Research Statement

Utkarsh Mall

I build computer vision tools to enable automatic scientific discovery from large-scale data. My research builds foundation vision models for expert domains. My research also improves these foundation models to make them more suitable for scientific applications enabling discovery. In close interdisciplinary collaboration with domain experts, I also apply these methods to a diverse set of real-world scientific problems.

Advances in computer vision (CV) have led to the development of several automation tools. At the same time, several scientific domains aim to automate their experiments. CV and machine learning (ML) tools can and are transforming scientific domains. As these experiments get more complex and the experimentation datasets become larger, the need for CV and ML tools that can automatically gain scientific insight from the data becomes essential.

Contemporary CV and ML models cannot be used for automation and insight discovery due to several challenges. First, several of the foundation CV models are trained on images (data) on the internet. Such vision models fail in scientific domains as there is a considerable domain shift. The current workaround for this problem is to obtain labeled training data in a specific scientific domain. However, manual annotations are too expensive and sometimes impractical to obtain at scale in each domain. Second, even if we have perfectly working computer vision tools for specific domains, discovering scientific insights is still challenging for several reasons. Discovering useful insights from a large amount of data is like finding a needle in a haystack. Trying to make sense of such large-scale data manually is infeasible. Typically scientists work in tandem with machines to gain new insights from such data. The scientific process could be sped up and in some cases made feasible with models that can also automate such discovery.

My research aims to solve both these challenges to build an automated framework for scientific discovery allowing experiment automation and discovery of novel insights. I aim to build domain-specific foundation models in a label-efficient manner, allowing for improved automation of data processing in such domains. To address the second challenge, my goal is to develop tools that adapt foundation models to possess key qualities essential for scientific discovery, including better interpretability, correctness, uncertainty estimation, expert editability, and more.

My framework building such discovery tools and domain-specific foundation models has led to several key insights in different areas. For example, my methods automatically uncover interesting connections between occasion and clothing styles in fashion anthropology [17]. My research also enables the discovery of powerful hypotheses for indicators such as population density and biomass estimation in the areas of demography and climate science, respectively [18].

With close interdisciplinary collaborations, my research also aims to find universal challenges across these domains, intending to develop new tools that can be applied broadly across many disciplines. For example, label-efficient visual foundation models for remote sensing [16, 19] can be used to automate several fields like agriculture science, urban planning, disaster management, etc. Similarly, building neurosymbolic methods to build interpretable-by-design models for scientific discovery can be applied to numerous fields like demography and climate science [18].

Past work

Label-efficient visual foundation models for science: Building large-scale visual foundation models requires billions of images. While images from cameras and phones are easily available online, collecting similar data for scientific domains is more difficult. In remote sensing, while the images are easily accessible, labeling or even obtaining weak supervision like captions is challenging. Medical diagnostic images are even more challenging due to privacy reasons.

This poses several problems: First, the methods for training foundation models for internet images are sub-optimal for data in specific domains as they cannot leverage the unique properties of that domain. Second, collecting large amounts of data is infeasible in several domains, making the typical way of training foundation models unusable.

My work leverages properties unique to a domain to learn label-efficient representations for that domain. For example, for remote sensing data, I proposed a self-supervised representation learning method that leverages the spatio-temporal nature of satellite images [16]. The key idea is to learn a representation that can differentiate between long-term permanent changes and be invariant to short-term seasonal changes. The learned representation is applicable to numerous tasks like landcover classification [8], segmentation [26], and change detection [3] supporting societal applications like monitoring and tracking changes to our planet for sustainable development [2]. The method performs significantly better than self-supervised methods developed for internet images, by leveraging the spatio-temporal properties of the domain.

Scientists often train different models to solve distinct problems in their domain. While image-only self-supervised representations are useful, mapping semantic concepts that an expert wants to recognize in satellite images still requires labeled examples. In my research, I created a vision-language model (VLM) [23] for satellite imagery, enabling zero-shot recognition of open-world concepts. The key challenge was that, unlike internet data, collecting image-text pairs for satellite images is challenging. Therefore I leveraged internet images with geotags as an intermediary between textual concepts and satellite images [19]. This results in a VLM called GRAFT, which can be used for numerous tasks such as classification, retrieval, segmentation, and visual question answering, without additional task-specific training (zero-shot).

