Subhransu Maji

Research Statement

I am a computer scientist specializing in computer vision (CV) and machine learning (ML). My research aims to make fundamental contributions towards developing AI systems with rich visual reasoning capabilities. My research is driven by the vast applications of computer vision and the potential for visual data analysis to provide insights into social, cultural, and natural factors that impact our lives. For instance, data collected from personal devices, citizen science platforms, social media, autonomous platforms, and medical devices can support innovative applications in e-commerce, robotics, manufacturing, and healthcare. Similarly, data from satellites, weather radars, and telescopes, and other sensor networks offer unprecedented opportunities to understand planet-scale phenomena, such as the effects of climate change, and to shed light on processes that govern the physical world.

While recent advances in AI have the potential to dramatically change how we interact with the world, make decisions, and design solutions for these applications, significant challenges remain. Current AI systems lack the ability to perceive details, handle multiple modalities, interact with humans to solve complex tasks, tackle novel tasks with limited supervision, and do not have the precision that scientists desire.

In addition to my core AI research which addresses these computational challenges, I engage in interdisciplinary collaborations to solve problems of societal interest and to advance science. I collaborate with ecologists to analyze bird migration [8, 28, 43], with astronomers to uncover scientific insights from images of galaxies [23, 29, 41], with chemists to develop representations of 3D materials [30], among others. Moreover, I am deeply involved in community building. I have been a long-term organizer of the Fine-Grained Visual Categorization workshops and recently organized the CV4Science workshop, which aim to foster collaborations between researchers in CS, humanities, natural sciences, and industry. I also mentor projects in collaboration with non-profits though the Data Science for Common Good initiative at the University.

Application areas My research is focused on the following areas:

- *Fine-grained recognition.* My earlier research developed state-of-the-art models for fine-grained categorization tasks such species indentification (e.g., [26, 27]), texture and material recognition (e.g., [12, 13]), attribute recognition of people and general objects (e.g., [3, 4, 35]). My recent research has focused on utilizing foundation models such as CLIP, GPT4, and self-supervised learning to enable part and attribute recognition with unlabeled and coarsely labeled data [11, 44–46, 55, 61, 62], and zero/few-shot learning [5, 54, 56]. With collaborators, we have developed techniques for utilizing multi-modal data such as geographic location, textual description, images and audio from iNaturalist and Wikipedia to enable better species indentification and range estimation.
- *3D shape understanding.* My earlier research contributed to architectures for 3D shape classification and segmentation tasks (e.g., [52, 53]), estimating 3D shapes from images (e.g., [18, 20, 32]), and methods for easier editing and manipulation of 3D shapes. My recent research extends these techniques with modern ingredients such as NeRFs and self-supervised learning for learning dense 3D representations [6, 7, 17, 19, 22, 47–50]. These methods can enable applications in graphics and robotics where 3D understanding is necessary.
- *Applications.* A significant component of my recent research involves collaborations with domain experts. I lead the computer vision efforts for the "dark ecology" project, where we have developed techniques [8, 28, 43] to study bird migration using radar imagery. This has enabled us to extract patterns of bird migration across the entire United States and over three decades by analyzing archived weather radar data. The work has resulted in novel scientific findings [2, 16] and high-impact publications [24], as well as coverage in popular media. I also collaborate with astronomers to analyze high-resolution images of galaxies to uncover the dynamics of star formation [23, 29, 41], with chemists to develop ways to analyze catalysts for chemical separation [30], and with civil engineers to monitor terrestrial water systems [31]. Common research themes across these collaborations include techniques to adapt CV models across heterogeneous domains, learning with limited amounts of supervision, and enabling scientists to draw conclusions from large datasets using imperfect AI systems.

Research themes In the context of these applications my research follows two themes:

- *AI with humans in-the-loop.* My earlier research focused on designing annotation tasks to discover interpretable parts and attributes [33,38], efficiently annotate data [36], and incorporate human feedback for deploying imperfect AI systems [60]. My recent research revisits these techniques to use modern AI systems for applications. For example, we have developed a technique called DISCOUNT [40], which enables accurate and efficient counting in large image collections by using humans to vet a fraction of the detector outputs. DISCOUNT was deployed for estimating damaged buildings for disaster planning for the Red Cross and for estimating bird migration trends from radar imagery. This work won the Best Paper Award at AAAI'24 (AI for Social Impact Track).
- *Efficiency in learning.* My earlier work contributed to improving the efficiency of recognition systems (e.g., [34, 37]), analysis of image representations for 3D shape understanding, and algorithms for few-shot learning. My recent research focusses on techniues for learning from coarse labeled datasets, developing self-supervised learning techniques for 3D data, understanding properties of deep networks theoretically [10, 22, 51], and modeling similarity between tasks for meta-learning [1, 15, 25, 57–59]).

Research collaboration, impact, and funding Within CS, I am a member of the CV, ML, and broadly AI communities. My research strategy is to publish fundamental and application-driven results in AI venues and collaborate with domain experts to publish novel scientific findings in ecology, astronomy, chemistry, and remote sensing journals. My PhD students find these collaborations fruitful, and several of them have won the Outstanding Synthesis Award, given to a few students each year in the Computer Science (CS) department. Besides this, I also engage in community building to identify interesting problems and develop benchmarks. I have organized the last nine Fine-Grained Visual Categorization workshops (FGVC3 – FGVC11), as well as the first CV4Science workshop, and contributed several datasets to the community.

Since joining UMass, I have published 70 articles in highly selective CV, ML, and AI conferences, including 35 at CVPR, ICCV, ECCV, KDD, SIGGRAPH, and AAAI, 18 articles at high-impact journals, and two book chapters. These include 23 conference papers and 11 journal articles that were published since my Tenure application in June 2019. My publications have been cited 26,419 times (h-index 47; i10-index 86) according to Google Scholar as of June 2024. Papers that I have co-authored have received the Best Paper Award at AAAI (AI for Social Impact Track) 2024 [40], Best Paper Honorable Mention at CVPR 2018 [52], and Best Student Paper at WACV 2015 [60].

