Research Statement

Cheng Perng Phoo (cpphoo@cs.cornell.edu)

The ability to perceive and understand the world is a major milestone for artificial intelligence. With strong perception capabilities, a model/artificial agent could analyze large collections of data and make informed decisions required to achieve its designated tasks. Over the past decades, we have observed tremendous progress in visual perception for Internet applications [5, 10, 14, 16, 17]. While these successes are laudable, successes in building perception models remain limited for other problem domains such as remote sensing, scientific discovery, and robotics. My research focuses on **building perception systems that are broadly useful for all problem domains**.

Toward achieving this goal, I have identified three major problems: label efficiency, deployment to novel domains, and trustworthiness. In this statement, I will describe a few of my past work on tackling the first two problems. Then, I would conclude this statement with a brief discussion of my future directions on tackling the three major problems, enabling perception for any problem domains.

1 Label-efficient Perception

The biggest problem in building perception systems for many problem domains, such as satellite imagery or egocentric imagery collected by embodied agents, is *the lack of large-scale annotated datasets.* Current perception systems are *not label-efficient*, often requiring a large amount of annotated examples to build; while this assumption holds for everyday internet images, it rarely holds for other problem domains. For example, building a perception system to identify new types of viruses would require a microbiologist to painstakingly annotate thousands or millions of microscopic images — an expensive affair.

In this section, I will present two lines of work, each taking a different approach to improving label efficiency. The first line of work uses self-training to bootstrap new models/representations from pre-trained models; the second line uses domain knowledge in satellite imagery to create perception models without human annotations.

1.1 Self-training to Bootstrap Perception Models for Any Problem Domains

Consider tasking a home robot to identify/remove sick plants in a greenhouse. While this agent might possess a perception model that could recognize common household items (e.g., furniture, utensils, food), it will have to learn to recognize new concepts (various plant diseases) rapidly with perhaps a limited amount of annotated examples to achieve its goal (say less than ten annotated examples). How could we enable the agent to learn new concepts in a new environment different from where it is supposed to operate?

My past work STARTUP [15] answers this question via a simple solution: adapting pre-trained models using unlabeled data from the target task. After all, it is often the case that labels are difficult to obtain, but unlabeled data are freely available (e.g., the robot could have access to past footage of the greenhouse).

To adapt the pre-trained model with unlabeled data, STARTUP leverages one simple observation: pseudo-labels from a pre-trained model could sometimes group unlabeled data into meaningful clusters (see Figure 1). By "self-training" a new classifier to replicate the pseudo-labels on the target unlabeled data, we can develop visual representations specialized to the new domain that are much stronger than state-of-the-art self-supervised visual representations [4]. This representation allows us to quickly learn a classifier with 5 to 10 examples per class. In addition, if we assume we have access to multiple pre-trained perception models (of different architectures/pre-training datasets), my recent work DistillNearest/DistillWeighted [3] shows that by using similarities between the pseudo-labels and the ground truth in a small labeled training set, we could distill the knowledge from different pre-trained models into a single efficient model for better perception.

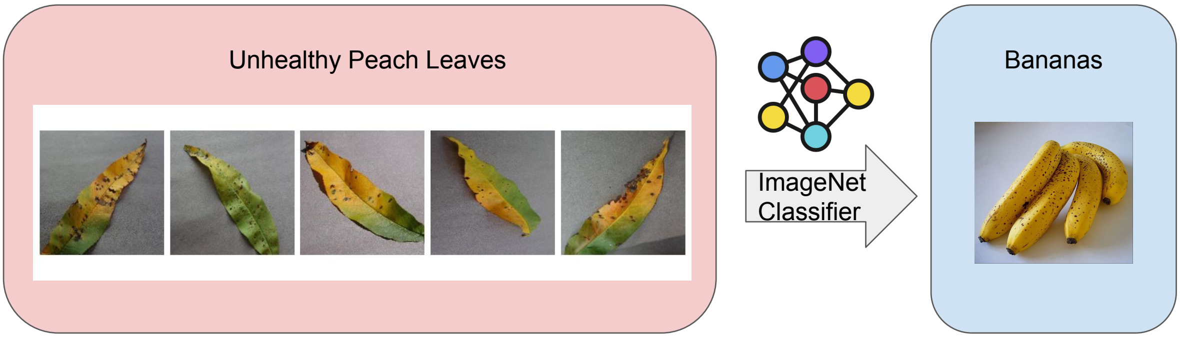

Figure 1: Predictions by an ImageNet classifier on five unhealthy peach leaves. While the classifier is not trained to classify unhealthy peach leaves, it recognizes all of them as bananas. The predictions are not semantically correct, but they correctly group unhealthy peach leaves into a single cluster. This indicates that the pseudo-labels by the classifier contain valuable signals that could be used for bootstrapping new perception models.

