

# Stephan Randle

📍 US    ✉️ [stephanrandle@hotmail.com](mailto:stephanrandle@hotmail.com)    ☎️ (248) 825-7866    🔗 [stephansama.info](http://stephansama.info)    🌐 [stephansama](https://stephansama.com)  
[in stephan-randle-38a30319a](https://www.linkedin.com/in/stephan-randle-38a30319a)    📺 [stephansama](https://www.youtube.com/channel/UCv33333333333333333333)    </> [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

- |  |  |
|--|--|
| <b>PhD</b> <b>Princeton University</b> , Computer Science <ul style="list-style-type: none"> <li>• Thesis: Efficient Neural Architecture Search for Resource-Constrained Deployment</li> <li>• Advisor: Prof. Sanjeev Arora</li> <li>• NSF Graduate Research Fellowship, Siebel Scholar (Class of 2022)</li> </ul> | Princeton, NJ<br>Sept 2018 – May 2023      |
| <b>BS</b> <b>Boğaziçi University</b> , Computer Engineering <ul style="list-style-type: none"> <li>• GPA: 3.97/4.00, Valedictorian</li> <li>• Fulbright Scholarship recipient for graduate studies</li> </ul>  | Istanbul, Türkiye<br>Sept 2014 – June 2018 |

## Experience

- |  |  |
|--|--|
| <b>Nexus AI</b> , Co-Founder & CTO <ul style="list-style-type: none"> <li>• Built foundation model infrastructure serving 2M+ monthly API requests with 99.97% uptime</li> <li>• Raised \$18M Series A led by Sequoia Capital, with participation from a16z and Founders Fund</li> <li>• Scaled engineering team from 3 to 28 across ML research, platform, and applied AI divisions</li> <li>• Developed proprietary inference optimization reducing latency by 73% compared to baseline</li> </ul>                       | San Francisco, CA<br>June 2023 – present<br>2 years 7 months |
| <b>NVIDIA Research</b> , Research Intern <ul style="list-style-type: none"> <li>• Designed sparse attention mechanism reducing transformer memory footprint by 4.2x</li> <li>• Co-authored paper accepted at NeurIPS 2022 (spotlight presentation, top 5% of submissions)</li> </ul>   | Santa Clara, CA<br>May 2022 – Aug 2022<br>4 months           |
| <b>Google DeepMind</b> , Research Intern <ul style="list-style-type: none"> <li>• Developed reinforcement learning algorithms for multi-agent coordination</li> <li>• Published research at top-tier venues with significant academic impact             <ul style="list-style-type: none"> <li>• ICML 2022 main conference paper, cited 340+ times within two years</li> <li>• NeurIPS 2022 workshop paper on emergent communication protocols</li> <li>• Invited journal extension in JMLR (2023)</li> </ul> </li> </ul> | London, UK<br>May 2021 – Aug 2021<br>4 months                |
| <b>Apple ML Research</b> , Research Intern <ul style="list-style-type: none"> <li>• Created on-device neural network compression pipeline deployed across 50M+ devices</li> <li>• Filed 2 patents on efficient model quantization techniques for edge inference</li> </ul>   | Cupertino, CA<br>May 2020 – Aug 2020<br>4 months             |
| <b>Microsoft Research</b> , Research Intern <ul style="list-style-type: none"> <li>• Implemented novel self-supervised learning framework for low-resource language modeling</li> <li>• Research integrated into Azure Cognitive Services, reducing training data requirements by 60%</li> </ul>   | Redmond, WA<br>May 2019 – Aug 2019<br>4 months               |

## Projects

---

<b>FlashInfer</b> Open-source library for high-performance LLM inference kernels <ul style="list-style-type: none"><li>Achieved 2.8x speedup over baseline attention implementations on A100 GPUs</li><li>Adopted by 3 major AI labs, 8,500+ GitHub stars, 200+ contributors</li></ul>	Jan 2023 – present
<b>NeuralPrune</b> Automated neural network pruning toolkit with differentiable masks <ul style="list-style-type: none"><li>Reduced model size by 90% with less than 1% accuracy degradation on ImageNet</li><li>Featured in PyTorch ecosystem tools, 4,200+ GitHub stars</li></ul>	Jan 2021

## Publications

---

<b>Sparse Mixture-of-Experts at Scale: Efficient Routing for Trillion-Parameter Models</b> <i>John Doe, Sarah Williams, David Park</i> <a href="#">10.1234/neurips.2023.1234</a> (NeurIPS 2023)	July 2023
<b>Neural Architecture Search via Differentiable Pruning</b> <i>James Liu, John Doe</i> <a href="#">10.1234/neurips.2022.5678</a> (NeurIPS 2022, Spotlight)	Dec 2022
<b>Multi-Agent Reinforcement Learning with Emergent Communication</b> <i>Maria Garcia, John Doe, Tom Anderson</i> <a href="#">10.1234/icml.2022.9012</a> (ICML 2022)	July 2022
<b>On-Device Model Compression via Learned Quantization</b> <i>John Doe, Kevin Wu</i> <a href="#">10.1234/iclr.2021.3456</a> (ICLR 2021, Best Paper Award)	May 2021

## 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

---

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

## Invited Talks

---

- Scaling Laws for Efficient Inference — Stanford HAI Symposium (2024)
- Building AI Infrastructure for the Next Decade — TechCrunch Disrupt (2024)
- From Research to Production: Lessons in ML Systems — NeurIPS Workshop (2023)
- 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.