My work has been supported by eight awards from the National Science Foundation (NSF), with three as sole PI and five as Co-PI, as well as grants from the National Aeronautics and Space Administration (NASA), Climate Change AI Foundation, and gifts from Dolby, Facebook, Adobe, and NVIDIA. Three NSF grants (one as sole PI and two as Co-PI) and the NASA grant (as Co-PI) were awarded since I applied for Tenure. The UMass portions of these awards total 5.7 million US dollars, of which 2.4 million are since my Tenure application.

On the education side, I have supervised twelve PhD students, seven of whom have graduated and taken up research positions in industry, or have remained in academia as post-docs. One of my students, Zezhou Cheng, will start as an Assistant Professor at the University of Virginia in Fall 2024.

Below, I describe my research contributions and future work organized into two thrusts: *advancing visual recognition* and *applications*, with a focus on work since my Tenure application.

1 Advancing Visual Recognition

The last few years have seen dramatic progress toward general intelligence with models such as GPT4, DALLE, and Gemini, that can answer natural language questions and generate images corresponding to them across a wide range of visual domains. We have also developed high-performance AI models for specific tasks such as playing Go, protein folding, and species identification from images and sound. These models are increasingly being deployed for analyzing vast datasets for scientific analysis and decision-making in novel ways.

However, deploying AI models for specific applications remains challenging. First, fields like astronomy and medical imaging are constrained by the amount of data they can acquire and label. Similarly, on citizen-science platforms like iNaturalist, despite millions of shared images, fewer than a few thousand species have sufficient obser-

vations out of the possible millions of plant and animal species that exist on our planet. This makes it challenging to train or adapt models with standard supervised learning. Second, current AI models are far from perfect. They may introduce bias or have unacceptable error rates. Even worse, their performance may be unpredictable when deployed in new domains. This makes it difficult to use AI models for scientific or high-stakes applications, where precise characterization of failure modes or statistical guarantees in performance is essential.

My ongoing research aims to improve the usability and precision of AI systems. I have explored the use of unlabeled and coarsely labeled data to train models for recognition tasks such as species recognition from images, audio, and text, as well as for segmenting parts and finding correspondences in both 2D and 3D objects [7, 11, 19, 22, 44–50, 55, 61, 62]. I have developed techniques to model and relate computer vision tasks that can be used to solve meta-tasks such as dataset discovery, model transfer, and task grouping for multi-tasking [1, 15]. Additionally, I have developed techniques for 3D shape generation that provide better control in the generative process through the use of shape handles, images, and even line drawings [6, 17]. Finally, I have developed human-in-the-loop recognition techniques that can estimate quantities of interest to any desired level of precision when deploying imperfect AI systems [40, 42]. I will briefly highlight each of these directions and explain how they connect to various applications in the next section.

Fine-grained categorization Recognizing species of birds, or makes and models of cars is challening because the subtle differences between categories are confounded by factors such as pose, viewpoint, and occlusion. My early research focused on texture understanding and generation and developed techniques that combined the benefits of deep learning with classical orderless texture representations [12, 13], greatly improving performance on texture and material understanding tasks. At ICCV 2015, we proposed bilinear CNN [26, 27], a deep architecture that offered the effectiveness of part-based representations, the dominant approach at that time, but did not require part annotations. The key idea was a factorization of the representation as a product of two learnable deep network representations designed to capture localized interactions. We also showed that the model is related to bag-of-words representations and their deep variants [13, 14] and could be trained in an end-to-end manner. The work was influential in the design of several architectures for fine-grained classification, and the idea of combining of information from multiple streams through product interactions has found its use in visual question answering and activity recognition. Interestingly, these product interactions are ubiqutous in the attention layers of Transformers.

Recent work from the community has emphasized the role of large-scale, high-quality training datasets for strong performance. In particular, the iNaturalist datasets (2018 to 2021) combined with ResNets have become a high-performance baseline. However, the iNat21 dataset only covers a few thousand species, leaving a large number of species in the long tail. Our ECCV'20 paper [56] explored the role of unlabeled datasets in improving the few-shot performance of these models. We found that existing benchmarks for semi and self-supervised learning are limited; they assume that images are evenly distributed across the (unknown) classes and that the unlabeled data does not contain images from classes not in the test set. These conditions are rarely met in practice, especially in fine-grained domains. For example, while it is easy to find images of Sparrows, it is more difficult to ensure that are images of Brewer's sparrow.

Our CVPR'21 paper [54] performed an extensive evaluation of existing semi-supervised learning methods on two novel benchmarks and found them to be brittle in the presence of class imbalance and novel classes. We proposed a distillation-based self-training approach that was more robust. These datasets were part of the challenges in the FGVC workshop in 2021 and 2022. In a follow-up BMVC'21 paper [55], we proposed ways to incorporate taxonomic labels within existing semi-supervised learning frameworks to further improve their robustness, as seen in Fig. 1.

Beyond categorization My recent research has explored the role of self-supervision and coarse supervision beyond classification tasks, extending to part segmentation and landmark detection tasks. Obtaining pixel labels or landmark annotations is time-consuming, but being able to estimate these labels from images can enable applications in image editing, animation, and body shape and pose estimation for animal monitoring. Our ICCV'21 work [11] developed a framework for learning landmarks from a collection of images of a category based on equivariant and invariant learning. We observed that most self-supervised learning techniques focus on learning representations invariant to photometric and geometric transformations. However, the emergence of invariance is

