Yoonho Lee

✉ yoonholee95@gmail.com | 🏠 yoonholee.com | 🎓 google scholar | Stanford, CA

Education

Stanford University, Ph.D. — *United States*
Department of Computer Science, Advisor: Chelsea Finn — *2021 - present*

POSTECH, M.S. — *South Korea*
Department of Computer Science and Engineering, Advisor: Seungjin Choi — *2018*

POSTECH, B.S. — *South Korea*
Department of Mathematics — *2016*

Publications

[25] Yoonho Lee, Joseph Boen, Chelsea Finn. *“Feedback Descent: Open-Ended Text Optimization via Pairwise Comparison”.* **arXiv:2511.07919** *(preprint)*

[24] Yuxiao Qu*, Anikait Singh*, Yoonho Lee*, Amrith Setlur, Ruslan Salakhutdinov, Chelsea Finn, Aviral Kumar. *“RLAD: Training LLMs to Discover Abstractions for Solving Reasoning Problems”.* **ICML 2025** Workshops: AI for Math, PRAL, ES-FoMo

[23] Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, Chelsea Finn. *“Test-Time Alignment via Hypothesis Reweighting”.* **ICML 2025** Workshop PUT

[22] Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Maximilian Du, Chelsea Finn. *“Bidirectional Decoding: Improving Action Chunking via Closed-Loop Resampling”.* **ICLR 2025** (13th International Conference on Learning Representations)

[21] Yoonho Lee, Michelle Lam, Helena Vasconcelos, Michael S. Bernstein, Chelsea Finn. *“Clarify: Improving Model Robustness With Natural Language Corrections”.* **UIST 2024** (ACM Symposium on User Interface Software and Technology), NeurIPS 2023 Workshops: XAIA, ICBINB

[20] Caroline Choi*, Yoonho Lee*, Annie S. Chen, Allan Zhou, Aditi Raghunathan, Chelsea Finn. *“AutoFT: Learning an Objective for Robust Fine-Tuning”.* Workshop on Distribution Shifts, **NeurIPS 2023**

[19] Annie S. Chen, Yoonho Lee, Amrith Setlur, Sergey Levine, Chelsea Finn. *“Confidence-Based Model Selection: When to Take Shortcuts for Subpopulation Shifts”.* Workshop on Distribution Shifts, **NeurIPS 2023**

[18] Caroline Choi*, Fahim Tajwar*, Yoonho Lee*, Huaxiu Yao, Ananya Kumar, Chelsea Finn. *“Conservative Prediction via Data-Driven Confidence Minimization”.* Transactions on Machine Learning Research (**TMLR 2024**), ICLR 2023 Workshops: TrustML, ME-FoMo

[17] Annie S. Chen*, Yoonho Lee*, Amrith Setlur, Sergey Levine, Chelsea Finn. *“Project and Probe: Sample-Efficient Domain Adaptation by Interpolating Orthogonal Features”.* **ICLR 2024, Spotlight** (top 5% of submissions), ICLR 2023 Workshops: TrustML (Oral), ME-FoMo

[16] Johnathan Wenjia Xie, Yoonho Lee, Annie S. Chen, Chelsea Finn. *“Self-Guided Masked Autoencoders for Domain-Agnostic Self-Supervised Learning”.* **ICLR 2024** (12th International Conference on Learning Representations)

[15] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, Chelsea Finn. *“DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature”.* **ICML 2023, Oral** (top 2% of submissions)

[14] Yoonho Lee*, Annie S. Chen*, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, Chelsea Finn. *“Surgical Fine-Tuning Improves Adaptation to Distribution Shifts”.* **ICLR 2023** (11th International Conference on Learning Representations)

[13] Yoonho Lee, Huaxiu Yao, Chelsea Finn. *“Diversify and Disambiguate: Out-of-Distribution Robustness via Disagreement”.* **ICLR 2023** (11th International Conference on Learning Representations)

[12] Yoonho Lee, Chelsea Finn, Stefano Ermon. *“Relaxing the Kolmogorov Structure Function for Realistic Computational Constraints”.* Workshop on Information-Theoretic Principles in Cognitive Systems, **NeurIPS 2022**

[11] Balhae Kim, Jungwon Choi, Seanie Lee, Yoonho Lee, Jung-Woo Ha, Juho Lee. *“On Divergence Measures for Bayesian Pseudocoresets”.* 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**)

[10] Huaxiu Yao*, Caroline Choi*, Bochuan Cao, Yoonho Lee, Pang Wei Koh, Chelsea Finn. *“Wild-Time: A Benchmark of in-the-Wild Distribution Shift over Time”.* 36th Conference on Neural Information Processing Systems (**NeurIPS 2022**), Datasets & Benchmarks track

