GAP approach, we propose a novel sensing model that uses the coarse-to-fine strategy to steer a retina-inspired log-polar sensor to search for task-relevant regions. Referred to as coarse-to-fine GAP (CF-GAP), our system does not process the entire image at once, but rather uses coarse glimpsing to find task-relevant regions, which are further refined by fine glimpsing. In this way, CF-GAP serves as a front-end for standard vision models delivering more targeted visual information to enhance their performance. In this work, we consider the problem of instance detection where the goal is to detect a specific instance of an object in a high-resolution scene given its few visual examples [5]. The visual examples guide the CF-GAP making thereby our sensing approach task-adaptive. The results show that CF-GAP front-end substantially improves performance of state-of-the-art vision models. Moreover, integrating it with small models enables them to outperform the larger ones making our approach well-suited to resource-constrained applications. Generally, CF-GAP can be seen as a control unit for sensor hardware that directs it toward task-relevant regions. Hence, our work can support a sensor-model co-design for active, task-driven sensing.

2 Model

The overall architecture (Figure 1) involves a hierarchical process guided by coarse and fine search maps, which highlight candidate regions where the search target is likely to be.

The sensing process begins with the generation of a coarse search map from a downscaled version of the scene launching a nested loop of coarse and fine glimpsing. Each coarse glimpse location is generated using a winner-takes-all (WTA) strategy followed by inhibition-of-return (IoR) to avoid revisiting the same areas, similar to [19, 20]. Each coarse glimpse location initiates a fine-glimpsing loop, where the log-polar sensor is steered to analyze the region around the coarse glimpse location in more detail, thereby compensating for the coarse search map's limited resolution. The sensor steering relies on the fine search map that is computed based on the log-polar visual content of each fine glimpse. The next fine glimpse location is determined as a 2D centroid of the fine search map, that is normalized by applying softmax along its spatial dimensions. At the end of each fine-glimpsing loop, a fixed-sized crop is cut from the original scene and passed to the downstream architecture to detect a candidate object and match it with the search target. After all glimpses are processed, the best-matched candidate object is returned. Our approach is independent of the downstream architecture, and during evaluation we employ selected state-of-the-art architectures, as described in Section 3.

Coarse and fine search maps are generated by comparing features extracted from the search target with features extracted from the down-scaled scene and the log-polar image, respectively. We employ a lightweight convolutional neural network (CNN) to extract feature maps of the scene and the search target. While for the coarse search map we use a pre-trained MobileNet-V3 [21], for the fine search

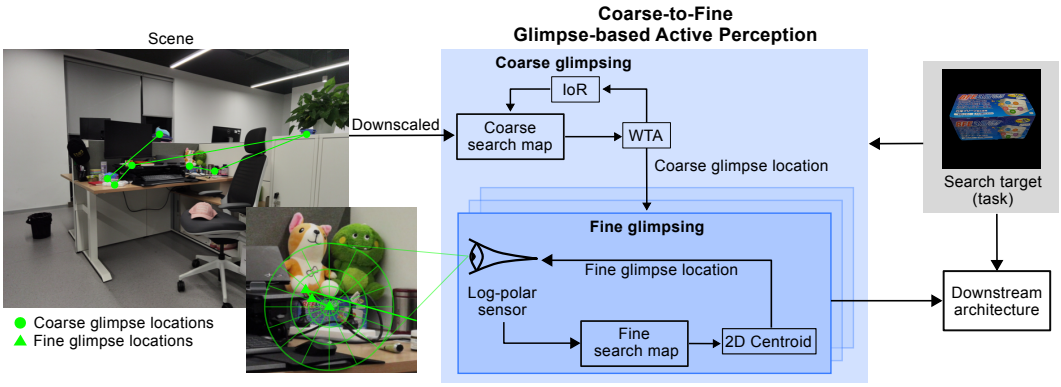

Figure 1: CF-GAP identifies candidate regions, passed to the downstream architecture for recognition.

map we train a custom CNN with deformable convolutions [22]. Deformable convolutions adapt the receptive field based on both visual content and spatial coordinates in log-polar space, which is crucial for handling the non-uniform resolution of log-polar images. The network is trained from scratch using Cut-and-Paste framework [23]: search targets are pasted onto random backgrounds, and the pasted regions serve as binary segmentation labels. The details about the search maps and the log-polar sensor are provided in Appendix A.