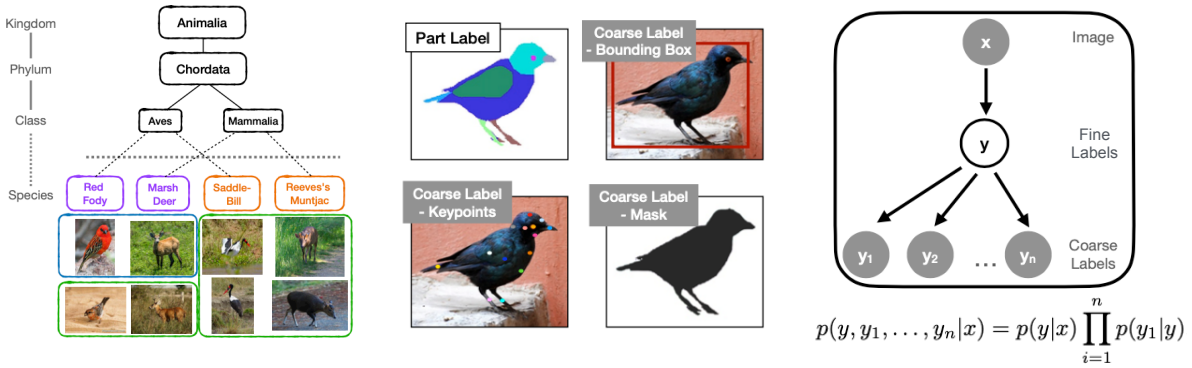

Figure 1: Learning from coarse supervision such as taxonomic labels (left) and annotation styles (middle). In both cases, we assume that the coarse labels are conditionally independent of the image given the fine label, as shown by the graphical model (right). Our recent work [8, 45, 55] learns to model the conditional distributions among labels, allowing learning from diverse and coarsely labeled data.

gradual in the layers of a convolutional network; early layers are nearly equivariant. By combining features from multiple layers of the network and training a lightweight network to be spatially equivariant, we obtained detailed spatial representations of objects. These representations can be used for image matching and landmark prediction. Our approach was significantly faster and more accurate than existing methods.

While unsupervised learning is attractive, there are often labeled datasets corresponding to related tasks that could be used to improve performance. For example, we found that different researchers had labeled bird roosts in radar images to conduct specific studies. However, the labels have widely different annotation styles; for example, some drew smaller bounding boxes than others, and they were across different subsets of data. Training a model by combining these datasets was not effective as the performance metrics and training objectives were affected by the heterogeneous annotation styles. In our AAAI'20 paper [8], we developed a latent variable model to account for the annotation styles, which resulted in a better model and meaningful performance metrics.

Our ECCV'22 paper [45] extended the framework to learn part segmentations by utilizing datasets consisting of coarse labels such as bounding boxes, figure-ground masks, and keypoint annotations (Fig. 1). We developed a probabilistic framework to infer the latent fine labels given coarse labels and the image in an iterative framework, where we also learned how to model the relationships between fine and coarse labels. We assumed that coarse labels are conditionally independent given the fine labels and parameterized these mapping using neural networks. The paper developed an amortized inference scheme, which made learning efficient and allowed the use of existing black-box models. This led to more accurate models than multi-tasking and transfer learning approaches.

Generative and contrastive learning In another line of work, we compared features from contrastively trained models with generative models such as GANs for image segmentation [44]. Both techniques learn from unlabeled data in different ways. At the time, several methods proposed using GANs to generate synthetic data for training but lacked a comparison with contrastive approaches. In our paper, we developed a strong baseline for using contrastively trained representations for segmentation and demonstrated that they are more effective than GAN-based representations. Additionally, contrastive representations were faster to train and simpler to use. We also highlighted that certain biases get amplified when training models on AI-generated datasets.

The debate over whether generative models are effective representation learners for images has resurfaced with the development of diffusion-based generative models, particularly stable diffusion. Recent research has shown that the features from stable diffusion models possess different properties than those from contrastively trained representations, such as DINO, and can be combined for added benefits.

Learning from text and large language models (LLMs) Arguably, the biggest improvement in image recognition systems has come from utilizing vast amounts of image and text data on the web as a source of supervision. For example, CLIP shows strong zero-shot and few-shot performance across existing benchmarks. Several techniques

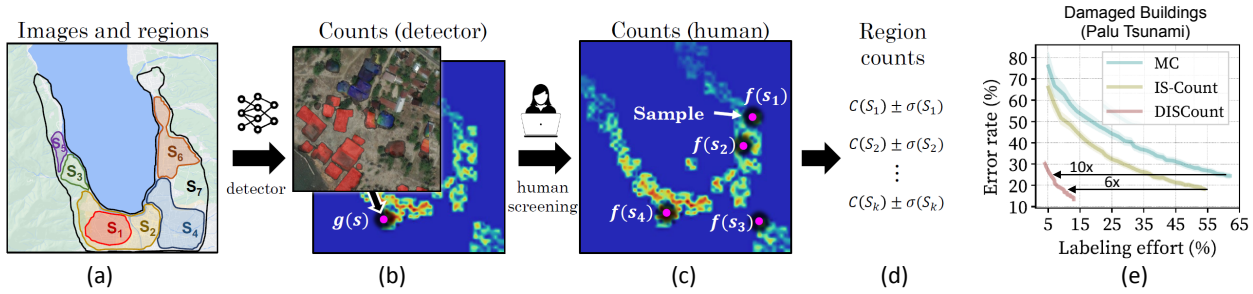

Figure 2: DISCOUNT uses detector-based importance sampling for estimation. **(a)** Geographical regions $S_1, S_2, \ldots, S_7$ where we want to estimate counts of damaged buildings. **(b)** A building damage detector is applied on satellite imagery to obtain *approximate* counts $g(s)$ for each tile (shown as dots). **(c)** Tiles selected for screening to determine true counts $f(s)$. **(d)** Counts (C) and confidence intervals (σ) are estimated for all regions. **(e)** DISCOUNT outperforms Monte Carlo and covariate-based sampling.

have been proposed to further improve performance using language data obtained from the web or large language models (LLMs) such as GPT4. However, training or fine-tuning CLIP when performance is poor requires large datasets of aligned language and image data, which can be hard to obtain.

Our recent CVPR'24 work [46] proposed combining fine-grained image datasets such as iNaturalist with summaries of categories obtained from LLMs to fine-tune CLIP and related vision-language models. This form of supervision is easy to generate (e.g., using Wikipedia pages or prompting LLMs appropriately); however, the supervision is coarse because sets of images are aligned with sets of text descriptions. We found that with a suitable change in the loss formulation, we can fine-tune CLIP, resulting in much higher zero-shot and few-shot performance in fine-grained domains.

More recently, we have been investigating ways to combine location data from observations, image data, and language data describing habitat and range preferences of animals to better estimate the distribution of particular species across the earth.

