Research Statement - Bharath Hariharan*

https://home.bharathh.info

July 20, 2023

1 Introduction

Behind the exceptional abilities of modern computer vision models are millions of hours of human labor in creating data (e.g., designing 3D models or artwork), curating data (e.g., filtering out poor or harmful content) and labeling data (e.g., carefully identifying the boundaries of every relevant object). This approach to building vision models using heavily annotated and curated data is expensive and often unethical (involving underpaid workers or data scraped without consent). The reliance on "big data" also leaves out rare or novel scenarios, damaging the robustness of these systems in deployment. These issues are exacerbated in scientific domains such as microscopy or "richer" tasks like 3D reconstruction where the equipment and expertise needed makes large annotated datasets prohibitively expensive.

I therefore seek to build learning-based computer vision systems that can learn from very limited data. To do so, I draw inspiration from human perception: after all, even two month-old infants can learn to recognize objects from just a few observations, and human adults can learn to recognize phenomena in completely new domains (X-rays, microscopy) with very little training. The key difference here is that unlike traditional learning systems, which treat every data point as independent and every learning task as a new task, for humans everything we see throughout our lives is connected because it is a view into the same underlying physical, dynamic world. Thus we interpret everything we see in the context of everything we have seen in the past, even when we encounter wholly alien domains like X-rays. My research incorporates these ideas into computer vision systems by connecting together disparate perception tasks, and disparate images within a task. This has yielded computer vision systems that can discover object classes without a single annotated example, model their shape and how they deform, and even bootstrap themselves in entirely new domains with very few labels. The impact of my work is visible in the 33k citations it has received, and has been recognized through an NSF CAREER and a PAMI Young Researcher award.

2 Bootstrapping recognition systems in new domains

How do we build recognition systems in new domains? Traditional machine learning techniques treat every new domain as a new learning problem, starting with a blank slate. As such, they need hundreds, thousands, or millions of labeled examples. This is in stark contrast with how humans learn new tasks. For instance, a biologist can quickly learn to interpret microscopy images by drawing on their experience on the macro scale: they may identify microscopic features as "comets" (knowing full well that these are not actual comets at all).

In two papers [9, 1] we instantiated this intuition for machine vision systems. When faced with a new domain, we first searched for similar domains for which off-the-shelf pre-trained classifiers are available. This nearest domain may not be similar enough: for example, if our target domain is images of sick plant leaves, we may only find a pre-trained classifier on generic objects. Nevertheless, we can use this classifier to produce fake annotations, called pseudo-labels, on the target domain. While the labels themselves might be incorrect, we observed that semantically similar images would still be assigned the same label. For instance, the object classifier would mistake sick peach leaves as "bananas", but *all* sick peach leaves would be labeled as bananas

[*] See https://home.bharathh.info for the most up-to-date research statement.

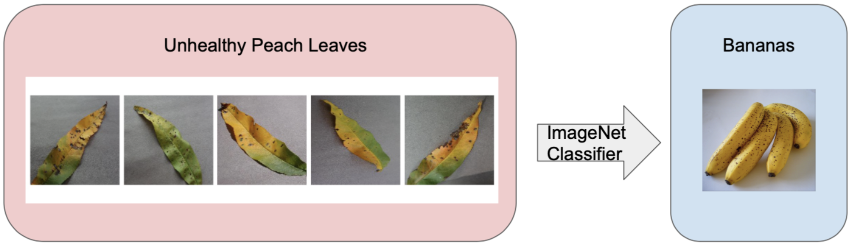

Figure 1: A classifier from one domain (ImageNet) when applied to images from another domain (images of sick plants) will assign incorrect labels but still group images in a semantically meaningful manner.

(Figure 1). We used these pseudo-labels as an auxilliary training signal in addition to the limited labels we have in the target domain. These pseudo-labels forced the classifier to recognize semantically meaningful features (such as roads in the example above) resulting in a more robust classifier when labels were few. The robustness of this approach has been corroborated by the independent discovery of this approach by other groups, as well as by a blind DARPA evaluation (in the Learning with Less Labels program) where this approach yielded the 2nd best results overall.