STARTUP and DistillNearest/DistillWeighted are general approaches that could be applied to any problem domain with sparse annotations and unlabeled data. While effective, they do not leverage characteristics of the problem domain that could potentially further enhance label efficiency. Next, I will discuss my work on label-efficient perception for satellite imagery via domain knowledge.

1.2 GRAFT: Aligning Ground Images and Satellite Images for Training VLMs without Annotations

Visual-language models (VLMs) such as CLIP [16] allow better open-vocabulary perception and better accessibility of perception models to non-AI-experts. Developing a VLM for satellite imagery would enable automatic analysis of large-scale satellite imagery for non-AI-experts. However, while we often upload images to the internet with textual descriptions, remote-sensed satellite imagery usually does not come with textual annotations since they are (semi-)automatically generated with less human involvement.

To build a VLM without textual annotations, I present my recent work GRAFT [13]. GRAFT sidesteps the need for textual annotations using two ingredients: (1) CLIP — a vision and language model that connects internet imagery to natural text, and (2) the observation that there might be multiple ground images associated with a single satellite image. More precisely, remotely sensed satellite imagery captures a single location on Earth. At the said location, multiple images could be captured on the ground and uploaded to the internet (see figure 2). By building a satellite image encoder that aligns with CLIP's image representation, GRAFT effectively uses ground images as an intermediary to establish the connection between satellite images and natural text, yielding a VLM for satellite imagery without needing individual textual annotations! Though simple, GRAFT can outperform many prior VLMs built with expensive textual annotations.

2 Deploying Perception Systems to Novel Domains

While label efficiency is crucial for building useful perception systems for different problem domains, safely deploying perception systems in a novel environment is also critical for adopting perception systems in any problem domain. Current perception systems are built using machine learning, which assumes identical training and testing environments. This assumption often does not hold in reality. Circling back to previous examples, if the microbiologist were to lend their model to their friend from another lab, they might find that their perception model fails to work properly because their friend uses a different microscope. This problem calls for mechanisms to adapt perception models to novel testing domains rapidly, ideally without any annotations (a.k.a the unsupervised domain adaption /

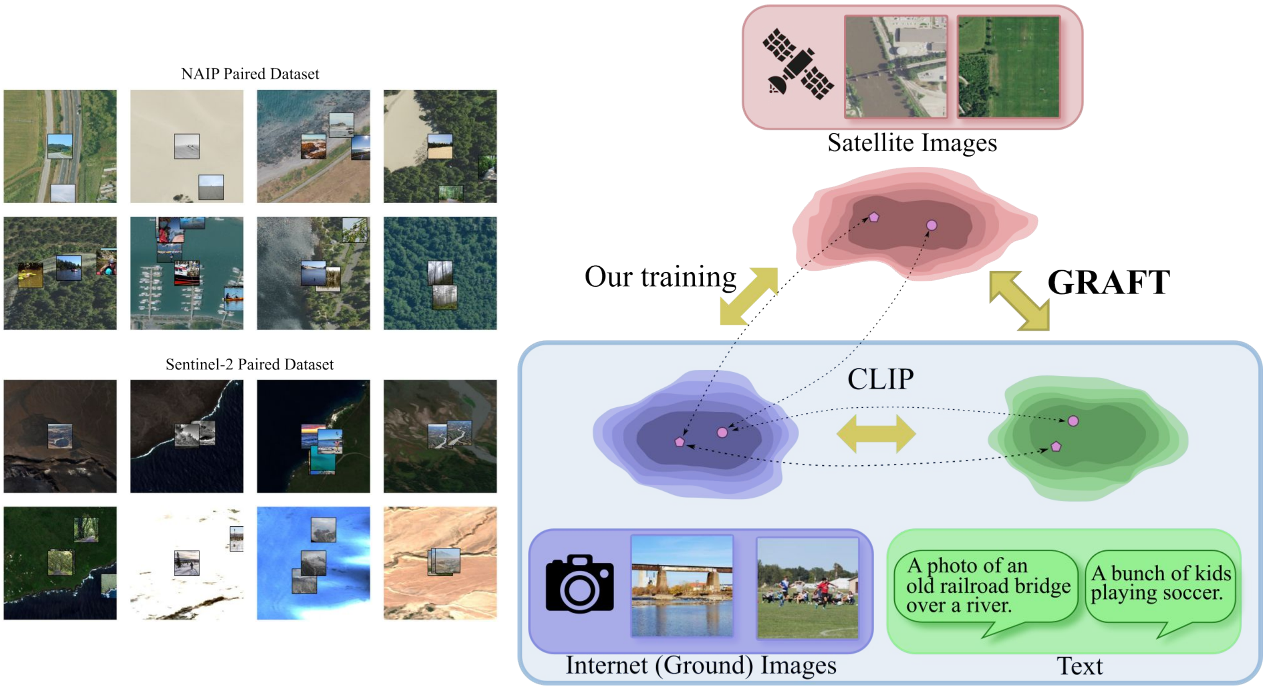

Figure 2: **Left:** Multiple ground images are associated with a single remotely sensed image. This applies to satellite images captured at different ground sampling distances (top vs bottom). **Right:** GRAFT VLM. We construct a representation of satellite images that aligns with CLIP's representation of their associated ground images. Coupled with CLIP's text encoder, we obtain a VLM without needing any annotations.