Statistical estimation with AI and humans in-the-loop Many applications use CV to detect and count objects in massive image collections. However, automated methods may fail to deliver accurate counts, especially when the task is very difficult or requires a fast response time. For example, during disaster response, aid organizations aim to quickly count damaged buildings in satellite images to plan relief missions, but pre-trained building and damage detectors often perform poorly due to domain shifts. Similarly many applications in ecology and remote sensing aim to make measurements (e.g., estimating populations across time, or surveying species diversity) to answer science or policy questions (e.g., which North American bird species are declining fastest?). In such cases, there is a need for human-in-the-loop approaches to accurately count with minimal human effort.

Our AAAI'24 work [40] developed DISCOUNT– a detector-based importance sampling framework for counting in large image collections (Fig. 2). DISCount uses an imperfect detector and human screening to estimate low-variance unbiased counts. We propose techniques for counting over multiple spatial or temporal regions using a small amount of screening and estimate confidence intervals. This enables end-users to stop screening when estimates are sufficiently accurate, which is often the goal in real-world applications. We demonstrated our method with two applications: counting birds in radar imagery to understand responses to climate change, and counting damaged buildings in satellite imagery for damage assessment in regions struck by a natural disaster (see Fig. 2). On the technical side we developed variance reduction techniques based on control variates and proved the (conditional) unbiasedness of the estimators. DISCOUNT leads to a 9-12× reduction in the labeling costs to obtain the same error rates compared to naive screening for tasks we consider, and surpasses alternative covariate-based screening approaches. The work was awarded the Best Paper in the AI for Social Impact Track at AAAI'24.

Our ECCV'24 paper [42] extended this framework for estimating cluster counts using a pairwise similarity model. This facilitates population size estimation by combining imperfect Re-ID systems with small amounts of human

vetting. We introduced a nested importance sampling approach which results outperforms existing active clustering approaches on a variety of animal Re-ID tasks.

Understanding deep networks Some of my research has focused on understanding the properties of neural networks to inform design choices, especially when data is limited. In CVPR'19 [10], we developed a Bayesian interpretation of the "deep image prior" work by analyzing the properties of convolutional networks with random parameters. We showed that random networks exhibit spatial smoothness properties in their generated outputs, which can be precisely characterized by a Gaussian process (GP) whose mean and covariance depend on the network architecture. This allowed us to design network architectures with different smoothness preferences and improve inference for denoising tasks.

In a follow-up work, we precisely characterized the spectral bias, i.e., the preference for learning smooth functions, and how to control it [51]. We also extended the work to describe manifold data such as point clouds and demonstrated how to estimate smooth surfaces from point clouds by fitting a neural network to them [22], or estimating 3D shapes from sparse views using silhouettes or depth maps [21]. Such methods have become increasingly common with the introduction of NeRFs, where the architecture itself provides a smoothness prior.

Modeling and relating visual tasks Our TASK2VEC paper [1] proposed a technique to map CV tasks into embeddings in Euclidean space, where distance in the embedding space reflects task similarity. A recently-awarded NSF grant aims to improve the framework to better undestand the space of tasks, and to solve meta-tasks such as dataset discovery, modeling transfer, and multi-tasking. With the growth of publicly available datasets and models, being able to quickly find similar tasks and their solutions can enable one to quickly develop solutions for the task in hand. Our CVPR'24 work [15] developed TASK2BOX which replaces Euclidean embeddings with box embeddings, allowing of assymetic relations between tasks such as transfer and containment. We used this framework to predict relations between novel tasks as well as to visualize publicly-available computer vision datasets. Unlike t-SNE, box embeddings allow better visualization of their hierarchical relationships.

3D shape understanding and generation There is a growing need to analyze and generate 3D shape data for applications in computer graphics, robotics and autonomous driving. However, existing datasets and techniques for processing 3D data are lacking in comparison to those for image data. My early research at UMass, in collaboration with several colleagues and graduate students advanced techniques for 3D shape analysis and synthesis. For example, we developed the multi-view CNN [53] for view-based representations of 3D shapes. The architecture is end-to-end trainable and benefits from transferable representations learned on large-scale image datasets. Although several new techniques were proposed since, a survey presented at an ECCV 2018 workshop [57] showed that multi-view architectures outperform these techniques when they are combined with the latest image classification networks. Multi-view representations were also effective for 3D shape segmentation and other tasks.

My recent work has focused on learning 3D representations for segmentation and correspondence tasks with less supervision. In our ECCV'20 paper [19], we demonstrated that using convex decomposition of 3D shapes as a proxy task leads to effective representations for part segmentation. This approach was based on the observation that most parts of natural and man-made objects are convex due to physical constraints or ease of assembly. Our results for few-shot segmentation on the ShapeNet dataset were state-of-the-art at the time of publication.

Building on this, we incorporated convex decomposition within the network architecture through an end-to-end primitive fitting [47]. This simplified the learning and offered flexibility in choosing primitives. Unlike our ECCV'20 work, which relied on an off-the-shelf approximate convex decomposition library, this new framework was based on our PARSENET approach [49], which developed a parametric surface fitting method for points.

In ECCV'22, we introduced MVDECOR [50], a framework that leverages view-based representations for 3D shape correspondence tasks. This approach utilized equivariant learning across images by rendering the 3D data and enforcing geometric consistency. However, the pipeline is most effective in domains with high-quality rendering engines, such as human bodies.

