# Samir Khaki

samir.khaki@mail.utoronto.ca | www.samirkhaki.com | Skhaki18 | SamirKhaki | Samir Khaki

## Education

**University of Toronto** *Ontario, Canada*

B.A.Sc. Electrical & Computer Engineering | Major GPA: 3.8 *Sep. 2019 - May 2024*

- Minors in: Artificial Intelligence (Machine Learning Emphasis) & Robotics (Vision Emphasis)
- Honors thesis/Research Supervisor: Dr. Kostas Plataniotis
- Research Supervisor (s): Dr. Steve Mann Dr. Mahdi Hosseini
- Research Project Advisor: Dr. Jimmy Ba

## Professional Experience (See Industry CV on my Website)

### *Academic Experience*

| | |
|---|---|
| 2023-Present | **Machine Learning + Vision Researcher**, *Song Han*, MIT Han Lab, Massachusetts Institute of Technology (MIT) |
| 2021-Present | **Machine Learning Researcher**, *Kostas Plataniotis & Mahdi Hosseini*, Multimedia Lab, University of Toronto (UofT) |
| 2021-Present | **ML Robotics Researcher**, *Steve Mann*, MannLab Canada |

### *Industry Experience*

| | |
|---|---|
| 2024-2024 | **Machine Learning Researcher**, Google Research |
| 2022-Present | **Machine Learning Researcher**, IBM |
| 2023-2023 | **Software Engineering**, Amazon AWS |

## Recent Publications (Full List on Google Scholar)

### Highlighted Publications (Non-Exhaustive) ** Equal Contribution First Author*

A. Sajedi *, **Samir Khaki ***, E. Amjadian, L. Liu, Y. Lawryshyn, K. Plataniotis. 2023. Efficient Dataset Distillation with Attention.
IEEE/CVF International Conference on Computer Vision (ICCV) and United States Patent Pending – Proceedings Link

**Samir Khaki**, W. Luo. 2023. CFDP: Common Frequency Domain Pruning.
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop – Proceedings Link

Y. Wang *, **Samir Khaki ***, W. Zheng *, M. Hosseini, K. Plataniotis. 2023. CONetV2: Efficient Auto-Channel Size Optimization
IEEE International Conference on Machine Learning and Applications (ICMLA) – Proceedings Link

A. Sajedi, **Samir Khaki**, Y. Lawryshyn, K. Plataniotis. 2024. Probabilistic Contrastive Learning for Multi-Label Classification
IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

### Under Review

**Samir Khaki**, K. Plataniotis. 2024. The Need for Speed: Pruning Transformers with One Recipe
International Conference on Learning Representations (ICLR) – Average Score: **6.0** – OpenReview Link

A. Sajedi *, **Samir Khaki ***, K. Plataniotis, M. Hosseini. 2024. End-to-End Supervised Multilabel Contrastive Learning
IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) – Arxiv 2023 Link

A. Sajedi *, **Samir Khaki ***, L. Liu, Y. Lawryshyn, K. Plataniotis. 2024. Data-2-Model Distillation: Why not use more knowledge?
IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)

## Invited Talks

Fall 2023. *Efficient Dataset Distillation with Attention*. Poster Presentation: International Conference on Computer Vision
**Paris, France**

Spring 2023. *Insider look at model compression*. Oral Presentation: Royal Bank of Canada (RBC) Borealis AI Group.
**Ontario, Canada**

Spring 2023. *Efficeint Computer Vision*. Poster Presentation: Computer Vision and Pattern Recognition Conference (CVPR).
**British Columbia, Canada**

Fall 2024. *Hosting the $1^{st}$ Workshop on Efficient Dataset Distillation at ECCV 2024*. Collaborating with the National University of Singapore (NUS), Massachusetts Institute of Technology (MIT), Princeton University, and University of Toronto (UofT)
**Milano, Italy**