In other applications like ornithology, zero-shot recognition might not be entirely possible without attribute labels. For example, an attribute-based zero-shot model [27] requires an expert to describe all relevant attributes for a new bird category [15]. My research proposes a solution where the model actively queries for a specific subset of attributes, making the process more annotation-efficient [13]. This approach is inspired by field guides, which describe new species by highlighting key differences with familiar ones.

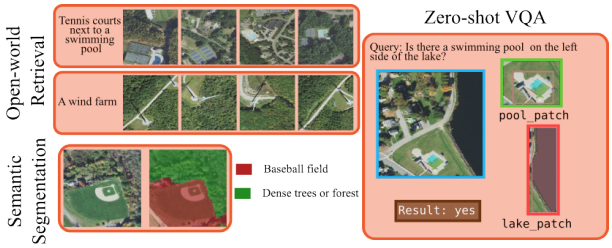

Figure 1: Applications enabled by GRAFT

Building such label-efficient domain-specific foundation models presents a novel challenge in AI research. At the same time, these foundation models have the potential to solve a wide range of problems in scientific domains.

Scientific program learning: A key step in science is building hypotheses based on collected observations. In fact, after collecting observations, scientists iterate over the hypotheses to build a theory. A good hypothesis that is 1) reliable: generalizable to unseen observation, 2) interpretable: provides insight into the problem while being accurate, and 3) sample-efficient: learnable with few good observations, forms the basis of the theory for a problem.

To learn such reliable, interpretable, and sample-efficient hypotheses, I built a framework for neuro-symbolic program learning called DiSciPLE [18]. The key insight was to learn hypotheses as Python programs built on powerful visual foundation models as primitives, resulting in reliability with interpretability. More specifically, DiSciPLE hypotheses are produced using LLMs and then tested on real-world data. Using an evolutionary search strategy, DiSciPLE builds new hypotheses by learning from the failures and successes of past hypotheses. On some of the real-world scientific problems, programs from DiSciPLE performed significantly better

than deep-learning models that tended to overfit while being uninterpretable [4]. The generated programs are interpretable by design and as a result, also provide scope for expert-in-the-loop interactions to impart domain-specific knowledge.

This work also highlights the interdisciplinary nature of my research. DiSciPLE can discover such interpretable and reliable hypothesis programs across a wide range of scientific domains ranging from demography [21] to climate science [22].

Virtuous cycle between Vision/AI and interdisciplinary fields: Close collaborations with domain experts allow bidirectional benefits. My research is guided by their needs, leading to novel problem exploration and the creation of tools that advance AI and CV. Conversely, my research has an impact on scientific domains. Advances in AI and CV tools in turn lead to better problem-solving in scientific domains.

For example, in remote sensing, clouds obstruct accurate observation, limiting insights into missing regions. My research introduced a large-scale benchmark and method for cloud removal, that fills spatio-temporal gaps in satellite images, enabling continuous signals for downstream applications [29]. Another problem when searching for concepts in a large region is the time-consuming process of iteratively searching over all high-resolution images. I developed a framework that first performs faster searches on low-resolution images, then selectively refines the search in high-resolution areas, significantly speeding up the search [24].

Conversely, my work has also led to interdisciplinary research, where domain experts apply the tools I developed to real-world problems in fashion anthropology [7] and public health [20].

Discovery from data: Domain-specific vision models can automate many steps of the scientific process but discovering insights requires tools beyond object recognition. Therefore, my research built spatio-temporal event discovery frameworks on top of these tools.

My research created a benchmark for the task of discovering "interesting" events [14] in remote sensing. For example, the goal is to discover spatio-temporal volumes corresponding to changes occurring due to particular real-world phenomena such as "forest fires" or "road constructions", useful for climate science, urban development, ecological impact studies, etc. Such methods effectively filter terabytes of spatio-temporal data into a manageable set of meaningful events.

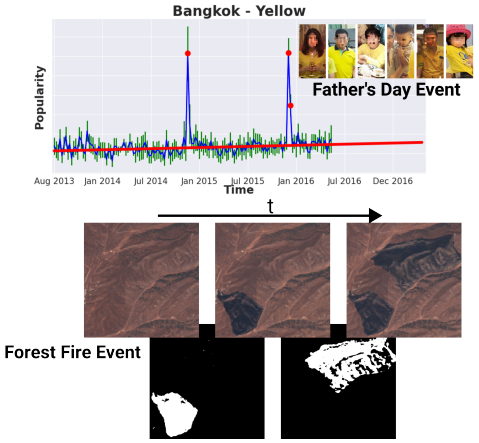

Figure 2: Automatic spatio-temporal event discovery from fashion (top) and satellite images (bottom).

