I develop computer vision and machine learning methods that learn structures from natural data, where I make learning mechanisms, models and learned representations all be aware of the structures of the data. Such methods could unleash new applications in medicine and active vision.

Computer vision (CV) and machine learning (ML), particular forms of artificial intelligence, are gradually changing the landscape of various fields and can be applied to the real world. For example, in healthcare and medicine, CV and ML have made tremendous development, especially when medical imaging is involved, such as disease diagnosis [1], [2], health monitoring [3], medical treatment [4], etc. I hold the view that CV and ML are crucial to facilitate accurate, efficient, and interpretable solutions to many science, medicine and engineering problems.

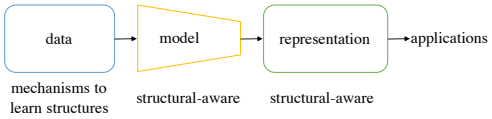
There are three main challenges towards the deployment of CV and ML models for real-world problems: **1)** While large-scale data is being produced daily, it needs to go through resource-consuming manual annotations before being ready to be used for training CV and ML models. **2)** Trade-offs between accuracy and efficiency exist in CV and ML models. The size of such models are growing exponentially for high accuracies in the past few years. However, for deployment and practical use in real-world problems, models need to be efficient: low latency with low computation cost. It is challenging to design small yet high-performance models. **3)** Existing CV models focus more on learning *what* (categories, semantics) rather than *where* (geometric structures) from images. However, such structures are important in many aspects, such as improving the interpretability of model predictions.

My research focuses on three perspectives to address these challenges:

1) Developing self-supervised and long-tailed mechanisms for learning structures from natural data

2) Improving training and inference efficiency with structure-aware models

3) Learning geometry-aware representations for understanding the environment and making actions



Structural-aware representation learning requires structural-aware models and mechanisms to preserve structures of natural data.

In the following, I summarize my prior works, followed by its application to medicine. I then introduce the future research problems that fascinate me, as well as my long-term research vision.

Self-Supervised Structure Learning from Natural Data (CVPR,TPAMI)

ML is a technique for recognizing patterns from data. Current works mostly rely on supervised learning to learn structures from data, i.e. human annotations are required for the collected data. In other words, each data sample of the dataset (e.g.images) should have a *correct answer*. Most high-accuracy supervised models require a large dataset that is completely annotated for a specific task to achieve satisfactory prediction accuracy. However, obtaining annotated data is often very costly or even infeasible in certain cases. Supervised models may also be limited to predicting predefined *correct answers*. Additionally, the model could be biased as labeled data may not cover the entire universe of potential labels.

We study self-supervised (unsupervised) models which have the following advantages: **1)** Data labels are not needed. Human labor, time and resources could be thus saved and the model can be trained on large-scale unlabeled datasets. **2)** Self-supervised models could go beyond predefined categories and discovers patterns from datasets in a data-driven manner.

Recently, self-supervised CV models have made great success in classifying object categories from an image, e.g. a cup is depicted in the image. In addition to the object category, we learn geometric information as well, e.g. the cup has a handle which is facing towards the user, such that the cup could be directly picked up. As a pioneer work, [5] **discovers the object category and pose from unlabeled images**. Applying self-supervised models to medicine also fascinates me. Our group [6] discovers novel phenotypes of meibomian gland (which resides along the rims of the eyelid) from images without any labels.

We study open long-tailed models to reduce model bias. Our visual world is inherently long-tailed [7], with a few common visual categories (i.e., head classes) and many more relatively rare categories (i.e., tail