Our past visual experience does not just give us a vocabulary of visual concepts. It also informs several expectations we have about the visual world. For instance, we know that objects can appear in many different backgrounds and from many different viewpoints. So when learning about a new kind of object (say a new bird species) we discount the background or object pose. In contrast, current recognition systems must learn to do this the hard way, by observing the new object in many different backgrounds and with many different poses. In multiple papers [13, 10, 14, 8, 5], we have shown repeatedly that explicitly encoding figure-ground separation and pose invariance in image classifiers leads to much higher accuracies with much fewer annotations. In fact, we found that shallow networks with this encoded invariance outperform much deeper black box networks with 5 times as many layers. Our most recent paper along these lines builds a classifier architecture that explicitly allows for object features to shuffle spatially in the image and is the current state-of-the-art for learning image classifiers from very few examples [14].

3 Discovering visual concepts from data

Even with the advances above, modern vision systems still need to be told what objects to recognize. But when these trained vision systems then encounter new kinds of objects in deployment, they might miss them entirely. For example, a self-driving car trained to recognize cars, cyclists and pedestrians may fail to detect a snowplow no matter how big it is. This is not an error that humans make: even if we have never seen a snowplow before, and even if we don't know the name, we would still recognize it as an object of interest.

What underlies this big distinction between humans and machines? We observed that where machine vision systems treats each image as an independent data point, the human experience is that of a continuous stream of visual input that reveals facts about the underlying physical world. For instance, we may have never seen the assortment of pixels that is the snowplow, but we watch a massive 3D shape *moving* relative to the static scene and know that we must avoid a collision. Then in subsequent encounters, we may see a stationary snowplow but recognize it immediately for what it is.

In a recent collaboration [17], we asked whether machine systems can use a similar mechanism to automatically train an object detector *without a single annotation*. In particular we looked at self-driving car systems operating on LiDAR input. Instead of treating unlabeled LiDAR data as a collection of independent data points, we observe that cars often traverse the same scene multiple times (because people drive the same routes everyday). We use these multiple traversals to automatically identify LiDAR points that are not persistent across all traversals of a scene as "moving" points. Clustering these points yields a noisy set of "moving" objects, which can then be used as supervision for an object detector. We thus train a 3D "moving object" detector that *outperforms* a supervised detector, but is trained with no labels at all.

This idea of discovering visual concepts (such as objects) without labels is also useful in scientific domains,

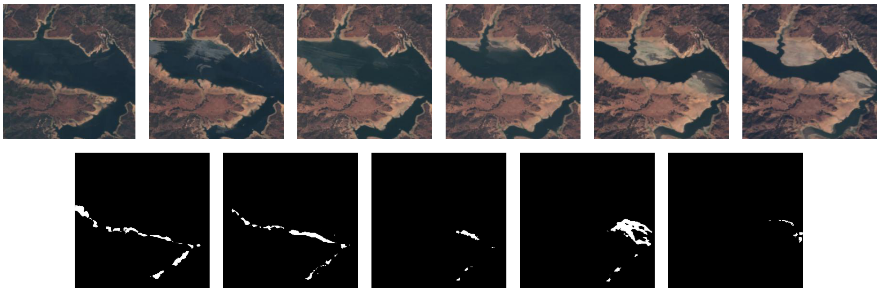

Figure 2: An event discovered by our approach. This event shows a lake drying up. The top row is the sequence of images, and the bottom row shows the pixels that were identified to have changed between each pair of images.

where we may have visual data but be unsure of what to look for. In another collaboration, we explored such discovery for satellite images. In satellite imagery, concepts of interest are *change events*, because these relate to human activity, natural disasters, climate change and so on. Therefore instead of treating satellite images as simply an unorganized collection of images, we use the spatiotemporal structure to focus on change. We design a first-of-its-kind discovery pipeline that (a) learns a feature representation sensitive to change, (b) uses this representation to detect spatiotemporally distributed change events, and (c) groups events into meaningful clusters that can be construed as classes [7, 6]. We find that our approach automatically discovers classes of events like “road construction”, or “drying of lakes” (names assigned post-hoc) without any annotations of the same (Figure 2).