UDA problem).

Numerous works have attempted UDA leveraging unlabeled data [9, 12, 20, 21]. While showing significant progress, these approaches are still limiting since they often do not use any domain knowledge to aid the adaption process. In the coming section, I will describe my past work on adapting the perception systems of self-driving cars leveraging unlabeled data and domain knowledge.

2.1 Rote-DA: Adapting Perception Systems for Autonomous Vehicles with Repeated Traversals.

Machine learning often assumes that data are independent and identically distributed (IID). However, real-world data are often not IID but rather correlated. Take autonomous vehicles as an example. Modern vehicles are often equipped with precise localization(GPS/INS). Thus, data collected by vehicles can be indexed and connected via the geo-locations they are collected.

Assuming data in the autonomous driving domain are IID could limit the possibility of developing stronger perception systems. In fact, my past work, Rote-DA [26], has shown that by leveraging this domain knowledge, we could better adapt a 3D object detector to a new domain using geo-indexed, non-IID, unlabeled data. Specifically, by comparing unlabeled LiDAR scans captured at the same location at different times (a.k.a. repeated traversals of the same location), we can effectively segment out dynamic LiDAR points. This segmentation signal allows us to effectively remove false positives when applying the source-detector to the target domain unlabeled scans, yielding cleaner pseudo-labels for better self-training adaptation.

3 Future Directions

Now that I have discussed my past work, I will outline my future directions.

3.1 Label-efficiency and UDA through Multiple Input Modalities

Humans observe the world through multiple senses. The complementary nature of different senses allows us to perceive the world holistically. This observation does not only human intelligence but also artificial intelligence. In particular, multiple works, including but not limited to camera-LiDAR fusion for self-driving cars [2, 7, 23, 24] and visual-audio fusion for fine-grained bird classification [22], have corroborated that multiple sensing modalities can enhance perception. Despite progress on multimodal perception, exploration of using multi-modal input to enhance label efficiency or domain adaption remains scarce. The success of my past work, GRAFT, has indicated the possibility of leveraging complementary sensor information (satellite and ground images) to enable building VLM without annotations. For future work, I will explore the limit of label efficiency brought by multimodal inputs. In particular, I would explore how we could construct label-efficient multimodal models from adapting unimodal frontier models of different modalities including but not limited to vision [10] and audio [1]. Solving this would enable more useful/performant perception models for various applications, such as audiovisual categorization of fine-grained species or perception for embodied agents.

3.2 Trustworthy Perception from Pre-trained Models

My prior work mostly focused on label efficiency or domain adaptation of pre-trained models. One aspect of useful perception systems that I did not explore is trustworthiness — how can we reliably trust the output produced by our perception model? While different aspects of trustworthy perception models have been explored previously in the literature (explainability[19], calibration [6, 11], OOD detection [8, 18]), they are often explored under the fully supervised setup, whereas investigation on label-efficient models remains scarce. However, with the democratization of various large-scale frontier models [25], we now have access to label-efficient learners that are pre-trained on a gigantic amount of data. Their exposure to a large amount of data (potentially unrelated to the target problem domain) could give rise to more opportunities to reassess and improve these different aspects of trustworthy models. For instance, with the advent of text-to-image generation [17] and LLMs [14], one could leverage both technologies to generate more near-distribution examples for training OOD detectors. In addition, these frontier models are strong label-efficient learners (some of them [16] are even zero-shot learners), which catalyze wide-spread adoptions, but it is apriori unclear how calibrated they are after fine-tuning or how to calibrate them effectively in a label-efficient manner. For future research, I am interested in tackling how we could build trustworthy perception models from these frontier models.

References

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hangbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2022.

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

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2020.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[7] Xiaofeng Han, Huan Wang, Jianfeng Lu, and Chunxia Zhao. Road detection based on the fusion of lidar and image data. *International Journal of Advanced Robotic Systems*, 14(6):1729881417738102, 2017.

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[9] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018.

[10] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[11] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.

[12] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[13] 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, 2023.

[14] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

[15] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. In *International Conference on Learning Representations*, 2020.

[16] 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.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[18] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. *arXiv preprint arXiv:2105.14148*, 2021.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[20] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[21] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[22] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. In *European Conference on Computer Vision*, pages 271–289. Springer, 2022.

[23] Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4603–4611, 2020.

[24] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11789–11798, 2021.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[26] Yurong You, Cheng Perng Phoo, Katie Luo, Travis Zhang, Wei-Lun Chao, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Unsupervised adaptation from repeated traversals for autonomous driving. *Advances in Neural Information Processing Systems*, 35:27716–27729, 2022.