Shen Zhuoran

cmsflash99@gmail.com | +1 425-428-3693 | cmsflash.github.io | github.com/cmsflash

Work Experience

Augment, San Francisco Bay Area, United States *Dec. 2023 – Present*
Research Scientist, Research

- Lead Augment's code LLM pre-training, matched state-of-the-art performance (DeepSeek-Coder).
- Post-trained several component models of Augment's state-of-the-art enterprise coding agent. An open-source version reached no. 1 on SWE-bench Verified.

Cruise, San Francisco Bay Area, United States *Jan. 2023 – Dec. 2023*
Senior ML/Robotics Engineer, Behaviors Data, AI

- Established a continuous training mechanism for Cruise's planning models.
- Lead an ML-based solution for misbehaviors around emergency vehicles.

Pony.ai, San Francisco Bay Area, United States *Nov. 2021 – Oct. 2022*
Software Engineer, Prediction Department

- Lead the motion prediction module's transition from heuristics to end-to-end deep learning.

Google, Seattle, WA, United States *Oct. 2019 – Aug. 2021*
AI Resident, Google Brain, Google Research

- Designed global self-attention networks (GSA-Nets), an early Transformer architecture for computer vision, with superior accuracy-latency trade-off vs. CNNs.
- Worked on vision Transformer for open-world localization (OWL-ViT), a state-of-the-art zero/few-shot detection framework that transfers from image-text pretraining. Published a paper at ECCV 2022.

SenseTime, Hong Kong *Jun. 2017 – Jun. 2019*
Research Intern, Intelligent Perception and Services Team, Smart City Group

- Proposed one of the first linear attention mechanisms, demonstrating superior performance on many image, video, and stereo vision tasks. First-author arXiv entry in Dec. 2018, published at WACV 2021.

Education

The University of Hong Kong, Hong Kong *Sep. 2015 – Jun. 2019*
BEng Computer Science; GPA: 3.85/4.30, Standing: 1/111.

Awards

- **First Runner-up**, *ACM-HK Programming Contest 2017*

Publications and Preprint

- M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, **Shen Z.**, X. Wang, X. Zhai, T. Kipf, N. Houlsby. (2022). *Simple Open-Vocabulary Object Detection with Vision Transformers.* ECCV 2022.
- **Shen Z.**, Zhang M., Zhao H., Yi S., Li H. (2021). *Efficient Attention: Attention with Linear Complexities*. WACV 2021.
- **Shen Z.**, I. Bello, R. Vemulapalli, Jia X., Chen C.-H. (2020). *Global Self-Attention Networks for Image Recognition*. arXiv: 2010.03019.
- Li Y.*, **Shen Z.***, Shan Y. (2020). *Fast Video Object Segmentation using the Global Context Module*. ECCV 2020. **Equal contribution.*

Skills

- **Languages**: Python, TypeScript, JavaScript, SQL, C++, Shell script, Markdown, LaTeX
- **Technologies**: PyTorch, TensorFlow, Keras, NumPy, Horovod, Slurm, Git, Bazel, Django
- **Skills**: Machine learning, large language models (LLMs), code LLMs, coding agents, post-training, computer vision, self-driving, motion prediction