3 Experiments

Instance detection task. Given a few visual examples from different viewpoints of an object (referred to as a search target), the task is to localize it in a scene. Unlike classical object detection, which identifies all instances of a category, this task focuses on detecting a particular object instance. We consider HR-InsDet dataset [5] that contains 100 object instances with 24 visual examples per instance and 160 high-resolution scenes from 14 indoor scenarios. For training, HR-InsDet dataset provides 200 images with random background to synthesize a training dataset following the cut-paste-learn strategy [23] where search targets are resized and pasted onto random background. The dataset is divided into subsets based on clutter and occlusion levels—*easy* and *hard*—and target sizes—*small*, *medium*, and *large*. Following the HR-InsDet evaluation protocols we report the performance for each of those subsets. The evaluation metrics are average precision (AP) at Intersection-over-Union (IoU) thresholds from 0.5 to 0.95 with the step size 0.05 and AP_{50} with IoU threshold of 0.5.

Baselines. The strongest baselines, OTS-FM [5] and IDOW [24], use a pre-trained object detector to process the entire scene identifying bounding boxes of all object-looking parts. This is followed by a feature extractor that generates embeddings of the detected proposals for matching with the search target's examples using Stable Matching [25]. Variants of these approaches differ in their choice of proposal detectors and feature extractors. Both baselines use either SAM [3] or GroundingDINO [26] for proposal detection and DINOv2 [4] for feature extraction. The only difference between OTS-FM and IDOW is that the latter employs a modified version of DINOv2 fine-tuned on objects from HR-InsDet dataset. Since we study how our CF-GAP front-end enhances baseline models, we exclude IDOW models from our evaluations due to unavailable fine-tuned weights of DINOv2. Hence, the considered baselines include OTS-FM_{SAM}, OTS-$FM_{GroundingDINO}$ depending on the proposal detector. In addition, we consider OTS-$FM_{MobileSAM}$ that uses a lightweight version of SAM, known as MobileSAM [27]. Since all baselines differ only by their proposal detectors, we refer to them simply as GroundingDINO, SAM, and MobileSAM.

Setup. We integrate our CF-GAP with each of the baselines, which become downstream recognition modules. When integrating with the MobileSAM and SAM baselines, CF-GAP introduces an important difference in the input provided to them. Specifically, standalone baselines receive a complete high-resolution input scene along with a coarse grid of 2D locations. This grid serves as a prompt to detect all possible object proposals at each of its locations. By contrast, CF-GAP provides the baselines only with a small scene crop along with a single fine glimpse location as a prompt at the end of each fine glimpsing loop. This enhances efficiency and reduces distraction. The GroundingDINO baseline does not allow for location-based prompting, therefore it receives only the small scene crops when combined with CF-GAP.

Original scenes are sized at 6144×8192 pixels, but baseline models require resizing and padding to 1024×1024. CF-GAP supports flexible input sizes; for faster experimentation, scenes are resized to 4096×5460. The coarse search map extractor processes scenes downscaled by a factor of 5. The radius of the log-polar sensor (Eq. 1 in Appendix A) is set to 1024 and the fine search map extractor receives the log-polar images resized to 402×128. Each fine glimpsing loop ends with a 600×600 crop passed to the downstream architecture. Each scene undergoes 30 coarse glimpses, with 7 fine glimpses per coarse one.

