Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Demonstrations Track

Probabilistic Feature Matching for Fast Scalable Visual Prompting

Thomas Frick*, **Cezary Skura***, **Filip M. Janicki***, **Roy Assaf**, **Niccolo Avogaro**, **Daniel Caraballo**, **Yagmur G. Cinar**, **Brown Ebouky**, **Ioana Giurgiu**, **Takayuki Katsuki**, **Piotr Kluska**, **Cristiano Malossi**, **Haoxiang Qiu**, **Tomoya Sakai**, **Florian Scheidegger**, **Andrej Simeski**, **Daniel Yang**, **Andrea Bartezzaghi**† and **Mattia Rigotti**†
IBM Research
{abt, mrg}@zurich.ibm.com

Abstract

In this work, we propose a novel framework for image segmentation guided by visual prompting, which leverages the power of vision foundation models. Inspired by recent advancements in computer vision, our approach integrates multiple large-scale pretrained models to address the challenges of segmentation tasks with limited and sparsely annotated data interactively provided by a user. Our method combines a frozen feature extraction backbone with a scalable and efficient probabilistic feature correspondence (soft matching) procedure derived from Optimal Transport to couple pixels between reference and target images. Moreover, a pretrained segmentation model is harnessed to translate user scribbles into reference masks and matched target pixels into output target segmentation masks. This results in a framework that we name *Softmatcher*, a versatile and fast training-free architecture for image segmentation by visual prompting. We demonstrate the efficiency and scalability of Softmatcher for real-time interactive image segmentation by visual prompting and showcase it in diverse visual domains, including technical visual inspection use cases.

1 Introduction

Foundation Models ushered in a significant shift in how machine learning models are developed and deployed, pivoting from a paradigm centered on training use case-tailored models on task-specific data to a paradigm where single generalist models are pretrained on diverse large-scale data, then fine-tuned for a wide range of tasks [Bommasani *et al.*, 2022]. Specifically in computer vision, models such as SAM [Kirillov *et al.*, 2023], CLIP [Radford *et al.*, 2021], and self-supervised backbones such as DINO [Caron *et al.*, 2021] and DINOv2 [Oquab *et al.*, 2023] have unlocked powerful and versatile visual functionalities like object detection, semantic segmentation and expressive embeddings that are at the core of a multitude of diverse applications. In particular, the possibility of using and combining these models in novel ways to address specific challenges in applied computer vision has been a topic of recent interest, including as a means to design new workflows in technical domains such as visual inspection (see e.g. [Rigotti *et al.*, 2023]).

In this work we take inspiration from the recent advancements driven by the approach of compositionally combining multiple Foundation Models to address sophisticated computer vision tasks. Specifically, we focus on the problem of image segmentation, which is a fundamental task in computer vision with a wide range of applications, including medical imaging, autonomous driving, and visual inspection, with a particular focus in developing a human-computer interaction workflow to facilitate open-world segmentation of images by visual prompting through sparse user annotations. For that, we largely build upon a previous architecture named *Matcher*, which was designed to perform training-free few-shot segmentation using *in-context examples* by means of off-the-shelf vision Foundation Models [Liu *et al.*, 2023]. Our framework enhances this approach's interactivity in two crucial ways: 1) we integrate a pretrained segmentation model to translate user scribbles on a representative sample of the object class to be segmented into reference masks, which are then passed to the few-shot segmentation architecture; 2) we develop a scalable probabilistic feature soft-matching procedure whose efficiency and low-latency allows us to embed few-shot segmentation in a real-time interactive workflow.

2 Related Work

The **Segment Anything Model (SAM)** [Kirillov *et al.*, 2023] has popularized the prompting paradigm in computer vision by enabling fine-grained image segmentation through interactive prompts in the form of points and/or bounding boxes.

Both **Visual Prompting via Inpainting** [Bar *et al.*, 2022] and **SegGPT/Painter** [Wang *et al.*, 2023] presented visual prompting models trained on few-shot image segmentation datasets. These models operate on a reference image and corresponding segmentation masks and generate a segmentation mask for a target image based on the reference.

