

# Stephan Randle

US    [stephanrandle@hotmail.com](mailto:stephanrandle@hotmail.com)    (248) 825-7866    [stephansama.info](http://stephansama.info)    [stephansama](https://github.com/stephansama)  
[stephan-randle-38a30319a](https://www.linkedin.com/in/stephan-randle-38a30319a)    [stephansama](https://www.youtube.com/stephansama)    [stephansama](https://github.com/stephansama)

## Welcome to RenderCV

---

RenderCV reads a CV written in a YAML file, and generates a PDF with professional typography.

See the [documentation](#) for more details.

## Education

---

### PhD **Princeton University**, Computer Science

- Thesis: Efficient Neural Architecture Search for Resource-Constrained Deployment
- Advisor: Prof. Sanjeev Arora
- NSF Graduate Research Fellowship, Siebel Scholar (Class of 2022)

Princeton, NJ

Sept 2018 – May 2023

### BS **Boğaziçi University**, Computer Engineering

- GPA: 3.97/4.00, Valedictorian
- Fulbright Scholarship recipient for graduate studies

Istanbul, Türkiye

Sept 2014 – June 2018

## Experience

---

### **Nexus AI**, Co-Founder & CTO

- Built foundation model infrastructure serving 2M+ monthly API requests with 99.97% uptime
- Raised \$18M Series A led by Sequoia Capital, with participation from a16z and Founders Fund
- Scaled engineering team from 3 to 28 across ML research, platform, and applied AI divisions
- Developed proprietary inference optimization reducing latency by 73% compared to baseline

San Francisco, CA

June 2023 – present

2 years 7 months

### **NVIDIA Research**, Research Intern

- Designed sparse attention mechanism reducing transformer memory footprint by 4.2x
- Co-authored paper accepted at NeurIPS 2022 (spotlight presentation, top 5% of submissions)

Santa Clara, CA

May 2022 – Aug 2022

4 months

### **Google DeepMind**, Research Intern

- Developed reinforcement learning algorithms for multi-agent coordination
- Published research at top-tier venues with significant academic impact
  - ICML 2022 main conference paper, cited 340+ times within two years
  - NeurIPS 2022 workshop paper on emergent communication protocols
  - Invited journal extension in JMLR (2023)

London, UK

May 2021 – Aug 2021

4 months

### **Apple ML Research**, Research Intern

- Created on-device neural network compression pipeline deployed across 50M+ devices
- Filed 2 patents on efficient model quantization techniques for edge inference

Cupertino, CA

May 2020 – Aug 2020

4 months

### **Microsoft Research**, Research Intern

- Implemented novel self-supervised learning framework for low-resource language modeling
- Research integrated into Azure Cognitive Services, reducing training data requirements by 60%

Redmond, WA

May 2019 – Aug 2019

4 months

## Projects

---

### FlashInfer

Jan 2023 – present

Open-source library for high-performance LLM inference kernels

- Achieved 2.8x speedup over baseline attention implementations on A100 GPUs
- Adopted by 3 major AI labs, 8,500+ GitHub stars, 200+ contributors

### NeuralPrune

Jan 2021

Automated neural network pruning toolkit with differentiable masks

- Reduced model size by 90% with less than 1% accuracy degradation on ImageNet
- Featured in PyTorch ecosystem tools, 4,200+ GitHub stars

## Publications

---

### Sparse Mixture-of-Experts at Scale: Efficient Routing for Trillion-Parameter Models

July 2023

John Doe, Sarah Williams, David Park

[10.1234/neurips.2023.1234](https://doi.org/10.1234/neurips.2023.1234) (NeurIPS 2023)

### Neural Architecture Search via Differentiable Pruning

Dec 2022

James Liu, John Doe

[10.1234/neurips.2022.5678](https://doi.org/10.1234/neurips.2022.5678) (NeurIPS 2022, Spotlight)

### Multi-Agent Reinforcement Learning with Emergent Communication

July 2022

Maria Garcia, John Doe, Tom Anderson

[10.1234/icml.2022.9012](https://doi.org/10.1234/icml.2022.9012) (ICML 2022)

### On-Device Model Compression via Learned Quantization

May 2021

John Doe, Kevin Wu

[10.1234/iclr.2021.3456](https://doi.org/10.1234/iclr.2021.3456) (ICLR 2021, Best Paper Award)

## Selected Honors

---

- MIT Technology Review 35 Under 35 Innovators (2024)
- Forbes 30 Under 30 in Enterprise Technology (2024)
- ACM Doctoral Dissertation Award Honorable Mention (2023)
- Google PhD Fellowship in Machine Learning (2020 – 2023)
- Fulbright Scholarship for Graduate Studies (2018)

## Skills

---

**Languages:** Python, C++, CUDA, Rust, Julia

**ML Frameworks:** PyTorch, JAX, TensorFlow, Triton, ONNX

**Infrastructure:** Kubernetes, Ray, distributed training, AWS, GCP

**Research Areas:** Neural architecture search, model compression, efficient inference, multi-agent RL

## Patents

---

1. Adaptive Quantization for Neural Network Inference on Edge Devices (US Patent 11,234,567)
2. Dynamic Sparsity Patterns for Efficient Transformer Attention (US Patent 11,345,678)
3. Hardware-Aware Neural Architecture Search Method (US Patent 11,456,789)

## Invited Talks

---

4. Scaling Laws for Efficient Inference — Stanford HAI Symposium (2024)
3. Building AI Infrastructure for the Next Decade — TechCrunch Disrupt (2024)
2. From Research to Production: Lessons in ML Systems — NeurIPS Workshop (2023)
1. Efficient Deep Learning: A Practitioner's Perspective — Google Tech Talk (2022)

## **Any Section Title**

---

You can use any section title you want.

You can choose any entry type for the section: `TextEntry`, `ExperienceEntry`, `EducationEntry`, `PublicationEntry`, `BulletEntry`, `NumberedEntry`, or `ReversedNumberedEntry`.

Markdown syntax is supported everywhere.

The `design` field in YAML gives you control over almost any aspect of your CV design.

See the [documentation](#) for more details.