[9] Yoonho Lee, Wonjae Kim, Wonpyo Park, Seungjin Choi. *“ Discrete Infomax Codes for Supervised Representation Learning”.* Special issue ”Theory and Applications of Information Processing Algorithms”, **Entropy 2022**

[8] Giung Nam*, Jongmin Yoon*, Yoonho Lee, Juho Lee. *“Diversity Matters When Learning From Ensembles”.* 35th Conference on Neural Information Processing Systems (**NeurIPS 2021**)

[7] Minkyo Seo*, Yoonho Lee*, Suha Kwak. *“On the Distribution of Penultimate Activations of Classification Networks”.* 37th Conference on Uncertainty in Artificial Intelligence (**UAI 2021**)

[6] Yoonho Lee, Juho Lee, Sung Ju Hwang, Eunho Yang, Seungjin Choi. *“Neural Complexity Measures”.* 34th Conference on Neural Information Processing Systems (**NeurIPS 2020**)

[5] Juho Lee*, Yoonho Lee*, Jungtaek Kim, Eunho Yang, Sung Ju Hwang, Yee Whye Teh. *“Bootstrapping Neural Processes”.* 34th Conference on Neural Information Processing Systems (**NeurIPS 2020**)

[4] Wonjae Kim, Yoonho Lee. *“Learning Dynamics of Attention: Human Prior for Interpretable Machine Reasoning”.* 33rd Conference on Neural Information Processing Systems (**NeurIPS 2019**)

[3] Juho Lee, Yoonho Lee, Yee Whye Teh. *“Deep Amortized Clustering ”.* Sets and Partitions Workshop at **NeurIPS 2019** (oral presentation)

[2] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, Yee Whye Teh. *“Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks”.* 36th International Conference on Machine Learning (**ICML 2019**)

[1] Yoonho Lee , Seungjin Choi. *“Gradient-based meta-learning with learned layerwise metric and subspace”.* 35th International Conference on Machine Learning (**ICML 2018**)

Fellowships and Grants

HAI Google Cloud Credits Award 2025–2026

Stanford HAI, $15,000 **OpenAI Superalignment Fellowship** 2024
OpenAI **HAI Google Cloud Credits Award** 2022–2023
Stanford HAI, $15,000 **KFAS Doctoral Fellowship** 2021–present
Korea Foundation for Advanced Studies **Korean Presidential Science Scholarship** 2012–2016
Korea Student Aid Foundation

Professional Service

Workshop organizer, NeurIPS Workshop on Distribution Shifts (2022, 2023)

Reviewer: NeurIPS (2018-2024), ICML (2019-2025), ICLR (2021-2025), NeurIPS workshop proposals (2024), ICML workshop proposals (2024), AAAI (2024-2025), AISTATS (2019-2022), IJCAI (2019-2021), ACML (2019-2020), ME-FoMo@ICLR (2023), TrustML@ICLR (2023).

Talks and Presentations

Moveworks, Mountain View, CA, USA	Dec. 2024
Centre for Frontier AI Research, Online	Aug. 2023
MosaicML, Online	Aug. 2023
ICLR 2023, Kigali, Rwanda	Apr. 2023
Deep Learning: Classics and Trends, Online	Mar. 2023
NeurIPS 2022, New Orleans, USA	Dec. 2022
ICML 2022, Baltimore, USA	Jul. 2022
NeurIPS 2021, Online	Dec. 2021
Post-NeurIPS Workshop @ KSC2020, Online	Dec. 2020
NeurIPS 2020, Online	Dec. 2020
NeurIPS 2019, Vancouver, Canada	Dec. 2019
Kakao Brain, South Korea	May. 2019, Oct. 2019
Second Korea-Japan Machine Learning Workshop, South Korea	Feb. 2019
ICML 2018, Stockholm, Sweden	Jul. 2018
Naver, South Korea	Apr. 2018

Teaching Experience

Teaching Assistant, CS330 Deep Multi-Task and Meta Learning, Stanford University	Sep. 2023 - Dec. 2023
Teaching Assistant, CS330 Deep Multi-Task and Meta Learning, Stanford University	Sep. 2022 - Dec. 2022
Teaching Assistant, Deep Learning, POSCO Group	Mar. 2017 - Jun. 2018
Teaching Assistant, Machine Learning for Business, Samsung Electronics	Sep. 2017 - Dec. 2017
Teaching Assistant, AI Job Training, POSTECH Institute of AI	Mar. 2017 - Jun. 2017
Teaching Assistant, CSED101 Programming and Problem Solving, POSTECH	Mar. 2017 - Jun. 2017