4 Results

As shown in Figure 2, CF-GAP consistently boosts the performance of baseline models, especially in cluttered scenes and scenes with small-sized objects demonstrating the effectiveness of the coarse-to-fine glimpsing. In addition, we measured that CF-GAP passes on average only 25% of the

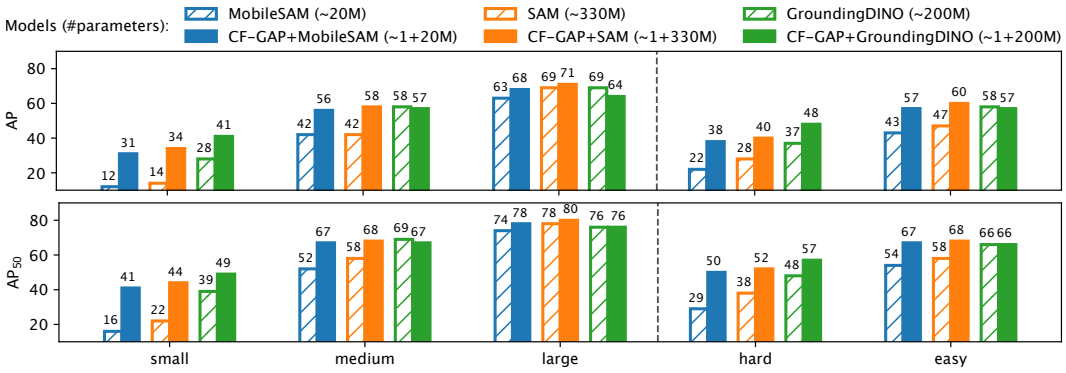

Figure 2: Performance across two dataset groupings: by search target sizes, and by scene types.

entire scene area to the baseline models (in form of crops) which is a significant reduction of irrelevant and potentially distracting information. Only for GroundingDINO, in few cases of low-clutter or large-target scenarios, its CF-GAP-extension slightly underperforms. The reason is that GroundingDINO tends to detect objects as well as their parts and the latter may be better matched with the search target examples due to imperfect features extracted by DINOv2. MobileSAM, the smallest baseline, benefits the most from CF-GAP, surpassing both SAM and GroundingDINO baselines, on difficult subsets and matching it on simpler ones.

Ablations. To assess the utility of the coarse-to-fine glimpsing, we evaluated the impact of the fine glimpsing stage. Table 1 reports how often at least one of the glimpse locations is on average inside the ground truth bounding box. Fine glimpsing proves crucial in difficult cases of cluttered scenes and small objects. This is attributed to the low resolution of the coarse search map that may produce glimpse locations slightly off from the actual object. The fine glimpsing corrects for this error.

Table 1: Percentage of times where at least one glimpse location corresponded to the search target.

	small	medium	large	easy	hard
w/o fine glimpsing	70%	87%	90%	86%	75%
with fine glimpsing	79%	91%	92%	89%	84%

To demonstrate the importance of the proposed CF-GAP, we conducted an additional study. Specifically, since the baselines cannot process input scenes in their original resolution, a naive solution would be to split them into smaller non-overlapping patches, and let the baselines process each patch separately. However, as shown in Table 2, this patched approach yields inferior results, underscoring the effectiveness of CF-GAP's task-driven selection strategy.

Table 2: Patched baselines

Model	AP	
	small	hard
MobileSAM	12	22
Patched MobileSAM	18	23
CF-GAP+MobileSAM	31	38
SAM	14	28
Patched SAM	19	24
CF-GAP+SAM	34	40
GroundingDINO	28	37
Patched GroundingDINO	22	28
CF-GAP+GroundingDINO	41	48

5 Conclusion

We presented a novel active sensing model that leverages a coarse-to-fine glimpsing process inspired from human vision. Our results demonstrate that our approach significantly improves the performance of common foundational models when applied to high-resolution images – particularly in scenarios characterized by visual clutter and small objects. Most foundational vision models are constrained by fixed input sizes. Our approach not only overcomes this bottleneck but also opens avenues for co-designing new sensor and model architectures optimized for active sensing. In this early study, we focused on the instance detection problem, which serves as an effective testbed for task-adaptive processing, where the task is defined by the search target. A natural extension of this work would involve generalizing the search target, for example, using textual prompts instead of specific visual examples. Our approach lays the groundwork for addressing such generalized tasks in future research.