My work also explored spatial and temporal discovery in another domain of fashion (anthropology). I proposed GeoStyle [17], a framework that models the fashion trends in a city over time, and discovers fashion anomalies occurring due to some real-world events when people wear different clothes. GeoStyle allows cultural anthropologists to explore these events globally, without physical exploration or prior knowledge (*i.e.* making discoveries). I also extended the discovery framework along the spatial dimension [12], resulting in discovering city neighborhoods with interesting fashion choices.

Future work

The North Star for my research is to enable discovery from massive amounts of scientific data being continuously captured around our planet, cities, and labs. To achieve such a goal, several

intermediate challenges need solving.

Better neuro-symbolic learning framework: DiSciPLE [18] showed that the proposed programs that are scientific hypotheses can be interpretable. However, DisciPLE is just the tip of the iceberg; there is huge potential for scientific discovery in our formulation of using LLMs for neuro-symbolic program learning. Domain experts also have other desiderata for these hypotheses. For example, in climate science and related fields, hypothesis models that are safe are preferred over extrapolative unsafe models [10, 28]. Similarly, most scientists prefer models that can provide predictions along with reliable uncertainty estimates. Improving on the neuro-symbolic framework proposed in DiSciPLE is the correct way to build models. For example, using general-purpose probabilistic programming languages [1] can be a way to automatically integrate the capability of sampling and consequently quantifying uncertainty. Programs also allow integration of other concepts from programming languages such as unit testing, and invariants [11] leading to a better understanding of the safety landscape.

Another direction I propose to explore is learning a library [5] of programs that can be reused for a wide array of problems in a domain. DiSciPLE can learn programs for one problem at a time. While this is useful, co-learning related programs for many problems can lead to better decisions. For example, modular subprograms that are used repeatedly across different problems can lead to finding unifying principles across problems. Such a framework can also lead to an interpretable understanding of the area as a whole rather than individual problems.

Building label-efficient scientific foundation models: While labeled data is hard to obtain for many scientific domains, unlabeled multimodal/multisensor data is becoming more available. For example, in climate science geo-spatial time-series data is publicly available [9], this can be combined with other sensors such as optical remote sensing, to obtain a multimodal foundation model that could help in both domains. Similarly, in many scientific domains such as social sciences [6], and public health, structured geospatial data can be found in the form of text, tables, CSVs, etc. All such data can also be combined with visual information such as street view images or images from the internet, to build foundation models in these areas. Like GRAFT, I plan to learn from such weakly paired information from various sensors. The tech industry is less incentivized to build such domain-specific foundation models, therefore I believe academia would be the right place to do this research.

Learning with scientists-in-the-loop: Neuro-symbolic programs also provide an intuitive interface for communication between machines and scientists. For instance, experts often struggle to impart their knowledge to deep learning models, typically relying on supervised data for multi-objective training [25]. However neuro-symbolic learning allows experts to provide information to the machine through code edits. I plan to improve this framework to enable scientists to easily incorporate their domain knowledge while the system learns.

Robust automated systems for scientific discovery: Besides collecting observations and building hypotheses, an often overlooked step in the scientific process is the cleaning and pre-processing of data. Scientists spend a lot of effort on this step and the conclusion of an experiment can vary significantly if this step is not done with care [30]. While these steps are important, different domains have their standards for them. For example, in remote sensing, there are standards for normalizing data from different sensors. I plan to build agentic systems that can obtain knowledge from prior works in related fields and use them for data cleaning. In interdisciplinary collaboration with scientists across areas, I aim to build an evaluation testbed for such systems, thereby establishing a virtuous cycle between AI and Scientists.

References

[1] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

[2] Sarah Carter and Martin Herold. Specifications of land cover datasets for SDG indicator monitoring. `https://ggim.un.org/` Accessed: 11-5-2024.

[3] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau. Urban change detection for multispectral earth observation using convolutional neural networks. In *IGARSS*, 2018.

