裴旻楠

主页： https://happypmn.github.io

邮箱： peiminnan19@mails.ucas.ac.cn

电话： +86-152-2930-7067

教育经历

2023 至今	中国科学院大学 · 自动化研究所（直博） · GPA：3.89/4.0 导师：程健、李钢；曾任人工智能学院学生会主席（2024）	北京
2019–2023	中国科学院大学 · 电子信息工程（本科） · GPA：3.93/4.0	北京
	2022 春季公派访学：UC Berkeley · EECS · GPA：4.0/4.0	伯克利
	北京市三好学生（2021）；国家奖学金（2022）；北京市优秀毕业生（2023）	

论文发表

GCC: A 3DGS Inference Architecture with Gaussian-Wise and Cross-Stage Conditional Processing.
Minnan Pei, Gang Li, Junwen Si, Zeyu Zhu, ……, Xiaoyao Liang, Jian Cheng — MICRO 2025

MemeBQ: Memory Efficient Binary Quantization of LLMs.
Yuanhui Wang, Kunlong Liu, **Minnan Pei**, Zhangming Li, Peisong Wang, Qinghao Hu — AAAI 2026

APEX: Integer-only Non-linear Function Approximation for Efficient Cross-Modal Inference.
Peihuan Ni, Zitao Mo, Tielong Liu, Hongli Wen, Zeyu Zhu, **Minnan Pei**, ……, Jian Cheng — DATE 2026

Boosting the Performance of Tree-Based Speculative Decoding of LLMs on FPGAs.
Tielong Liu, Gang Li, Zitao Mo, Zeyu Zhu, **Minnan Pei**, Jian Cheng — DATE 2026

项目经历

GCC: 3D Gaussian Splatting 推理加速体系结构 · 手机端渲染系统加速 2024–

从 GPU 处理 3DGS 渲染的传统低效模式出发，提出高斯级数据流与跨阶段条件跳过机制，减少冗余计算与访存压力；在体系结构层实现流水化算子调度和可重构片上缓存策略，实现推理速度与能效的整体提升。进一步面向移动端特性，将推理数据流与算子划分嵌入 Vulkan 渲染框架，在华为手机上实现 3DGS 实时渲染加速，显著降低功耗并保持画面质量稳定。

SpatialViz-Bench: 大模型三维空间认知能力评测基准 2024–2025

构建自动化空间任务生成框架，覆盖旋转、折叠、视图合成等空间理解场景，并建立跨模型评测体系揭示多模态模型在空间认知能力上的差异；数据集与评测代码开放用于模型训练策略与可解释性研究。

LLaMA 系列大预言模型的 FPGA 端到端部署 2023–2024

面向资源受限的边缘智能场景，完成 LLaMA 模型在 Xilinx VHK 系列 FPGA 板卡上的端到端部署与推理加速。设计比特序列化矩阵乘单元，实现大模型在 FPGA 平台上的低比特量化稳定推理。

基于机器学习的高性能算术逻辑电路结构搜索与优化 2022–2023

针对乘法器的多目标优化问题，构建基于 Chisel 的自动化结构生成与评估框架，形成可扩展的架构性能数据集。利用图神经网络提取算术电路的结构特征实现帕累托最优预测，加速逻辑电路设计空间探索过程。

专业技能

编程语言	Python, C/C++, CUDA, Verilog, MATLAB, Chisel
框架工具	PyTorch, Git, Linux, Docker, CACTI, Ramulator, Vulkan
研究领域	三维视觉算法、神经渲染加速、体系结构设计、访存与算子分析、软硬件协同优化

研究方向

- 神经渲染算法
- 计算机体系结构
- 软硬件协同优化