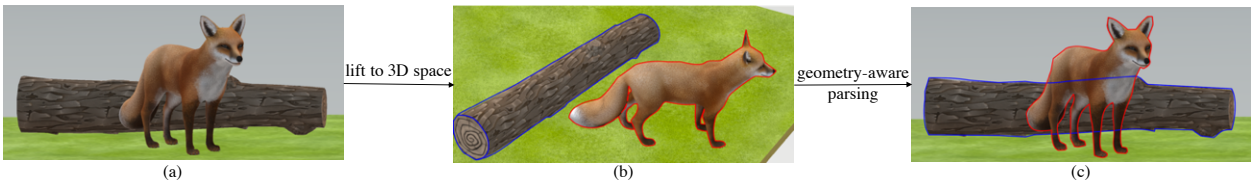
classes). Most observations are of common categories, and rare ones have very limited observations. For example, for meibomian gland atrophy, 86% subjects have no or very low atrophy, while only 1% subjects have severe atrophy [8]. The insufficient data leads to insufficient training of the model and low accuracy for rare categories. We for the first time formally studied *Open Long-Tailed Recognition* (OLTR) for natural data in computer vision, which was studied separately as imbalanced classification, few-shot classification and open-set recognition [9], [10]. We formally **define the OLTR task, develop benchmark datasets and evaluation protocols, as well as a novel system** that outperforms state-of-the-art approaches. We apply the OLTR framework to applications that deal with natural data, such as wildlife recognition [11].

Structure-Aware Models (CVPR,WACV)

The past few years have witnessed the exponential growth of the size and computation of ML models [12], [13]. While it is possible to use a much larger model to achieve similar performance, a leaner model would be more time and resource efficient, both in terms of model training and deployment. Our key intuition is that to efficiently parse the structures of data, the model should be aware of its structures. We develop methods specifically designed to preserve data structures such as orthogonality and recurrence [13], [14].

Our lean and efficient models have the following advantages: **1)** They take **less time to train and optimize** as we use constrained and structure-aware neural optimization. **2)** They take **less computation power and responds faster during inference** as the model is much smaller. **3)** They are **robust to adversarial attacks** as such models are aware of data structures and less prone to overfit or be fooled. Applying the technique to existing medical AI models (supervised like [8] or unsupervised like [6]) could greatly reduce the model training and inference time.

Structure-Aware Representations (TPAMI)



Principles behind geometry-aware representation learning: A 2D image is lifted to 3D space by hallucination, and object organizations and their geometric relationships can be understood.

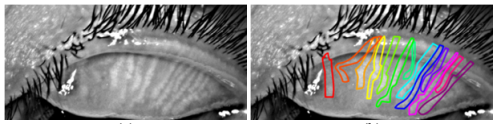
Two-streams hypothesis [15], a model of the neural processing of vision, argues that humans possess two visual systems, *what* (semantics) and *where* (geometric structures) pathways. We study joint *what* and *where* representation learning, or geometry-aware representation learning, unlike existing works [16], [17] that learn *what* representations from images. For example, in the figure above, existing representation learning methods label (a) as fox. Our geometry-aware representation considers the holistic structure (b) of the scene and understands it as a fox in front of a log (c).

There are two levels of geometry-aware representation learning: object level and scene level. Our previous work focuses on object-level representation learning. Specifically, we learn geometry-aware representations by equivariance in the learning mechanism [5], [18]. **1)** We demonstrate it is possible to **learn a joint representation that encodes both semantic and geometric information of an object without any labels. 2)** We show **the awareness of geometry and the awareness of semantics is benefiting each other**. For example, the approach unsupervisedly learns the symmetry of digits, which improves the semantic classification as symmetry serves as an indicator for discriminating digits like 6 and 9.

The geometry-aware representations not only enable downstream tasks like semantic and pose classification, but also generative tasks: We are able to generate objects and scenes with high fidelity from such representations [19], [20].

Medical Applications (Scientific Reports,TVST)

As a specific application, we develop CV and ML models for meibography, infrared imaging of the human eyelid for observation meibomian glands, whose dysfunction causes dry eye syndrome. Figure (a) on the right is a meibography image. Clinicians examine the morphology and atrophy of meibomian glands (highlighted in color in (b)) and decide the atrophy level.



(a) (b)

Learning morphology and geometry of glands from meibography images helps build an interpretable model - associations between morphological features and patient demographics, symptoms and more can be understood and provide knowledge for clinicians.

