Sangdon Park

sangdon.github.io
sangdon@postech.ac.kr
github.com/sangdon
Google Scholar

Research Interests

Artificial Intelligence (AI), Trustworthy AI, and AI Alignment — My research interest focuses on designing trustworthy AI systems by understanding from theory to implementation and by considering practical applications in computer security, computer vision, natural language processing, robotics, and cyber-physical systems.

Education

University of Pennsylvania Philadelphia, USA
Ph.D. in Computer and Information Science 2021

- Advisors: Insup Lee and Osbert Bastani
- Thesis: *Uncertainty Estimation Toward Safe AI*
- Committee: Kostas Daniilidis, Nikolai Matni, Edgar Dobriban, and Kilian Q. Weinberger

Seoul National University Seoul, Korea
M.S. in Electrical and Computer Engineering 2012

- Advisor: Kyoung Mu Lee
- Thesis: *Abnormal Object Detection by Transformed-Canonical Scene Generation*

Seoul National University Seoul, Korea
B.S. in Computer Science and Engineering 2010

- Thesis Advisor: Byoung-Tak Zhang
- Thesis: *Behavioral Intelligence for Crowd Avatar in 3D Virtual Worlds*

Employment

POSTECH Pohang, Korea
Assistant Professor Aug. 2023-Now

Georgia Institute of Technology Atlanta, USA
Postdoctoral Researcher (Mentor: Taesoo Kim) Sept. 2021-July 2023

Google Cloud AI Sunnyvale, USA
Research Intern (Host: Kihyuk Sohn) Summer 2020

Biointelligence Laboratory, Seoul National University Seoul, Korea
Undergraduate Researcher 2008-2010

Republic of Korea Army Korea
Military Service 2006-2008

Publications

[1] J. Hwang, J. Park, H. Park, D. Kim, **S. Park**, and J. Ok, “Retrieval-augmented generation with estimation of source reliability”, in *Empirical Methods in Natural Language Processing (**EMNLP**)*, 2025.

[2] K. Kim, Y. Choi, H. Kim, D. Kim, and **S. Park**, “Chronobias: A benchmark for evaluating temporal group bias in retrieval-augmented language models”, in *Empirical Methods in Natural Language Processing (**EMNLP Findings**)*, 2025.

[3] S. Moon, M. Lee, **S. Park**, and D. Kim, “Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning”, in *International Conference on Computer Vision (**ICCV**)*, Oct. 2025.

[4] H. Wang, Z. Yang, **S. Park**, Y. Yang, S. Kim, W. Lunardi, M. Andreoni, T. Kim, and W. Lee, “SOUNDBOOST: Effective RCA and Attack Detection for UAV via Acoustic Side-Channel”, in *Proceedings of the 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (**DSN**)*, 2025.

[5] M. Lee, K. Kim, T. Kim, and **S. Park**, “Selective Generation for Controllable Language Models”, in *Neural Information Processing Systems (**NeurIPS**)*, 2024.

[6] S. Li, **S. Park**, I. Lee, and O. Bastani, “TRAQ: Trustworthy Retrieval Augmented Question Answering via Conformal Prediction”, in *Annual Conference of the North American Chapter of the Association for Computational Linguistics (**NAACL**)*, 2024.

[7] H. Park, J. Hwang, S. Mun, **S. Park**, and J. Ok, “MedBN: Robust Test Time Adaptation against Malicious Test Samples”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2024.

[8] W. Si, **S. Park**, I. Lee, E. Dobriban, and O. Bastani, “PAC prediction sets under label shift”, in *The Twelfth International Conference on Learning Representations (**ICLR**)*, 2024.

[9] W. Si, S. Li, **S. Park**, I. Lee, and O. Bastani, “Angelic Patches for Improving Third-Party Object Detector Performance”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)*, 2023.

[10] **S. Park**, O. Bastani, and T. Kim, “ACon^2: Adaptive Conformal Consensus for Provable Blockchain Oracles”, in *Proceedings of the 32nd USENIX Security Symposium (**Security**)*, 2023.

[11] R. Kaur, K. Sridhar, **S. Park**, Y. Yang, S. Jha, A. Roy, O. Sokolsky, and I. Lee, “CODiT: Conformal out-of-distribution detection in time-series data for cyber-physical systems”, in *Proceedings of the 14Th ACM/IEEE International Conference on Cyber-Physical Systems (**ICCPS**)*, 2023.

[12] **S. Park**, X. Cheng, and T. Kim, “Unsafe’s Betrayal: Abusing Unsafe Rust in Binary Reverse Engineering via Machine Learning”, *arXiv preprint arXiv:2211.00111*, 2023.

[13] **S. Park**, E. Dobriban, I. Lee, and O. Bastani, “PAC Prediction Sets for Meta-Learning”, in *Neural Information Processing Systems (**NeurIPS**)*, 2022.

[14] S. Li, **S. Park**, X. Ji, I. Lee, and O. Bastani, “Towards PAC multi-object detection and tracking”, *arXiv preprint arXiv:2204.07482*, 2022.

[15] S. Jang, **S. Park**, I. Lee, and O. Bastani, “Sequential covariate shift detection using classifier two-sample tests”, in *Proceedings of the 39th International Conference on Machine Learning (**ICML**)*, 2022.

[16] R. Kaur, S. Jha, A. Roy, **S. Park**, E. Dobriban, O. Sokolsky, and I. Lee, “iDECODe: In-distribution equivariance for conformal out-of-distribution detection”, in *Association for the Advancement of Artificial Intelligence (**AAAI**)*, 2021.

[17] **S. Park**, S. Li, I. Lee, and O. Bastani, “PAC confidence predictions for deep neural network classifiers”, in *International Conference on Learning Representations (**ICLR**)*, 2021.