These are but the first steps towards discovery without labels. I envisage future recognition systems that automatically go through vast troves of unlabeled data to uncover potential visual concepts of interest, and then interact with a human expert to validate these concepts and name them. In addition to reducing the amount of labels required, this approach will additionally help automate one of the most significant and yet underappreciated challenges of building recognition systems in novel domains: defining the taxonomy of classes in the first place.

4 Seeing the physical world behind the pixels

We trained a ”moving object” detector from unlabeled data without any supervision. But can we discover finer-grained categories, like cars, pedestrians, cyclists or snowplows? This question depends on what the underlying basis of categorization is, and what we will ultimately use this categorization for. For humans, often the simplest basis of categorization is object shape. For a robot or other embodied agent, categories primarily serve as a convenient shorthand for how different objects behave and how the robot should interact with them. Thus, if a machine vision system can not just detect useful objects but also estimate their shape and how they will behave (move, articulate or deform when acted upon), then it has enough information to discover meaningful categories. The automatic discovery of object categories thus requires the *reconstruction* of the 3D world and its motion over time.

The problem of reconstruction is a mainstay of computer vision and has seen decades of research. While off-the-shelf solutions (e.g., COLMAP) exist, these solutions make myriad assumptions: e.g., a primarily static scene, and the availability of multiple views. In the wild, where these assumptions are not met, these solutions produce noisy or incomplete reconstructions. In my research, I have been working towards more precise and complete reconstruction of shape and motion of dynamic scenes from a few sparse views.

Reconstructing shape Classical techniques for reconstructing 3D shape yield noisy or incomplete results when there are few views and when the scene is challenging (e.g., with textureless surfaces). With collaborators, I have worked on learning-based techniques that improve the classical pipeline: we trained better descriptors for correspondence from unlabeled data [12] and are working on learning to estimate relative camera pose for widely different or ambiguous views [2, 3].

In addition, I have also built generative models of 3D shape [16, 4] (Figure 3) to help denoise noisy or incomplete shape reconstructions in the common form of sparse point clouds. This is challenging because while generative models over vectors or tensors are common, there is no way to handle arbitrary sets such as point clouds. Our insight is to treat point clouds as *samples* from an underlying probability density that is high on the object surface and low everywhere else. Thus each 3D shape corresponds to a *probability distribution over 3D points*, and a point cloud is a population of samples from this distribution. We represent this probability distribution using a neural network that takes two inputs, a 3D point and a vector "code" that specifies the shape identity, and outputs a probability value of sampling that point. The "code" reduces each shape to a single vector, allowing us to use standard generative modeling tools to model priors over shapes. This line of work was one of the first generative models that could be trained on arbitrary point clouds, and allows us to complete and de-noise point clouds obtained from traditional reconstruction pipelines.

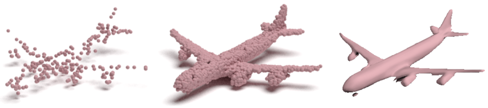

Figure 3: Our generative model [4] can complete sparse point clouds.

The idea of using a neural network to represent a 3D shape is now a burgeoning area of research; today such neural networks are called *neural fields*. While our work was not the first instantiation of this idea, it was the first instantiation that could work with point clouds, the most common way in which 3D shape is processed by 3D sensors like LiDAR or 3D reconstruction pipelines. As such, our work has since been heavily cited (> 400 citations), both in its capacity as a generative model of shape and as one of the first instantiations of a neural field.

Reconstructing motion and deformation Much of our world is dynamic, filled with moving, articulated and deformable objects. Modeling how different objects will articulate or deform is essential for many robotics applications. In my recent work I am therefore looking at modeling and reconstructing motion or deformations from unlabeled input.

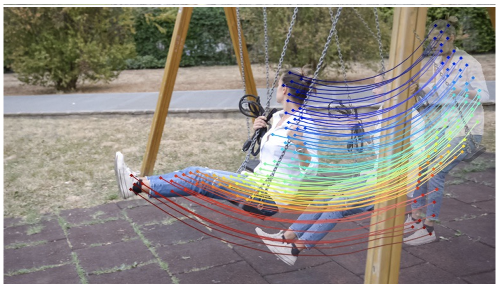

Figure 4: Our approach [11] can track points through occlusions.