In our work [8], [21], we segment individual glands from the image, and use it for classification purposes, rather than directly learning predictions from images. The morphological features provide a **fast, quantitative and repeatable** way to measure the morphology of meibomian glands, e.g. the exact number of glands and the exact length of a specific gland. With the morphology analysis algorithm, we also learn to associate the morphological features with demographics, patient symptoms, clinical signs and diagnoses [2], [22]. For example, with the data-driven method, we gain knowledge such as glands with faint appearance are top indicators for predicting the patient's age.

Future Directions

To make a difference in the fields of healthcare, medical imaging and other real-world problems, we must shift the CV and ML paradigms. Currently, the research focuses on passive learning: learning patterns from existing data without considering much about actions and how actions affect the world. This could be not optimal as the world is dynamically changing, and ignoring actions and world states may lead to inaccurate predictions. Rather than re-discovering patterns from existing data, the future is active learning: a real-time system multi-modality decision system that is able to quickly understand the situation (the world state) and act. Towards that end, we need representation learning methods that learn effective actionable representations from dynamic scenes and multi-modality data. With such representations, a world model that understands world states conditioned on different actions can be built. It has wide applications ranging from healthcare to robotics.

My prior works on learning structure-aware representations provide basis for developing world models. Specifically, from natural data, we use structure-aware models to learn object-level geometry-aware representations in a self-supervised fashion. As a fundamental step towards world models, such representations need to be extended to scenes with multiple objects and multi-modality sensory inputs. In the following, I elaborate on how to make the learned representation more generalizable to encode dynamic world states, followed by how a preliminary world model could be built. I conclude the section with the potential applications to medicine and robotics.

Geometry-Aware Representations for Multi-Object Scenes. The first and essential step towards active vision is to build geometry-aware representations of the environment where multiple objects exist. Our previous work [5] unsupervisedly learns geometry-aware representations for a single object, which is not practical enough for modeling complicated environment. I plan to continue this line of research, and first extend the representation learning for multi-object systems in a simulated environment. I will then move on to learning representations of static scene images, where multiple objects and complicated backgrounds are present. Based on these experiences, I will finally learn representations from dynamic scenes/environments.

Learning from Multi-Modality Data. We live in a world where sensory data of different modalities (e.g. image, text and sound) helps us make decisions and actions. Towards active vision, an AI system must be able to make use of multi-modality data. One advantage is that multi-modality data provides a comprehensive and complementary source for understanding the environment and its geometry (e.g. learn to navigate with audio like a bat [23], where the visual input and audio input both helps reconstruct the geometry of the scene). Recently, researchers have made successes in multi-modality CV models [24], [25]. However, such methods are primarily developed for parsing image semantics, not geometric

information. Based on our prior work that builds representation learning methods from single-modality inputs [5], we plan to develop geometry-aware learning methods that learn from multi-modality data, based on both multi-modality CV models and our previous efforts on self-supervised learning.

World Model: From Passive Learning to Active Learning. We humans are able to learn enormous amounts of knowledge about the world through observation and interactions in a self-supervised way. A collection of models of the world tell us what is likely and what is plausible. With the *world model*, we predict the consequences of actions. We can also reason, plan, explore, and imagine new solutions to problems. With our geometry-aware multi-modality representation learning works discussed above as the basis, we plan to develop preliminary world models that estimate the current state of the world and perform the optimal action by imagining possible action sequences and consequent state changes. The efforts will also prepare us for the long-term goal towards world-models that could reason, have common sense and general intelligence. The world model, even a preliminary version, would enable many applications ranging from medicine to robotics.

Applications to Medicine and Robotics. The world model would shift the CV and ML paradigm from passive learning to active learning, and enables many applications. For medicine, world models unleash possibilities for systems to take actions in response to the changing environment. As a more clinical application, it can be used for precision medicine, e.g. to assign optimal regimes to patients with distinct characteristics. We could also train customized treatment plan models on longitudinal patient data with treatment and consequences. Such models can also be used for simulations in offline settings to investigate high-risk treatments and identify when the state of patients' health reaches a critical point. It will help identify what treatments to avoid without a medical dead-end, where a patient will expire, regardless of all potential future treatment sequences [26].