[Zhang *et al.*, 2023] introduced a training-free method for one-shot segmentation leveraging pretrained image encoders in conjunction with SAM. The labeled pixels within the annotated mask on a reference image are assigned to pixels on target images thanks to a cosine similarity matrix of their corresponding encoded patches. The target patch of maximum

†corresponding authors

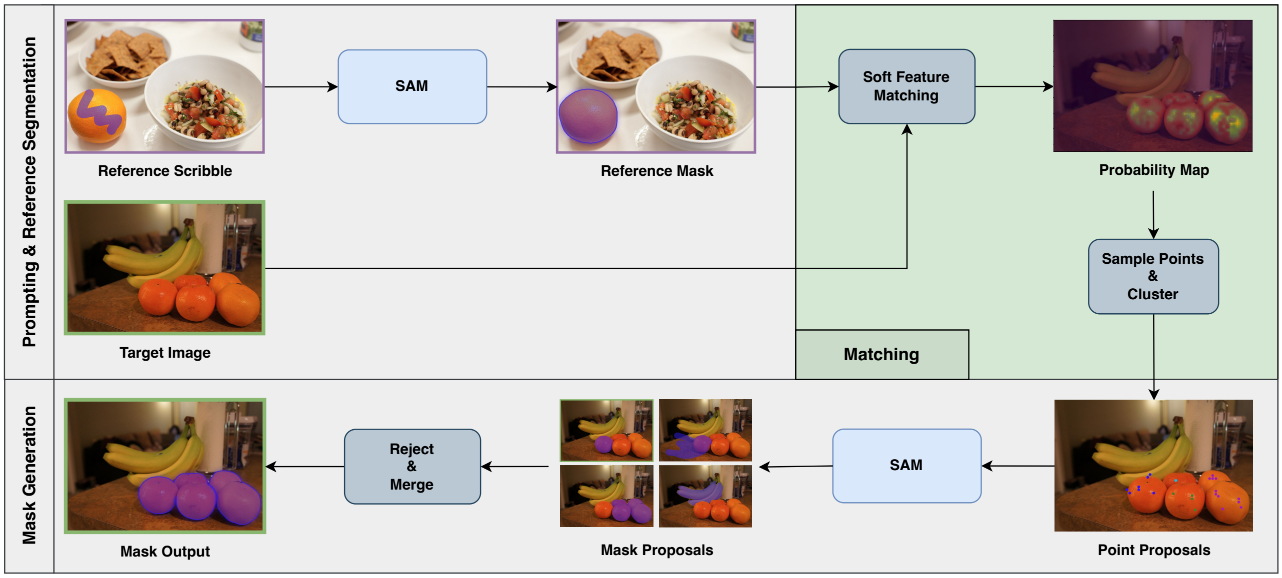

Figure 1: **Visual Prompting Framework:** 1) *Prompting & reference segmentation:* Coarse user annotations (scribbles) are converted to reference segmentation mask using SAM. 2) *Matching:* Image features are extracted using DINOv2 from reference and target images. The feature patches within the reference mask are matched to all patches in the target through our probabilistic matching procedure, resulting in a probability map over target images. This is sampled to obtain points, which are then clustered. 3) *Mask generation:* For each cluster, the respective points are passed to SAM to generate mask proposals. Each mask proposal is scored and discarded based on SAM-predicted IOU or merged into the final output mask.

similarity is then utilized by SAM to generate a segmentation mask for the target object.

[Gupta and Kembhavi, 2022] presented a neuro-symbolic approach for solving complex visual tasks given natural language instructions by leveraging the in-context learning ability of LLMs to generate modular programs that combine pretrained models leveraging their *compositionality*, a feature that has received recent interest for enabling flexible generalization (see e.g. [Ito *et al.*, 2022]).

[Liu *et al.*, 2023] introduced **Matcher**, an approach that uses a bidirectional matching procedure to match the encoded reference and target image patches using the Hungarian algorithm, an accurate but slow assignment algorithm with worst-case complexity cubic in the size of the problem [Crouse, 2016]. Similarly to [Zhang *et al.*, 2023], one-shot (or few-shot) segmentation is implemented by assigning annotated encoded pixels on reference images to encoded target pixels, which then serve as prompts for SAM to produce segmentation mask proposals on the target images. The set of mask proposals is finally scored and either accepted or rejected.

[Janouskova *et al.*, 2023] proposed a framework for model-assisted labeling of visual inspection defects through an interactive annotation process leveraging gradient-based explainability to improve the efficiency of the provided labels.