My recent work has also focused on providing users with better control when generating 3D data. While unconditional and text-based 3D shape generation have been successful, content creators use a wide range of tools for

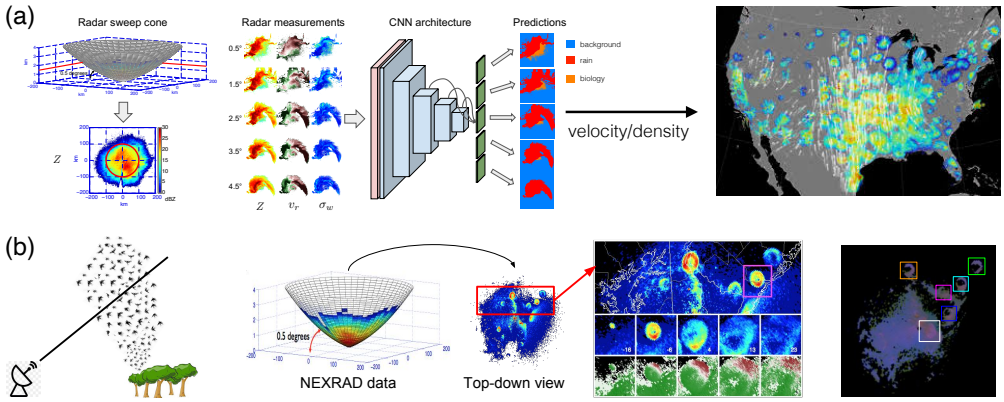

Figure 3: Extracting biological information from weather radars. (a) We developed MistNet [28], a deep network capable of separating biology from rain, allowing us to estimate the density and velocity of flying animals at the continental scale and across decades. (b) The exodus of birds, bats, and insects can be detected on weather radars using CV models we developed [8, 43], allowing us to study phenomena tied to specific species.

3D editing, as indicated by the complexities of software such as 3DS Max and Maya. In our CVPR'20 paper [17], we developed a generative model for shape handles—lightweight primitives that allow for explicit control over the 3D shape. Examples include rigging cages and primitives, which are often used by 3D content creators to manipulate shapes efficiently. Our paper introduced an auto-regressive model to generate sets of primitives that approximate the 3D shape. The model was trained without explicit supervision in an end-to-end manner using a reconstruction objective. A side effect of this learning process was that the model implicitly learned correspondences and semantic parts across a shape collection.

More recently, our ECCV'22 work [6] developed a method to edit 3D shapes using images, sketches, and text in a unified manner. Our model learned a latent space over the 3D shapes and mappings from individual modalities to the latent space, allowing one to combine feedback across multiple modalities seamlessly.

Ultimately, these models are limited by the availability of 3D data, which is much harder to collect than images. In our "Accidental Turntables" project [9], we explore the use of videos of rotating objects to collect 3D data using off-the-shelf object detection, tracking, and structure-from-motion techniques. While these methods sometimes fail, we demonstrate that the data can be used to train models for relative pose estimation, offering robustness that training on synthetic data does not provide. We have also explored how 3D shapes can be estimated from a few views of an object by aligning local NeRF-based representations and synchronizing poses.

2 Applications

Radar aeroecology Weather radar networks offer an unprecedented opportunity to study dynamic, continental-scale movements of animals over a long time. For example, the U.S. National Weather Service operates a network of 143 radars in the contiguous U.S, with data archive dates from the early 1990s. These radars were designed to study weather phenomena, such as precipitation and severe storms, but are very sensitive and turn out to also detect flying animals, including birds, bats, and insects. Radars can also detect phenomena that can be matched to individual species, such as insect hatches and departures of large flocks of birds or bats from roosting locations. However, the scope of ecological research using weather radar has historically been limited due to the difficulty of accessing the biological information in the data.

I lead the computer vision efforts for a NSF funded dark ecology project — a research collaboration between the University of Massachusetts and the Cornell Lab of Ornithology. Together with collaborators, in particular Prof. Daniel Sheldon at UMass, we developed new methods to measure and predict biological activity in the U.S. weather radar data and conducted long-term large-scale analyses of bird migration. We developed MistNet [28], an AI system to discriminate between precipitation and biology in radar data and enable massively scalable analyses

(Fig. 3(a)). The project led to a number of high impact publications, including the first continent-scale radar analyses to document multi-decadal decline shifts in timing of bird migration in North America [24]. The work has been covered by dozens of news outlets and inspired community-wide conservation efforts.

Another, recently funded NSF project in collaboration with University of Colarado and Oklahoma State University, aims to (1) understand how global environmental change has impacted seasonal timing and population abundance of aerial insectivores over the past twenty-five years and (2) determine drivers of recent within and between seasonal variation in timing and abundance. Aerial insectivore populations have shown precipitous declines in the last half century — often at much steeper rates than other aerial taxa. Understanding mechanisms driving these changes would have broad implications for hundreds of species of birds, bats, and insects, and serve as an indicator of terrestrial and aquatic ecosystem health.

Over the last few years we have developed a CV system to detect signatures of roosts in radar data [8], and a pipeline to process large-scale and long-term data in the Great Lakes area in the U.S. (Fig. 3(b)). This work proposed a benchmark and developed ways to train Faster R-CNN detectors using annotations with varying labeling styles as we described earlier. We also developed a model that uses spatio-temporal information (e.g., previous two scans) which has better precision at the task [43]. The data from the models has also led to two scientific publications [2, 16] in high-impact journals.

While CV made it possible to perform the analysis, it still took more than 180 human hours to screen the model ouputs across nearly 600,000 radar scans in the Great Lakes region. Despite years of effort from the team the detector performance is far from perfect. This is often a sitution a scientist faces when using off-the-shelf AI system which does not have high-enough precision for their needs. This motivates methods like DISCOUNT which allow end users to estimates quantities of interest with statistical guarantees for a given amount of effort. Our ongoing work aims to deploy DISCOUNT within the labeling UI to enable scaling of the study to the entire US archive to unlock new biological information.

Astronomy I collaborate with Astronomer Prof. Daniela Calzetti to develop algorithms for search, classification, and shape measurement of young star clusters in high resolution images of galaxies. We developed STARCNET, a deep network capable of classifying star clusters into various morphological types and achieving accuracy comparable to human experts [41]. Our initial tests were performed on the two closest galaxies to our own Milky Way (M31 and M33), and then extended to M51 and NGC628, which are further away from us. These are well-studied galaxies for which high-fidelity catalogs already exist, and can be used for training and evaluation of automatic systems. Subsequently, we used the model to analyze star cluster formation and evolution in the M101 galaxy [29].

The recent launch of the James Webb Space Telesope (JWST) is paving the way to answer some of the remaining questions in star formation: 1) what determines the (low) efficiency of star formation in galaxies? Is this regulated at the local (cloud) or global (galaxy) level? 2) Which are the dominant mechanisms that enable recently formed stars to emerge from their natal cocoons of dust and gas? 3) How do those mechanisms depend on the physical parameters of the stellar populations and on the galactic environment? A recently funded NSF proposal will help us answer some of these questions by combining the data from ALMA, JWST and Hubble Space Telescope and our previously developed AI tools to search for and identify star clusters across various stages of evolution. We piloted the tool in some initial studies [39] to study the feedback mechanisms in star formation [23].