References

[1] P. Wu and S. Xie, "V?: Guided visual search as a core mechanism in multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[2] J. Zhang, M. Khayatkhoei, P. Chhikara, and F. Ilievski, "MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs," in *The Thirteenth International Conference on Learning Representations*, 2025.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[5] Q. Shen, Y. Zhao, N. Kwon, J. Kim, Y. Li, and S. Kong, "A high-resolution dataset for instance detection with multi-view object capture," *Advances in Neural Information Processing Systems*, vol. 36, pp. 42064–42076, 2023.

[6] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in cognitive sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[7] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.

[8] X. Yu, Z. Zhou, S. I. Becker, S. E. Boettcher, and J. J. Geng, "Good-enough attentional guidance," *Trends in Cognitive Sciences*, vol. 27, no. 4, pp. 391–403, 2023.

[9] M. Rolfs, "Microsaccades: small steps on a long way," *Vision research*, vol. 49, no. 20, pp. 2415–2441, 2009.

[10] E. L. Schwartz, "Spatial mapping in the primate sensory projection: analytic structure and relevance to perception," *Biological cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[11] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, pp. 1195–1201, 2011.

[12] V. Javier Traver and A. Bernardino, "A review of log-polar imaging for visual perception in robotics," vol. 58, no. 4, pp. 378–398, 2010.

[13] F. Faramarzi, B. Linares-Barranco, and T. Serrano-Gotarredona, "A 128$\times$ 128 electronically multi-foveated dynamic vision sensor with real-time resolution reconfiguration," *IEEE Access*, 2024.

[14] V. Mnih, N. Heess, A. Graves, and k. kavukcuoglu, "Recurrent Models of Visual Attention," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2048–2057, PMLR, 07–09 Jul 2015.

[16] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[17] A. Pardyl, M. Wronka, M. Wołczyk, K. Adamczewski, T. Trzciński, and B. Zieliński, "Adaglimpse: Active visual exploration with arbitrary glimpse position and scale," in *European Conference on Computer Vision*, pp. 112–129, Springer, 2025.

[18] S. Woźniak, H. Jónsson, G. Cherubini, A. Pantazi, and E. Eleftheriou, "On the visual analytic intelligence of neural networks," *Nature Communications*, vol. 14, no. 1, p. 5978, 2023.

[19] O. Kolner, T. Ortner, S. Woźniak, and A. Pantazi, "Mind the GAP: Glimpse-based active perception improves generalization and sample efficiency of visual reasoning," in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," vol. 20, no. 11, pp. 1254–1259, 1998. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[21] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

[22] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017.

[23] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 1301–1310, 2017.

[24] Q. Shen, Y. Zhao, N. Kwon, J. Kim, Y. Li, and S. Kong, "Solving instance detection from an open-world perspective," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9901–9910, 2025.

[25] D. G. McVitie and L. B. Wilson, "The stable marriage problem," *Communications of the ACM*, vol. 14, no. 7, pp. 486–490, 1971.

[26] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*, pp. 38–55, Springer, 2024.

[27] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S.-H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[28] C. Esteves, C. Allen-Blanchette, X. Zhou, and K. Daniilidis, "Polar transformer networks," *arXiv preprint arXiv:1709.01889*, 2017.

Appendix

A Model details

Coarse and fine search maps. The coarse and fine search maps are generated by correlating visual features extracted from the search target with features extracted from the input image, see Figure 3. The input image for the coarse map is a downscaled version of the original scene where the search target has to be found. The input image for the fine map is a log-polar image extracted from the original scene around a glimpse location.