3 Visual Prompting Framework

System architecture. Figure 1 presents our **Sofmatcher** framework for interactive image segmentation guided by visual prompting on a reference image. This consists of 3 steps: **1) Prompting & reference segmentation**, where a user provides scribbles on the reference image indicating the object class to be labeled on the target images, and where the scribbles are used as sparse prompt for SAM which then is used to output a reference mask; **2) Matching**, where *soft probabilistic matching* (detailed below) outputs a probability map over pixels of each target image quantifying their match to pixels in the reference mask; points are then sampled from the probability map, clustered and used for **3) Mask generation**, where clustered points are used as sparse prompts to SAM to generate mask proposals; these are filtered based on SAM's IoU predictions and aggregated into the mask output.

The key innovations of our framework compared to previous approaches like Matcher [Liu *et al.*, 2023] are aimed at producing an architecture that is amenable to being embedded in an interactive object segmentation workflow where users can provide visual prompts by coarsely annotating reference images through scribbles and interact in real-time with the resulting segmentation masks, possibly by correcting or complementing them with additional annotations.

Our first innovation for this is the **Prompting & reference segmentation** step in Fig. 1, which, while conceptually simple, provides a way for the user to directly and intuitively prompt the segmentation pipeline with *coarse visual prompts (scribbles)* instead of requiring detailed segmentation masks.

Our second major innovation is a computationally efficient version of the **Matching** step in Fig. 1, and was dictated by the requirement of low-latency segmentation and the observation that feature matching procedure used in the past, like the Hungarian algorithm (see e.g. [Liu *et al.*, 2023]), display a worst-case computational complexity that scales *cubically* with image sizes (number of patches) [Crouse, 2016], making them unpractical for an interactive workflow. Instead

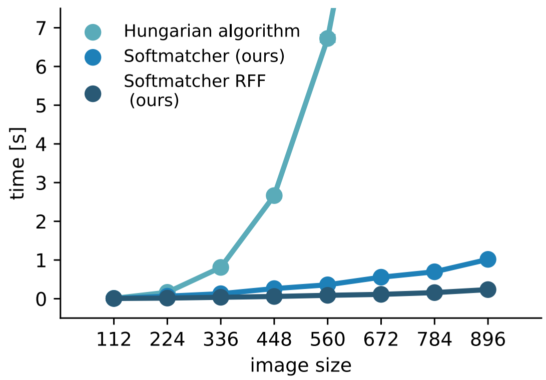

Figure 2: **Relative timing of different matching procedures** computed on 1 CPU core on a Dual AMD EPYC 7003/7002 Series Processors, assuming a featurization based on a VIT encoder with patch size of 14, feature size of 768.

of using (Hungarian) bipartite matching based on the cosine similarity between reference and target features, we opt for an Optimal Transport (OT) approach based on the quadratic cosine similarity matrix as a cost matrix. While very related, this method allows us to motivate a sequence of approximations for an efficient implementation of the matching procedure: we first introduce an entropic regularization, then consider the case of large regularization limit where the solution to the OT problem converges to the geometric mean of softmaxed cosine similarity maps between individual reference features and target feature maps (where the averaging is conducting across reference features) [Dognin *et al.*, 2019], an operation which only has *quadratic* complexity in the number of image patches complexity and results in our *Softmatcher* procedure. Moreover, it affords an even more scalable implementation by approximating the softmax computation of reference-target feature similarities through Random Fourier Features [Rahimi and Recht, 2007; Choromanski *et al.*, 2020], which we call *Softmatcher RFF*.

Figure 2 compares the timing of matching reference and target image features with the Hungarian algorithm, compared to our proposed soft matching methods as a function of image size assuming a featurization based on a VIT encoder with patch size of 14, feature size of 768. *Softmatcher* is around 6x faster than the Hungarian algorithm at image size 448, and this discrepancy quickly increases with image size due to its better computation complexity scaling. *Softmatcher RFF* is slightly faster and displays even better scalability.

We evaluate our visual prompting pipeline on FSS-1000 [Li *et al.*, 2020], which consists of 1000 object classes with pixel-wise annotations. FSS-1000 contains many objects that are not part of any previously annotated dataset (e.g., tiny daily objects, merchandise, and cartoon characters). As this disentangles previous knowledge from pretrained models to a certain degree, it lends itself well as a few-shot benchmark.

