Kai Yi

Actively Seeking **Research Internships** and a **Full-Time** Research Scientist Job
williamyi96@gmail.com ◇ Google Scholar ◇ kaiyi.me ◇ (+966) 54-9585759
3113-WS05, Building 12 (Library), KAUST, Thuwal, Saudi Arabia. 23955-6900

SUMMARY

I am a Ph.D. candidate in Computer Science at KAUST, supervised by Prof. Peter Richtárik, and expecting to graduate in 2025. I have interned at Sony AI, Vector Institute, Tencent AI, and SenseTime. My research primarily focuses on **Centralized/Federated LLM Compression**. As the primary author, I have co-authored over 15 papers, accumulating 320+ citations. My work is highly interconnected, featuring significant projects such as LLM post-training compression algorithm **PV-Tuning** (preprint), with more on the way; communication-efficient federated learning methods **CohortSqueeze** (preprint), **FedP3** (ICLR), and **EF-BV** (NeurIPS); and multimodal language model projects **DACZSL** (ICCVW), **HGR-Net** (ECCV), and **VisualGPT** (CVPR).

EDUCATION

King Abdullah University of Science and Technology (KAUST) Dec 2021 - Present
Ph.D. Candidate supervised by Prof. Peter Richtárik
Research Interests: LLM Compression, Federated Learning, Distributed Optimization

King Abdullah University of Science and Technology (KAUST) Sep 2020 - Dec 2021
M.S. of Vision-CAIR, supervised by Prof. Mohamed Elhoseiny
Research Interests: Zero-Shot Learning, Vision and Language
Thesis: Domain-Aware Continual Zero-Shot learning

Xi'an Jiaotong University (XJTU), Xi'an, China Aug 2015 - Jun 2019
B.S. of Software Engineering
Thesis: Accurate Object Detection and Weakly-Supervised Perception in Complex Scenes, supervised by Prof. Nanning Zheng and rated as A+ (Top 1%)

HIGHLIGHTED PUBLICATIONS

[1] **Kai Yi**, Peter Richtárik. Enhancing Distributed LLM Training: Gradient Low-Rank Projection with Error Feedback for Memory Efficiency. *under preparation*, 2024.

[2] **Kai Yi**, Timur Kharisov, Igor Sokolov, Peter Richtárik. Cohort Squeeze: Beyond a Single Communication Round per Cohort in Cross-Device Federated Learning. *arXiv*, 2024.

[3] **Kai Yi**, Nidham Gazagnadou, Peter Richtárik, Lingjuan Lv. FedP3: Federated Personalized and Privacy-friendly Network Pruning under Model Heterogeneity. **ICLR**, 2024.

[4] Exploring Hierarchical Graph Representation for Large-Scale Zero-/Few-Shot Image Classification. **Kai Yi**, Xiaoqian Shen, Yunhao Gou, Mohamed Elhoseiny. **ECCV**, 2022.

[5] Laurent Condat, **Kai Yi**, Peter Richtárik. A Unified Theory of Error Feedback and Variance Reduction Mechanisms for Controlling Biased and Unbiased Gradient Compressors in Distributed Optimization. **NeurIPS**, 2022.

[6] Jun Chen, Han Hao, **Kai Yi**, Boyang Li, Mohamed Elhoseiny. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. **CVPR**, 2022.

RESEARCH EXPERIENCE

Sony AI Jun 2023 - Sep 2023
Research Intern, supervised by Dr. Nidham Gazagnadou and Dr. Lingjuan Lyu Tokyo, Japan

- Innovated federated learning strategies for one-for-all foundation models, leading to significant advancements detailed in FedP3 (ICLR'24, [3]).

Vector Institute May 2023 - Sep 2023
Research Intern, supervised by Prof. Yaoliang Yu Remote

- Federated stochastic bilevel optimization and Newton methods for bilevel optimization. Designed efficient fully single-loop variance reduced methods based on L-SVRG for stochastic bilevel optimization ([4]).

Tencent AI Lab Dec 2020 - Apr 2021
Research Intern, mentored by Dr. Jiaxiang Wu Shenzhen, China

- Developed ML algorithms tailored for bioinformatics data, enhancing commercial products at Tencent.

Sensetime Group Limited Mar 2019 - Jun 2019
Research Intern with Dr. Wentao Liu Beijing, China

- Created accurate and fast object detection methods for commercial embedded chips at SenseTime.

Institute of Artificial Intelligence and Robotics Jul 2017 - Feb 2019
Research and Engineering Intern with Prof. Nanning Zheng Xi'an, China

- Developed cognition-based small object detection methods for autonomous driving, leading to publications at AIAI'18 ([17]), GlobalSIP'18 ([16]), and preprint ([15]), and enhancing the Pioneer I autonomous vehicle.

OTHER PUBLICATIONS

[1] Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Pavlovich Burlachenko, **Kai Yi**, Dan Alistarh, Peter Richtárik. PV-Tuning: Beyond Straight-Through Estimation for Extreme LLM Compression. *arXiv*, 2024.

[2] Georg Meinhardt, **Kai Yi**, Laurent Condat, Peter Richtárik. Prune at the Clients, Not the Server: Accelerated Sparse Training in Federated Learning. *arXiv*, 2024.

[3] **Kai Yi**, Georg Meinhardt, Laurent Condat, Peter Richtárik. FedComLoc: Communication-Efficient Distributed Training of Sparse and Quantized Models. *arXiv*, 2024.

[4] **Kai Yi**, Yaoliang Yu. Efficient Fully Single-Loop Variance Reduced Methods for Stochastic Bilevel Optimization. *under review*, 2023.