## Teaching Experience

| | |
|---|---|
| 2022-2024 | **ECE516: Intelligent Image Processing – *Lead Teaching Assistant* – (Master's Course)**, |
| 2022-2024 | **ECE1724: Adv. Intelligent Image Processing – *Lead Teaching Assistant* – (Ph.D Course)**, |

## Efficient Deep Learning Projects (Model-Centric)

**Research, ICLR 2024, The Need for Speed: Pruning Transformers with One Recipe**,

$1^{st}$ Author
2023-2024

- Earned an average score of **6**; First Author and only student author.
- Developed a one-shot compression pipeline for transformers across multiple modalities and tasks: language processing, language generation, vision classification, transfer learning, and semantic segmentation.
- Leveraged intermediate feature error to compress transformers dynamically across several architectures including BERT, ViTs, DeiTs, Mask2Former, and GPT2.
- Notably achieved a **≤2%** degradation from natural language baselines, and generalizable performance on semantic segmentation with a **24%** reduction in FLOPS and a **13%** reduction in latency on Mask2Former.

*University of Toronto*

**Research, CVPR-ECV 2023, CFDP: Common Frequency Domain Pruning**,

$1^{st}$ Author
2023-2023

- Published independent of graduate students or direct faculty supervisors.
- Under the guidance of Prof. Jimmy Ba, this paper addresses the computational bottlenecks in deep cnns architectures using Fourier analysis techniques.
- Computes the magnitude of the discrete cosine transforms over the intermediate features, regularized by its frequency distribution to quantify the weight importance.
- Achieved a **+0.2%** performance improvement in GoogleNet on CIFAR-10 with a **54%** reduction in parameters while achieving baseline performance with ResNet-50 on ImageNet-1K at a **40%** parameter reduction.

*University of Toronto*

**Design Project, L'Oreal ModiFace, Model Compression for Real-Time Webcam Detection**,

Team Lead
2023-2024

- Leading a model compression project on maximizing inference speed (frame-per-second) of a facial localization model for a landmark regression task.
- Implementing weight-norm structured pruning on the feature encoder led to device-agnostic acceleration with a **46%** memory reduction and **27%** latency improvement, further optimized by dynamic quantization for a net **1.8×** speedup.
- To be presented at the University of Toronto Research Conference in April 2024.

*University of Toronto*

**Research, MIT HAN Lab, Efficient High-Resolution Segmentation**,

Member
2023-2024

- Accelerating high-resolution semantic segmentation by refining low-resolution predictions using a sparse high-resolution inference.
- Surpassed High-Resolution baseline on Pascal VOC segmentation by **+0.4%** using refined low-resolution predictions, improving net latency by **2×**, for HRNet-48W.

*Massachusetts Institute of Technology*

## Efficient Deep Learning Projects (Data-Centric & Training Paradigms)

**Research, ICCV 2023, DataDAM: Efficient Dataset Distillation with Attention Matching**,

Equal $1^{st}$ Author
2023-2023

- A **patent** is filed with Royal Bank of Canada and the United States Patent Office.
- Leveraging randomly sampled networks, DataDAM circumvents the costs of bi-level optimization while harnessing intermediate feature attention to capture knowledge with a distance measure in the kernel representation embedding.
- The distilled data achieved a **7%** improvement in accuracy with a **5**× acceleration and a **3**× reduction in GPU memory consumption.
- Developed derivative work incorporating generative models for dataset distillation with constant space complexity, submitted for review at CVPR 2024.

*University of Toronto*

**Research, CVPR 2024, End-to-End Supervised Multilabel Contrastive Learning**,

Equal $1^{st}$ Author
2022-2024

- This work focuses on improving the efficiency of training by incorporating contrastive loss dynamics enabling smaller architecture, such as TResNet-M, to outperform their larger counterparts, TResNet-L.
- Leverages a custom loss function to conjoin contrastive kernel representations and multi-modal feature distributions to capture epistemic and aleatoric uncertainties enabling a faster training convergence profile.
- Outperforms competitive methods on the Microsoft-COCO and Pascal-VOC datasets with a **35%** reduction in computational complexity.

*University of Toronto*