Recently, in robotics, modern world models have shown great promise for data-efficient learning in video games and physical robots [27], [28]. Learning world models from past experience enables robots to imagine the future outcomes of potential actions, reducing the amount of trial and error in the real environment needed to learn successful behaviors. However, such systems are far from being practical, and may require heavy human intervention or repair. We anticipate improved performance and stability with our geometry-aware multi-modality self-supervised world models.

References

Part I - My Selected Publications

[2] T. Kothpalli, J. Wang, A. D. Graham, S. X. Yu, and M. C. Lin, "Machine learning approach to predicting dry-eye related signs, symptoms and diagnoses," *In Perpetration for the Journal of Ophthalmology*, 2023.

[5] J. Wang, Y. Chen, S. X. Yu, and Y. LeCunn, "Geometry-aware self-supervised learning," in *In Preparation for Proceedings of the International Conference on Computer Vision*, 2023.

[8] J. Wang, T. N. Yeh, R. Chakraborty, X. Y. Stella, and M. C. Lin, "A deep learning approach for meibomian gland atrophy evaluation in meibography images," *Translational vision science & technology*, vol. 8, no. 6, pp. 37–37, 2019.

[9] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2537–2546.

[10] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and X. Y. Stella, "Open long-tailed recognition in a dynamic world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[11] Z. Miao, K. M. Gaynor, J. Wang, *et al.*, "Insights and approaches using deep learning to classify wildlife," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.

[13] J. Wang, Y. Chen, S. X. Yu, B. Cheung, and Y. LeCun, "Compact and optimal deep learning with recurrent parameter generators," *In Preparation for Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[14] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 505–11 515.

[18] J. Wang, R. Chakraborty, and X. Y. Stella, "Spatial transformer for 3d point clouds," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, 2021.

[19] J. Wang, J. Lin, Q. Yu, R. Liu, Y. Chen, and S. X. Yu, "3d shape reconstruction from free-hand sketches," in *Proceedings of the European Conference on Computer Vision Workshop*, 2022.

[20] J. Wang, S. Jeon, S. X. Yu, X. Zhang, H. Arora, and Y. Lou, "Unsupervised scene sketch to photo synthesis," in *Proceedings of the European Conference on Computer Vision Workshop*, 2022.

[21] J. Wang, S. Li, T. N. Yeh, *et al.*, "Quantifying meibomian gland morphology using artificial intelligence," *Optometry and Vision Science*, vol. 98, no. 9, pp. 1094–1103, 2021.

[22] J. Wang, A. D. Graham, S. X. Yu, and M. C. Lin, "Predicting demographics from meibography using deep learning," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

Part II - Other Publications

[1] P. A. Keane and E. J. Topol, *With an eye to ai and autonomous diagnosis*, 2018.

[3] M. Porumb, S. Stranges, A. Pescapè, and L. Pecchia, "Precision medicine and artificial intelligence: A pilot study on deep learning for hypoglycemic events detection based on ecg," *Scientific reports*, vol. 10, no. 1, pp. 1–16, 2020.

[4] E. Huynh, A. Hosny, C. Guthier, *et al.*, "Artificial intelligence in radiation oncology," *Nature Reviews Clinical Oncology*, vol. 17, no. 12, pp. 771–781, 2020.

[6] C.-H. Yeh, X. Y. Stella, and M. C. Lin, "Meibography phenotyping and classification from unsupervised discriminative feature learning," *Translational vision science & technology*, vol. 10, no. 2, pp. 4–4, 2021.

[7] W. J. Reed, "The pareto, zipf and other power laws," *Economics letters*, vol. 74, no. 1, pp. 15–19, 2001.

[24] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[25] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[26] M. Fatemi, T. W. Killian, J. Subramanian, and M. Ghassemi, "Medical dead-ends and learning to identify high-risk states and treatments," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4856–4870, 2021.

[27] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak, "Planning to explore via self-supervised world models," in *International Conference on Machine Learning*, PMLR, 2020, pp. 8583–8592.

[28] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "Daydreamer: World models for physical robot learning," *arXiv preprint arXiv:2206.14176*, 2022.