The key challenge with estimating motion or deformation is that one needs to track every 3D point in the scene over time. This is fundamentally a correspondence problem: we need to match points in one time point (or frame) with other time points. Unfortunately, this match may be ambiguous as many scene points may look the same. The ambiguity is further exacerbated by phenomena like occlusion, which are common in real world scenes. In recent work [11] we regularized this problem by representing a video as a canonical scene smoothly deforming over time, with both the scene and the deformation represented using a neural field. This approach allows us to consistently and correctly track 3D scene points even through large deformations and extended periods of occlusion (Figure 4). In a related work [15], we explored more expressive, physics-based regularizers, allowing for as-rigid-as-possible deformations of neural fields. This work was the first to explore such physics-based deformations for neural fields, and has since spawned several follow-up papers on using neural fields for physical simulation of deforming and articulated objects.

While we are far from robust reconstruction of dynamic scenes, this work already hints towards the possibility of identifying moving objects and understanding their behavior from raw unlabeled video without supervision. In the future, I plan to use this reconstruction to discover object categories from video and build a full machine vision system (reconstruction + recognition) without any human annotation at all.

5 Future work

Going forward, I am interested in further expanding the capabilities of vision systems while reducing their need for data. Some of the directions I am interested in particular are:

Expert-in-the-loop recognition: In modern recognition pipelines, humans are primarily data labelers. However, in challenging domains where labeling requires expertise, the experts may have important domain knowledge. My work has shown that explicitly incorporating this domain knowledge leads to more accurate, more data-efficient systems. What is missing, however, is a unified framework for experts to specify many different kinds of domain knowledge to the learner, and a system that can understand and learn quickly from these specifications so as to minimize expert effort. Advances in language models show a way forward, provided we can correct for erroneous output,

Discovery through reconstruction: In my work, we are expanding the capabilities of 3D reconstruction pipelines to handle unconstrained, dynamic scenes. As discussed above, I believe that this reconstruction can be used to discover categories of interest and develop recognition engines without labels, by clustering together objects with similar shapes and behaviors. The discovered categories can then allow us to further produce priors over object shape and behavior, feeding back into the reconstruction pipeline. Thus, I envisage a virtuous loop between recognition and reconstruction that reduces, if not eliminates, the need for human annotations.

Perception for robotic manipulation: Perception is not an end task; rather it is a means to an end. To build better perception systems we need to engage with the downstream application. I am interested in taking the techniques we have built and are building, to design perception modules for robots. I am especially interested in robots performing fine-grained manipulation of deformable or articulated objects, where fine-grained categorization and reconstruction becomes important.

References

[1] Kenneth Borup, Cheng Perng Phoo, and Bharath Hariharan. Distilling from similar tasks for transfer learning on a budget. In *ICCV*, 2023.

[2] Ruojin Cai, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Extreme rotation estimation using dense correlation volumes. In *CVPR*, 2021.

[3] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *ICCV*, 2023.

[4] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. In *ECCV(**Spotlight**)*, 2020.

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

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

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

[8] Cheng Perng Phoo and Bharath Hariharan. Coarsely-labeled data for better few-shot transfer. In *ICCV*, 2021.

[9] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *ICLR(**Oral**)*, 2021.

[10] Luming Tang, Davis Wertheimer, and Bharath Hariharan. Revisiting pose-normalization for fine-grained few-shot recognition. In *CVPR*, 2020.

[11] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023.

[12] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *ECCV(**Oral**)*, 2020.

[13] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *CVPR(**Oral**)*, 2019.

[14] Davis Wertheimer, Luming Tang, and Bharath Hariharan. Few-shot classification with feature map reconstruction networks. In *CVPR*, 2021.

[15] Guandao Yang, Serge Belongie, Bharath Hariharan, and Vladlen Koltun. Geometry processing using neural fields. In *NeurIPS*, 2021.

[16] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *ICCV(**Oral**)*, 2019.

[17] Yurong You, Katie Luo, Cheng Perng Phoo, Wei-Lun Chao, Wen Sun, Bharath Hariharan, Mark Campbell, and Kilian Weinberger. Learning to detect mobile objects from lidar scans without labels. In *CVPR*, 2022.