[4] Mia Chiquier, **Utkarsh Mall**, and Carl Vondrick. Evolving interpretable visual classifiers with large language models. In *ECCV*, 2024.

[5] Kevin Ellis, Lucas Morales, Mathias Sablé-Meyer, Armando Solar-Lezama, and Josh Tenenbaum. Learning libraries of subroutines for neurally–guided bayesian program induction. *NeurIPS*, 2018.

[6] Social Explorer. Social explorer, 2013. `https://socialexplorer.com` Accessed: 11-5-2024.

[7] Rachel Rose Getman, Denise Nicole Green, Kavita Bala, **Utkarsh Mall**, Nehal Rawat, Sonia Appasamy, and Bharath Hariharan. Machine learning (ml) for tracking fashion trends: Documenting the frequency of the baseball cap on social media and the runway. *Clothing and Textiles Research Journal*, 2020.

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTAEORS*, 2019.

[9] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

[10] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 2024.

[11] K Rustan M Leino and Peter Müller. Modular verification of static class invariants. In *FM 2005: Formal Methods: International Symposium of Formal Methods Europe, Newcastle, UK, July 18-22, 2005. Proceedings*. Springer, 2005.

[12] **Utkarsh Mall**, Kavita Bala, Tamara Berg, and Kristen Grauman. Discovering underground maps from fashion. In *WACV*, 2022.

[13] **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Field-guide-inspired zero-shot learning. In *ICCV*, 2021.

[14] **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Change event dataset for discovery from spatio-temporal remote sensing imagery. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[15] **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Zero-shot learning using multimodal descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022.

[16] **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Change-aware contrastive learning for satellite images. In *CVPR*, 2023.

[17] **Utkarsh Mall**, Kevin Matzen, Bharath Hariharan, Noah Snavely, and Kavita Bala. Geostyle: Discovering fashion trends and events. In *ICCV*, 2019.

[18] **Utkarsh Mall**, Cheng Perng Phoo, Mia Chiquier, Bharath Hariharan, Kavita Bala, and Carl Vondrick. DiSciPLE: Learning interpretable programs for scientific discovery. In *arXiv preprint (in submission)*, 2024.

[19] **Utkarsh Mall**, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *ICLR*, 2024.

[20] **Utkarsh Mall**, Carl Vondrick, Marianthi Anna Kioumourtzoglou, and Robbie M Parks. How physical neighborhood features drive differences in health impacts of tropical cyclones. *ISEE Conference Abstracts*, 2024.

[21] Nando Metzger, John E Vargas-Muñoz, Rodrigo C Daudt, Benjamin Kellenberger, Thao Ton-That Whelan, Ferda Ofli, Muhammad Imran, Konrad Schindler, and Devis Tuia. Fine-grained population mapping from coarse census counts and open geodata. *Scientific Reports*, 12(1):20085, 2022.

[22] Juan Nathaniel, Gabrielle Nyirjesy, Campbell D Watson, Conrad M Albrecht, and Levente J Klein. Above ground carbon biomass estimate with physics-informed deep network. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1297–1300. IEEE, 2023.

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[24] Shreelekha Revankar, Cheng Perng Phoo, **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Scale-aware recognition in satellite images under resource constraint. *arXiv preprint (in submission)*, 2024.

[25] Jennifer J Sun, Ann Kennedy, Eric Zhan, David J Anderson, Yisong Yue, and Pietro Perona. Task programming: Learning data efficient behavior representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2876–2885, 2021.

[26] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *CVPR*, 2022.

[27] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018.

[28] Sungduk Yu, Walter M Hannah, Liran Peng, Mohamed Aziz Bhouri, Ritwik Gupta, Jerry Lin, Björn Lütjens, Justus C Will, Tom Beucler, Bryce E Harrop, et al. Climsim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. *arXiv preprint arXiv:2306.08754*, 2023.

[29] Hangyu Zhou, Chia-Hsiang Kao, Cheng Perng Phoo, **Utkarsh Mall**, Bharath Hariharan, and Kavita Bala. Allclear: A comprehensive dataset and benchmark for cloud removal in satellite imagery. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[30] Stephen T. Ziliak and Deirdre N. McCloskey. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. University of Michigan Press, Ann Arbor, MI, 2008.