Predicting material properties for separation processes Separation of chemical mixtures accounts for more than 10% of global energy consumption, driven heavily by thermal processes such as distillation. Replacing these energy-intensive traditional separation processes with more efficient alternatives is projected to eliminate 100 million tonnes of CO2 emissions and save billions of dollars in energy costs. Among the alternatives, separation using nanoporous materials, for example in an adsorption or membrane-based setup, can be an order of magnitude more efficient. Despite their great promise, identifying the optimal material out of the large pool of candidate structures for a given separation task requires significant resources and prolongs the development cycle.

Ongoing work in collaboration with Prof. Peng Bai in Chemical Engineering, initially funded by Climate Change AI Foundation, develops AI techniques to predict optimal materials for targeted separation applications (Fig. 4). Our primary focus is on nanoporous zeolites, for which we already have several existing datasets and the capability

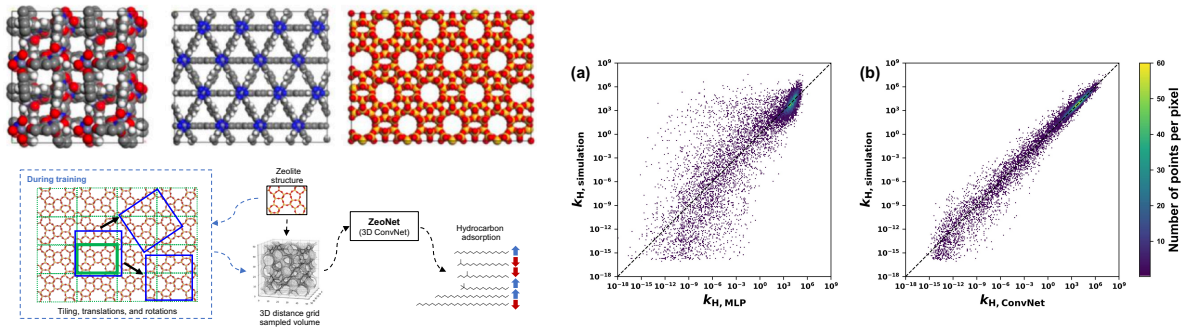

Figure 4: Zeolites have nanometer-sized pores that can adsorb and react with molecules selectively. We developed ZeoNet [30], a 3D ConvNet that can predict the adsorption properties of target molecules. Compared to the geometric features shown on the right (a), the ConvNet's predictions are much more accurate (b).

to generate a new dataset relevant for CO2 capture. We developed ZeoNet [30], a representation learning framework using ConvNets and 3D volumetric representations for predicting adsorption in zeolites. ZeoNet was trained on the task of predicting Henry's constants for adsorption, k_H, of n-octadecane in more than 330,000 known and predicted zeolite materials. Employing a 3D grid based on the distances to solvent-accessible surfaces, a volumetric representation that can be generated efficiently, the best-performing model achieved a correlation coefficient $r^2 = 0.977$ and a mean-squared error MSE = 3.8 in $\ln k_H$, which corresponds to an error of 9.3 kJ/mol in adsorption free energy. In comparison, a model based on hand-designed geometric features has values of $r^2 = 0.777$ and MSE = 36.6. ZeoNet is also relatively efficient and can process 8 structures per second on an Nvidia RTX 2080TI GPU, orders of magnitude faster than forcefield simulations.

These results provide benchmark quality data and comprehensive guidelines for using 3D ConvNets to model porous materials. Our ongoing work aims to extend these ideas to a broader class of materials such as metal organic frameworks (MOFs) and learn to predict their properties for other tasks such as CO2/N2 and ethanol/water separation tasks at multiple pressures. On the representation learning side we are investigating equivariant representations and self-supervised learning to make training label efficient.

Mapping and monitoring rivers networks NASA's recently launched SWOT mission promises a sea change for terrestrial hydrology. Principally, SWOT's reservoir/lake volume change observations and SWOT's derived river discharge product are each unprecedented in terms of their resolution, scale, and frequency. I lead the computer vision efforts within a large collaborative project across Umass Amherst, University of Pittsburgh, University of North Carolina, and NASA, which seeks to integrate data from the soon-to-be launched SWOT mission with traditional optical imagery from Landsat and Sentinel-2 into a common platform to dramatically and uniquely advance our understanding of the world's river water quality and quantity, informing the management and use [31].

Beyond these projects, my research group is developing methods to estimate damaged buildings for disaster response planning, landcover mapping, biodiversity estimation from remote sensing data, as well as general tools that can enable interactive estimation building on DISCOUNT*, test-time adaptation, active learning, and language models.*

3 Future Work and Conclusion

While we have made significant progress in advancing AI in recent years, several technological and social challenges must be addressed for its widespread adoption. My research tackles several key aspects: generalization, robustness, and usability of AI systems in visual domains.

In core computer vision, I aim to improve our understanding of spatio-temporal domains such as video, audio, and remote sensing data. This includes object detection, tracking, 3D shape estimation, depth estimation, as well as various temporal reasoning tasks. Active learning, labeling, and statistical estimation in these domains present unique challenges as individual frames cannot be treated independently. I also plan to continue our research on how to combine language and vision for multi-modal reasoning and for learning generalizable representations.

I plan to strengthen collaboration with domain experts to better understand the role of AI within scientific fields. AI, particularly computer vision, has benefited from benchmarks that allow systematic evaluation and adoption of methods. However, the utility of standard benchmarks such as CIFAR and ImageNet is becoming limited. I intend to introduce better benchmarks and tasks inspired by real-world use cases and problems of societal interest, both through publications and workshops such as FGVC and CV4Science, which I help organize.

