Zhizhen Pan

+86-15111023061 | panzhizhen@bupt.edu.cn | https://ddsacu.github.io/ | Beijing

EDUCATION

Beijing University of Posts and Telecommunications September 2023 – Present

B.Eng. in Electronic Information Engineering *GPA of 3.79/4.0 (Ranking of 3/101)*

- **Core Courses**: Signal and System (96), Telecommunication Principal (96), Digital Signal Processing (94), Introduction of Artificial Intelligence (97), Machine learning (91), Digital Circuit Design (94), Computer Network (96), Data Design (92), Advanced Mathematics (90), Linear Algebra (95), Probability Theory and Stochastic Processes (98)

PUBLICATION

QVGGT: Post-Training Quantized Visual Geometry Grounded Transformer

Zhizhen Pan, Hesong Wang, Huan Wang†

The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026

PROJECTS

BlockVid-2 kv cache compression March 2026 - Present

- Participated in the long-video generation inference optimization project, designed and evaluated KV cache retrieval and compression strategies to reduce computational overhead in long-sequence generation. Explored context reuse schemes with different page/span granularities, and verified their effectiveness via latency and generation quality metrics.
- Advised by Prof. Bohan Zhuang.

Post-Training Quantized Visual Geometry Grounded Transformer Augest 2025 - Present

- We propose QVGGT, a post-training quantization framework, to address the deployment bottleneck of VGGT, a SOTA model for single-pass 3D attribute estimation. We analyse the heterogeneous quantization sensitivity of VGGT and perform selective mixed-precision quantization, camera information compensation and task-aware scale searching mechanism.
- Achieved near-lossless W4A16 quantization across multiple 3D perception benchmarks, reducing VRAM usage by 3.0× 4.9× and boosting speed by 2.8× on RTX 4090 over the FP32 baseline.
- Advised by Prof. Huan wang. First author accepted by CVPR '26.

Acceleration and Operator Optimization of LLM Based on the Unsloth Framework April 2025 - Juliy 2025

- To tackle the VRAM bandwidth bottleneck in LLM fine-tuning and inference, I researched the acceleration mechanism, optimizing low-level operators to boost LLM training and inference efficiency based on *Unsloth* framework.
- Advised by Prof. Huan wang.

Intelligent Pet Feeding System Based on Facial Recognition September 2024 - April 2025

- **Core Responsibilities**: Developed a pet facial recognition model with over 92% accuracy by fine-tuning *YOLOv5s* model; deployed the model on Raspberry Pi 5, and constructed a hardware system integrating infrared sensors and servo control modules to achieve multi-pet identity recognition and automatic feeding.
- **Contributions**: Collected raw data and built a dedicated *dataset*, then pre-trained the model to implement reliable facial recognition functionality; Achieved YOLOv5s model light-weighting (30% reduction in GFLOP) through model pruning and module replacement techniques.
- Advised by Prof. Sihai Wang.

EXPERIENCE

Westlake University ENCODE Lab April 2025 – Present

Visiting student advised by Prof. Huan Wang Hangzhou, Zhejiang

Zhejiang University ZIP Lab March 2026 – Present

Research assistant advised by Prof. Bohan Zhuang Hangzhou, Zhejiang

Last updated in June 2026.

SKILLS

Programming : C/C++, Python, Java, HTML, CSS, JavaScript, Matlab, CUDA/Triton (currently learning).
Languages : English (CET-6 574), Mandarin (Native).
Other Skills: LaTeX, Linux, Git, Shell, STM32

AWARDS

Second Prize Scholarship of Beijing University of Posts and Telecommunications. 2024, 2025
Third Prize in the Chinese Mathematics Competitions (CMC). 2024
Second Prize in the China International College Students' Innovation Competition (Beijing Division). 2024

Last updated in June 2026.