[5] Wenxuan Zhang, Paul Janson, **Kai Yi**, Ivan Skorokhodov, Mohamed Elhoseiny. Continual Zero-Shot Learning through Semantically Guided Generative Random Walks. **ICCV**, 2023.

[6] **Kai Yi**, Paul Janson, Wenxuan Zhang, Mohamed Elhoseiny. Domain-Aware Continual Zero-Shot Learning. **ICCV** *OOD-CV Workshop*, 2023.

[7] **Kai Yi**, Laurent Condat, Peter Richtárik. Explicit Personalization and Local Training: Double Communication Acceleration in Federated Learning. *arXiv*, 2023.

[8] Grigory Malinovsky, **Kai Yi**, Peter Richtárik. Variance Reduced ProxSkip: Algorithm, Theory and Application to Federated Learning. **NeurIPS**, 2022.

[9] **Kai Yi**, Divyansh Jha, Ivan Skorokhodov, Mohamed Elhoseiny. Language-Guided Imaginative Walks: Generative Random Walk Deviation Loss for Unseen Class Recognition using Text Descriptions. **CVPR** *L3D-IVU Workshop*, 2022.

[10] Divyansh Jha, **Kai Yi**, Ivan Skorokhodov, Mohamed Elhoseiny. Creative Walk Adversarial Networks: Novel Art Generation with Probabilistic Random Walk Deviation from Style Norms. **ICCC**, 2022.

[11] **Kai Yi**, Yungeng Zhang, Jianye Pang, Xiangrui Zeng, Min Xu. Learning To Disentangle Semantic Features From cryo-ET with 3D Spatial Generative Network. *Technical Report*, 2021.

[12] Yuchen Zeng, Xiangrui Zeng, **Kai Yi**, Jie Jin, Jing Zhang, Yi-Wei Chang, Yang Ge, Min Xu. Unsupervised Domain Alignment based Open Set Structural Recognition of Macromolecules Captured by Cryo-Electron Tomography. **ICIP**, 2021.

[13] Mohamed Elhoseiny*, **Kai Yi***, Mohamed Elfeki*. CIZSL++: Creativity Inspired Generative Zero-Shot Learning. *T-PAMI under review*, arXiv.

[14] Jianye Pang, **Kai Yi**, Wanguang Yin, Min Xu. Experimental Analysis of Legendre Decomposition in Machine Learning. *Technical Report*, 2020.

[15] **Kai Yi**, Zhiqiang Jian, Shitao Chen, Nanning Zheng. Feature Selective Small Object Detection via Knowledge-based Recurrent Attentive Network. *Technical Report*, 2019.

[16] Tiannan Dong, Jianji Wang, Meng Yang, **Kai Yi**, Nanning Zheng. Affine LBG for Codebook Training of Univariate Linear Representation. *IEEE Global Conference on Signal and Information Processing (**GlobalSIP**)*, 2018.

[17] **Kai Yi**, Shitao Chen, Yu Chen, Chao Xia, Nanning Zheng. Cognition-based Deep Learning: Progresses and Perspectives. *Artificial Intelligence Applications and Innovations (**AIAI**)*, 2018 (Oral).

TEACHING & SERVICES

Conference Reviewer:
NeurIPS'22-24, ICLR'23-25, ICML'22-24, AISTATS'23, CVPR'22-24, ICCV'23
ECCV'22,24, AAAI'22-24, WACV'21-25, BMVC'20-23, ITSC'20-21, IV'18-21

Journal Reviewer:
T-PAMI, IJCV, CVIU, T-IP, T-SP, T-NNLS

Teaching Assistant:
CS283: Deep Generative Modeling (KAUST)
Introduction to Machine Learning, Computer Architecture (XJTU)

TALKS

- Invited talk at SonyAI presenting our federated pruning project. 2023.09.29
- Invited talk at SonyAI-PPML talking about Accelerated LT Methods in FL. 2023.08.23
- Invited talk at Vector Institute Demo Day talking "Optimal and Efficient Variance Reduced Methods for Stochastic Bilevel Optimization" 2023.08.17
- Invited presenter at KAUST VCC Open House 2023 talking ProxSkip-VR. 2023.03.02
- Spotlight talk of EF-BV at KAUST Rising Stars in AI Symposium 2023. 2023.02.21
- Representing our group to present ProxSkip-VR at KAUST VCC Showcase Event. 2023.01.29
- Invited speaker at ECCV2022-AI TIME talking about our HGR-Net. 2022.12.07
- Spotlight talk of CIZSL++ at KAUST Conference on Artificial Intelligence. 2021.04.28

AWARDS & HONORS

- KAUST Graduate Scholarship 2020-
- Outstanding Graduates of XJTU (top 5%) 2019
- Zeng Xianzi Scholarship (37/4100, top 0.9%) 2016-2018
- Candidate of 6th Excellent Student Model of XJTU (3/37) 2018
- Outstanding Leader of the Students' Union (top 2%) 2016
- Excellent Student Award (top 5%) of XJTU 2016-2018

ACTIVITIES

- KAUST Orientation Leader 2022 Fall
- KAUST CEMSE Student Ambassador Sep 2021 - Now
- Member of SIAM, IEEE, CVF
- Volunteer of ICML 2021; NeurIPS 2020, 2021.

ADDITIONAL INFORMATION

Skills: Proficient in Python, Pytorch, and Android Developments, Master JAX, TensorFlow, C++
Hobbies: Fond of long-distrance running and hiking