Shen Zhuoran

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

Work Experience

Cruise, San Francisco Bay Area, United States *Jan. 2023 – Present*
Senior Machine Learning Engineer, Behaviors Data, AI

- Deploying the first large language model (LLM) on premise for data-driven enhancement of onboard models.
- Working on data-driven machine learning transition of the planning stack.

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

- Developed the next-generation, end-to-end, general-purpose trajectory prediction model for self-driving.

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

- Designed global self-attention networks (GSA-Nets), a novel meta-architecture for computer vision that uses efficient attention mechanisms to fully replace convolution with superior accuracy-cost trade-offs.
- Worked on vision Transformer for open-world localization (OWL-ViT), a simple zero/few-shot detection framework that transfers from image-text pretraining. Set a new state-of-the-art for one-shot detection by a wide margin. To publish a paper at ECCV 2022.
- Developed an on-device age detector using cross-domain knowledge distillation. Deployed the model to user devices to support privacy-preserving data filtering for a confidential project.

Tencent, Shenzhen, China *Jul. 2019 – Sep. 2019*
Research Intern, Applied Research Center, Platform and Content Group

- Proposed a novel efficient attentive memory mechanism for an arbitrarily long video with constant complexity w.r.t. video length. Presented a first-author paper at ECCV 2020.

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

- Proposed a novel efficient attention mechanism with linear complexities. Significantly improved performance-cost trade-offs on many tasks including object detection, instance segmentation, stereo depth estimation, and temporal action localization. Presented a first-author paper 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, C++, Shell script, Markdown, LaTeX
- **Technologies**: TensorFlow, Keras, PyTorch, NumPy, OpenCV, Horovod, Slurm, Git, Bazel, Django
- **Skills**: Deep learning, machine learning, neural networks, computer vision, motion prediction, self-driving