We integrate this improved matching pipeline into an interactive Visual Prompting platform that allows users to segment object classes of interest by merely highlighting representative objects in one or more reference images with scribbles. Given the improved computation complexity, our method allows the user to iterate in real-time with the segmentation outputs, adding additional scribbles on additional references to improve segmentation in case the model missed something, resulting in an intuitive and seamlessly interactive workflow.s

FSS-1000	Matcher	SM (ours)	SM RFF (ours)
one-shot	**87.0**	85.5 ± 0.7	85.9 ± 0.6
five-shot	**89.6**	87.1 ± 0.1	87.1 ± 0.3

Table 1: **Few-shot evaluation on FSS-1000:** We compare performance in terms of IOU of Matcher with our Softmatcher (SM) and Softmatcher RFF (SM RFF) methods on FSS-1000.

Deployed service and front-end. The interactive web interface is designed to provide seamless interaction between the user and the Softmatcher pipeline. It consists of a front-end built with Angular, a Python API back-end, and an inference service using Torch Serve. Users add scribbles to any image to mark objects of interest. The visual prompting pipeline then highlights similar objects with precise segmentation masks the target images. If the user is not satisfied with the initial results, they can refine the outputs by iteratively adding or deleting scribbles. Alternatively, instead of adding more scribbles, users can add additional prompts by converting output segmentation masks from a previous run into reference masks. These reference masks will skip step 1 of the pipeline (see Fig. 1). The system also allows for scribbles to be classified into different categories, enabling the creation of segmentation masks for multiple classes.

The process of repeatedly adding and adjusting scribbles provides users with a deeper understanding of how the model operates. By understanding the model's capabilities and limitations, users learn to collaborate with the model more effectively, leading to better outcomes. We've also started to enhance our framework's interactivity with vision-language models like CLIP, enabling the use of text prompts in addition to reference scribbles. This opens up the possibility of combining visual and text prompts to refine masks mutually and address scenarios where scribbling alone is not enough.

Demonstration. We illustrate how users typically engage with our web interface and the visual prompting pipeline through three sample projects. The first two projects illustrate a general use case on everyday objects, while the third shows a domain-specific proprietary defect detection dataset. Our demonstration covers the interactive process of adding scribbles to images, executing the pipeline to receive segmentation masks, and then enhancing the results by adding additional scribbles. Furthermore, we showcase the capability for users to process images with references from various classes.

Acknowledgments

We would like to thank Trafikverket for their collaboration, specifically for collecting image data of civil infrastructures. We are also grateful to Youssef Mroueh for insightful discussions. This work is partly funded by the European Union's Horizon Europe research and innovation program under grant agreements No. 101070408 (SustainML) and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00295.

References

[Bar *et al.*, 2022] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual Prompting via Image Inpainting, September 2022.

[Bommasani *et al.*, 2022] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. *arXiv:2104.14294 [cs]*, April 2021.

[Choromanski *et al.*, 2020] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking Attention with Performers. *arXiv:2009.14794 [cs, stat]*, September 2020.

[Crouse, 2016] David F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, August 2016.

[Dognin *et al.*, 2019] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Cicero Dos Santos, and Tom Sercu. Wasserstein Barycenter Model Ensembling, February 2019.

[Gupta and Kembhavi, 2022] Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional visual reasoning without training, November 2022.

[Ito *et al.*, 2022] T. Ito, T. Klinger, D.H. Schultz, J.D. Murray, M.W. Cole, and M. Rigotti. Compositional generalization through abstract representations in human and artificial neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2022.

[Janouskova *et al.*, 2023] K. Janouskova, M. Rigotti, I. Giurgiu, and C. Malossi. Model-Assisted Labeling via Explainability for Visual Inspection of Civil Infrastructures. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 244–257. Springer, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything, April 2023.

[Li *et al.*, 2020] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2866–2875, June 2020.

[Liu *et al.*, 2023] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment Anything with One Shot Using All-Purpose Feature Matching, May 2023.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

[Rahimi and Recht, 2007] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2007.

[Rigotti *et al.*, 2023] M. Rigotti, D. Antognini, R. Assaf, K. Bakirci, T. Frick, I. Giurgiu, K. Janoušková, F. Janicki, H. Jubran, C. Malossi, A. Meterez, and F. Scheidegger. Towards Workflows for the Use of AI Foundation Models in Visual Inspection Applications. *ce/papers*, 6(5):605–613, 2023.

[Wang *et al.*, 2023] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. SegGPT: Segmenting Everything In Context, April 2023.

[Zhang *et al.*, 2023] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize Segment Anything Model with One Shot, May 2023.