For scientific tasks, the goal often involves performing measurements on a large, but finite collection of images. My ongoing and future work aims to better incorporate human feedback to improve estimations when using AI models. Human effort can be utilized in multiple ways, such as active learning to improve model performance, data collection in conjunction with semi-supervised learning, providing detailed feedback or labels, or direct statistical estimation, as we proposed in DISCOUNT. Each of these tasks has different associated costs, and currently, no framework allows end-users to explore these options for their specific problems without requiring significant expertise. Through collaborations, I aim to understand these trade-offs for use cases in radar aeroecology, astronomy, and remote sensing tasks.

I am also interested in advancing tools for fine-grained recognition to better monitor biodiversity. Two ongoing collaborations with Prof. Van Horn and the iNaturalist team focus on building benchmarks and models for recognizing animal species based on sound, and incorporating observation data across thousands of species and their habitat and range preferences extracted from text to develop better species distribution models.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task Embedding for Meta-Learning. In *International Conference on Computer Vision (ICCV)*, 2019.

[2] Maria Carolina TD Belotti, Yuting Deng, Wenlong Zhao, Victoria F Simons, Zezhou Cheng, Gustavo Perez, Elske Tielens, Subhransu Maji, Daniel Sheldon, Jeffrey F Kelly, et al. Long-term analysis of persistence and size of swallow and martin roosts in the us great lakes. *Remote Sensing in Ecology and Conservation*, 9(4):469–482, 2023.

[3] Lubomir Bourdev, Subhransu Maji, Thomas Brox, and Jitendra Malik. Detecting People using Mutually Consistent Poselet Activations. In *European Conference on Computer Vision (ECCV)*, 2010.

[4] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing People: Poselet-Based Approach to Attribute Classification. In *International Conference on Computer Vision (ICCV)*, 2011.

[5] Zhaowei Cai, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Zhuowen Tu, and Stefano Soatto. Exponential moving average normalization for self-supervised and semi-supervised learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. Cross-modal 3d shape generation and manipulation. In *European Conference on Computer Vision (ECCV)*, 2022.

[7] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. LU-NeRF: Scene and Pose Estimation by Synchronizing Local Unposed NeRFs. In *International Conference on Computer Vision (ICCV)*, 2021.

[8] Zezhou Cheng, Saadia Gabriel, Pankaj Bhambhani, Daniel Sheldon, Subhransu Maji, Andrew Laughlin, and David Winkler. Detecting and tracking communal bird roosts in weather radar data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[9] Zezhou Cheng, Matheus Gadelha, and Subhransu Maji. Accidental turntables: Learning 3d pose by watching objects turn. In *International Conference on Computer Vision Workshops (ICCW)*, 2023.

[10] Zezhou Cheng, Matheus Gadelha, Subhransu Maji, and Daniel Sheldon. A Bayesian Perspective on the Deep Image Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *International Conference on Computer Vision (ICCV)*, 2021.

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep Filter Banks for Texture Recognition, Description, and Segmentation. *International Journal of Computer Vision (IJCV)*, January 2016.

[14] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep Filter Banks for Texture Recognition and Description. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[15] Rangel Daroya, Aaron Sun, and Subhransu Maji. Task2Box: Box Embeddings for Modeling Asymmetric Task Relationships. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[16] Yuting Deng, Maria Carolina TD Belotti, Wenlong Zhao, Zezhou Cheng, Gustavo Perez, Elske Tielens, Victoria F Simons, Daniel R Sheldon, Subhransu Maji, Jeffrey F Kelly, et al. Quantifying long-term phenological patterns of aerial insectivores roosting in the great lakes region using weather surveillance radar. *Global Change Biology*, 29(5):1407–1419, 2023.

[17] Matheus Gadelha, Giorgio Gori, Duygu Ceylan, Radomír Mech, Nathan Carr, Tamy Boubekeur, Rui Wang, and Subhransu Maji. Learning generative models of shape handles. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] Matheus Gadelha, Subhransu Maji, and Rui Wang. 3D Shape Induction from 2D Views of Multiple Objects. In *International Conference on 3D Vision (3DV)*, 2017.

[19] Matheus Gadelha, Aruni RoyChowdhury, Gopal Sharma, Evangelos Kalogerakis, Liangliang Cao, Erik G. Learned-Miller, Rui Wang, and Subhransu Maji. Label-efficient learning on point clouds using approximate convex decompositions. In *European Conference on Computer Vision (ECCV)*, 2020.

[20] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[21] Matheus Gadelha, Rui Wang, and Subhransu Maji. Shape Reconstruction using Differentiable Projections and Deep Priors. In *International Conference on Computer Vision (ICCV)*, 2019.

[22] Matheus Gadelha, Rui Wang, and Subhransu Maji. Deep manifold prior. In *International Conference on Computer Vision Workshops, (ICCVW)*, 2021.

[23] Benjamin Gregg, Daniela Calzetti, Angela Adamo, Varun Bajaj, Jenna E Ryon, Sean T Linden, Matteo Correnti, Michele Cignoni, Matteo Messa, Elena Sabbi, et al. Feedback in emerging extragalactic star clusters, feast: The relation between 3.3μm pah emission and star formation rate traced by ionized gas in ngc 628. *The Astrophysical Journal (ApJ)*, 2024.

[24] Kyle G Horton, Frank A La Sorte, Daniel Sheldon, Tsung-Yu Lin, Kevin Winner, Garrett Bernstein, Subhransu Maji, Wesley M Hochachka, and Andrew Farnsworth. Phenology of nocturnal avian migration has shifted at the continental scale. *Nature Climate Change*, 10(1):63–68, 2020.

[25] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN Models for Fine-grained Visual Recognition. In *International Conference on Computer Vision (ICCV)*, 2015.

[27] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017.

[28] Tsung-Yu Lin, Kevin Winner, Garrett Bernstein, Abhay Mittal, Adriaan M. Dokter, Kyle G. Horton, Cecilia Nilsson, Benjamin M. Van Doren, Andrew Farnsworth, Frank A. La Sorte, Subhransu Maji, and Daniel Sheldon. MistNet: Measuring historical bird migration in the US using archived weather radar data and convolutional neural networks. *Methods in Ecology and Evolution*, 2019.