[18] **S. Park**, O. Bastani, J. Weimer, and I. Lee, "Calibrated prediction with covariate shift via unsupervised domain adaptation", in *International Conference on Artificial Intelligence and Statistics (**AISTATS**)*, 2020.

[19] **S. Park**, O. Bastani, N. Matni, and I. Lee, "PAC confidence sets for deep neural networks via calibrated prediction", in *International Conference on Learning Representations (**ICLR**)*, 2020.

[20] **S. Park**, R. Ivanov, J. Weimer, and I. Lee, "From verification to learning for defense against adversarial examples in neural networks", *Korea Cyber-security Competition*, 2018.

[21] **S. Park**, J. Weimer, and I. Lee, "Resilient linear classification: An approach to deal with attacks on training data", in *International Conference on Cyber-Physical Systems (**ICCPS**)*, 2017.

[22] J. Oh, T. M. Howard, M. R. Walter, D. Barber, M. Zhu, **S. Park**, A. Suppe, L. Navarro-Serment, F. Duvallet, A. Boularias, *et al.*, "Integrated intelligence for human-robot teams", in *International Symposium on Experimental Robotics (**ISER**)*, 2016.

[23] **S. Park**, W. Kim, and K. M. Lee, "Abnormal object detection by canonical scene-based contextual model", in *European Conference on Computer Vision (**ECCV**)*, 2012.

Scholarships and Awards

• DARPA AIxCC Winner as Team Atlanta ($4M Team Prize)	2025
• Best Paper Finalist from CKAIA)	2025
• Bang Seung-Yang Graduate Fellowship from POSTECH CSE (awardee: Jaewoo Jeong)	2025
• POSTECH GSAI BK21 Best Paper Award	2025
• NeurIPS'24 Spotlight Paper (top 2.08%)	2024
• DARPA AIxCC Finalists as Team Atlanta ($2M Team Prize)	2024
• NeurIPS'23 Outstanding Reviewer Award	2023
• ICML'23 TEACH Workshop Best Paper Award	2023
• ICCPS'23 Best Paper Award finalist	2023
• NeurIPS'21 Outstanding Reviewer Award (top 8% of reviewers)	2021
• Korea cyber-security paper competition Best Paper Award ($4,500)	2018
• PhD fellowship at University of Pennsylvania	2013-2021
• Distinguished MS Dissertation Award at Seoul National University	2012
• Academic Performance Scholarship	2009
• National Science and Engineering Undergraduate Scholarship	2003–2008

Service

- **Area Chair**
  NeurIPS'24-25, ICML'25, ICLR'26
- **Reviewer**
  NeurIPS'21-23, ICML'21-23, ICLR'22-24, AAAI'25, Journal of the Royal Statistical Society: Series B
- **External Reviewer**
  S&P'21, S&P'22, Security'22, Security'23, Security'24, NDSS'24

Teaching

- **Instructor** at POSTECH — Fall 2025
  ML (CSED515)
- **Instructor** at POSTECH — Fall 2023, Fall 2024, Spring 2025
  Trustworthy ML (AIGS703L / CSED703L)
- **Instructor** at POSTECH — Spring 2024
  Discrete Mathematics (CSED261)
- **Teaching Assistant** at University of Pennsylvania — Spring 2015
  Machine Perception (CIS580)
- **Teaching Assistant** at University of Pennsylvania — Fall 2014
  Computer Vision and Computational Photography (CIS581)
- **Teaching Assistant** at Seoul National University — Fall 2010
  Linear Algebra for Electrical Systems
- **Instructor** at Seoul National University — Feb. 2008
  1st Free Computer Education for Gwanak-gu Community Youth

Talks

- **Introduction to Trustworthy LLMs**
  KISC (한국통신학회 단기강좌) — July. 2025
- **Trustworthy LLMs**
  Best of the Best (BoB) — July. 2025
- **Selective Generation Towards Trustworthy LLMs**
  NYU Frontier Lab — July. 2025
  KIISC AI Security Group (정보보호학회 AI보안 연구회) — July. 2025
- **Toward Trustworthy Large Language Models**
  PKNU AI — Dec. 2024
  POSTECH AI Day — Dec. 2024
- **Rethinking and Harnessing Trustworthiness of Generative AI for Security**
  Samsung Security Tech Forum (SSTF) – *Invited Talk* — Sep. 2024
- **Trustworthy Military-AI: AI Controllability**
  REAIM — Sep. 2024
- **Trend on Trustworthy Language Models**
  National Statistics Development Forum — Sep. 2024
- **Trustworthy AI: A Compositional Perspective**
  POSTECH AI/CSE — April 2024
- **Conformal Prediction for Trustworthy AI**
  Korean AI Association Winter School — Feb. 2024
- **Uncertainty Learning for Trustworthy and Secure AI**
  POSTECH AI/CSE — Mar. 2023
  KAIST EE — April 2023
  SNU CSE — April 2023
  Korea University CSE — July 2023
  SNU IPAI — July 2023
  SNU Frontier Summer School — Aug. 2023
  UNIST IB — Oct. 2023
  KAIST — Jan. 2024

CAU AI May 2024
Korea GSS June 2024

- **PAC Prediction Sets for AI Safety**
  ICML Workshop DFUQ 2022 – *Invited Talk* Jul. 2022
- **Uncertainty Quantification via PAC Prediction Sets**
  DGIST Dec. 2021
- **From Verification to Learning for Defense against Adversarial Examples in Neural Networks**
  KAIST CS Aug. 2018
  Hanyang University Aug. 2018
  KIISC Aug. 2018