We employ a lightweight CNN to produce feature maps both for the scene and the search target. The feature maps of the search target are averaged over the spatial dimension forming a convolutional kernel. This kernel is normalized and convolved over the normalized scene feature maps producing the search map. For the coarse search map, we employ a pre-trained MobileNet-V3 [21] as a feature extractor. For the fine search map we use a custom CNN with deformable convolutions [22]. Deformable convolutions adapt the receptive field based on both visual content and spatial coordinates in log-polar space, which is crucial for handling the non-uniform resolution of log-polar images. The network consists of one deformable convolution layer with receptive size of (15, 9) and stride of (4, 2) followed by four 1×1 convolution layers. The hidden size of all layers is 256. We train this network from scratch using Cut-and-Paste framework [23]: search targets are pasted onto random backgrounds, and the pasted regions serve as binary segmentation labels. The log-polar images are produced by applying the log-polar sensor at randomly sampled locations. The training is done with a combination of Dice and Focal losses in ratio 1:20.

Log-polar sensor for fine glimpsing. The log-polar sensor samples pixels based on the log-polar layout around the glimpse location, oversampling regions that are closer to the glimpse location and undersampling ones that are farther. More specifically, given an image I of size $H \times W$ and a glimpse location (x_i, y_i), the sensor samples pixels from the image I according to the log-polar coordinate transform as in [28]:

$$x_i^s = x_i + e^{\log(r)x_i^t/W}\cos(\frac{2\pi y_i^t}{H}), \quad y_i^s = y_i + e^{\log(r)x_i^t/W}\sin(\frac{2\pi y_i^t}{H}) \tag{1}$$

where (x_i^s, y_i^s) denote the sampled points from (x_i^t, y_i^t) coordinates in the image I, and r is a hyper-parameter that defines the radius of the region around the glimpse location in I from which the pixels are to be sampled. Compared to the cartesian grid-based layout, the log-polar design offers a better resolution-field of view balance. In particular, it magnifies regions near glimpse locations facilitating fine search of the search target.

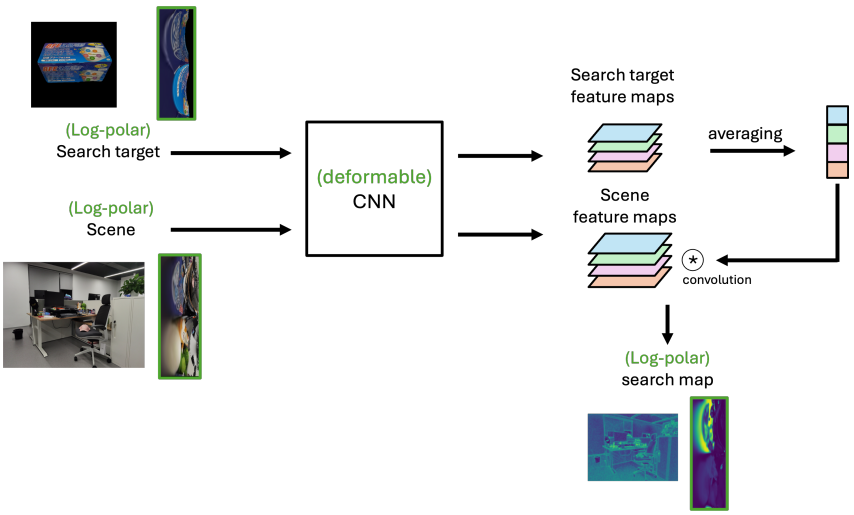

Figure 3: Search map extraction. Green text and frames highlight the difference between coarse and fine search map extraction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We evaluate our approach on the specified in the abstract task in Section 4 and compare our method to relevant approaches in the literature.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we discuss the limitation of the scope of our work in Section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: we do not provide any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided detailed description of the most important aspects of the model. Although given the limited size of the workshop submission some small details may be missing but we will be open to share them upon request.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use publicly available dataset and intend to release the code after we submit the work to the main track of another upcoming conference.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes] .

   Justification: we specify the most improtant details in Section 3

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No] .

   Justification: because of computational constraints caused by high-resolution images we were not able to provide multiple seeds for all results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: given the limited size of the workshop submission, we considered these details as unimportant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the code of ethics and comply.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work is at a too early stage to have any

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: we cite the works of the authors whose publicly available dataset we use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are introduced.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing or research with human subjects was conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: this research does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The development of our approach did not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.