[29] ST Linden, G Perez, Daniela Calzetti, S Maji, Matteo Messa, BC Whitmore, Rupali Chandar, Angela Adamo, K Grasha, DO Cook, et al. Star cluster formation and evolution in m101: An investigation with the legacy extragalactic uv survey. *The Astrophysical Journal*, 935(2):166, 2022.

[30] Yachan Liu, Gustavo Perez, Zezhou Cheng, Aaron Sun, Samuel C Hoover, Wei Fan, Subhransu Maji, and Peng Bai. Zeonet: 3d convolutional neural networks for predicting adsorption in nanoporous zeolites. *Journal of Materials Chemistry A*, 11(33):17570–17580, 2023.

[31] Luisa Vieira Lucchese, Rangel Daroya, Travis Simmons, Punwath Prum, Subhransu Maji, Tamlin Pavelsky, Colin Gleason, and John Gardner. Modeling suspended sediment concentration using artificial neural networks, an effort towards global sediment flux observations in rivers from space. Technical report, Copernicus Meetings, 2024.

[32] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3d shape reconstruction from sketches via multi-view convolutional networks. In *International Conference on 3D Vision (3DV)*, 2017.

[33] Subhransu Maji. Discovering a Lexicon of Parts and Attributes. In *Second International Workshop on Parts and Attributes, ECCV*, 2012.

[34] Subhransu Maji, Alexander Berg, and Jitendra Malik. Classification using Intersection Kernel Support Vector Machines is Efficient. In *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[35] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action Recognition from a Distributed Representation of Pose and Appearance. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[36] Subhransu Maji, Tamir Hazan, and Tommi Jaakkola. Efficient Boundary Annotation using Random Maximum A-Posteriori Perturbations. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

[37] Subhransu Maji and Jitendra Malik. Object Detection using a Max-margin Hough Transform. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[38] Subhransu Maji and Gregory Shakhnarovich. Part Discovery from Partial Correspondence. In *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[39] Gustavo Perez, Sean Linden, Timothy McQuaid, Matteo Messa, Daniela Calzetti, and Subhransu Maji. An ai-assisted labeling tool for cataloging high-resolution images of galaxies. In *AI for Science: Progress and Promises, NeurIPS*, 2022.

[40] Gustavo Pérez, Subhransu Maji, and Daniel Sheldon. Discount: Counting in large image collections with detector-based importance sampling. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

[41] Gustavo Pérez, Matteo Messa, Daniela Calzetti, Subhransu Maji, Dooseok E Jung, Angela Adamo, and Mattia Sirressi. Starcnet: Machine learning for star cluster identification. *The Astrophysical Journal*, 907(2):100, 2021.

[42] Gustavo Perez, Grant Van Horn, Daniel Sheldon, and Subhransu Maji. Human-in-the-loop visual re-id for population size estimation. In *European Conference on Computer Vision (ECCV)*, 2024.

[43] Gustavo Perez, Wenlong Zhao, Zezhou Cheng, Maria Carolina TD Belotti, Yuting Deng, Victoria F Simons, Elske Tielens, Jeffrey F Kelly, Kyle G Horton, Subhransu Maji, et al. Using spatiotemporal information in weather radar data to detect and track communal roosts. *Remote Sensing in Ecology and Conservation*, 2024.

[44] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. GANORCON: are generative models useful for few-shot segmentation? In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[45] Oindrila Saha, Zezhou Cheng, and Subhransu Maji. Improving few-shot part segmentation using coarse supervision. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *European Conference on Computer Vision (ECCV)*, 2022.

[46] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, 2024.

[47] Gopal Sharma, Bidya Dash, Aruni Roy Chowdhury, Matheus Gadelha, Marios Loizou, Liangliang Cao, Rui Wang, Erik G. Learned-Miller, Subhransu Maji, and Evangelos Kalogerakis. Prifit: Learning to fit primitives improves few shot point cloud segmentation. *Comput. Graph. Forum*, 41(5):39–50, 2022.

[48] Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. Neural shape parsers for constructive solid geometry. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(5):2628–2640, 2022.

[49] Gopal Sharma, Difan Liu, Subhransu Maji, Evangelos Kalogerakis, Siddhartha Chaudhuri, and Radomír Mech. Parsenet: A parametric surface fitting network for 3d point clouds. In *European Conference on Computer Vision (ECCV)*, 2020.

[50] Gopal Sharma, Kangxue Yin, Subhransu Maji, Evangelos Kalogerakis, Or Litany, and Sanja Fidler. Mvdecor: Multi-view dense correspondence learning for fine-grained 3d segmentation. In *European Conference on Computer Vision (ECCV)*, 2022.

[51] Zenglin Shi, Pascal Mettes, Subhransu Maji, and Cees G. M. Snoek. On measuring and controlling the spectral bias of the deep image prior. *Int. J. Comput. Vis.*, 130(4):885–908, 2022.

[52] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. SPLATNet: Sparse lattice networks for point cloud processing. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[53] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view Convolutional Neural Networks for 3D Shape Recognition. In *International Conference on Computer Vision (ICCV)*, 2015.

[54] Jong-Chyi Su, Zezhou Cheng, and Subhransu Maji. A realistic evaluation of semi-supervised learning for fine-grained classification. In *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[55] Jong-Chyi Su and Subhransu Maji. Semi-supervised learning with taxonomic labels. In *British Machine Vision Conference (BMVC)*, 2021.

[56] Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? In *European Conference on Computer Vision (ECCV)*, 2020.

[57] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *Second Workshop on 3D Reconstruction Meets Semantics, ECCV*, 2018.

[58] Jong-Chyi Su and Subhransu Maji. Adapting models to signal degradation using distillation. In *British Machine Vision Conference (BMVC)*, 2017.

[59] Jong-Chyi Su, Chenyun Wu, Huaizu Jiang, and Subhransu Maji. Reasoning about fine-grained attribute phrases using reference games. In *International Conference on Computer Vision (ICCV)*, 2017.

[60] Catherine Wah, Subhransu Maji, and Serge Belongie. Learning Localized Perceptual Similarity Metrics for Interactive Categorization. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2015.

[61] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[62] Chenyun Wu, Mikayla Timm, and Subhransu Maji. Describing textures using natural language. In *European Conference on Computer Vision (ECCV)*, 2020.