Yixin Zhang

zyxcambridge@gmail.com

+86 17521398109

Shanghai, China

Algorithm Engineer

Work Experience

VLN Deployment Algorithm Engineer | Robotics Subsidiary of a Listed Company | Sep. 2025 - Present

- **Project:** Vision-Language Navigation (VLN) Algorithm Deployment and Autonomous Following Robot System
- **Core Algorithm Capabilities:**
  - **Long-range Planning:** Support navigation over 150 meters, enabling path planning in complex environments
  - **Zero-shot Generalization:** Achieve autonomous adaptation in unfamiliar environments, stable operation in new scenarios without pre-training
  - **Dense Obstacle Navigation:** Break through dense obstacle challenges, safely pass through extreme scenarios with obstacle spacing less than 50cm
  - **Dynamic Obstacle Avoidance:** Real-time perception and avoidance of moving obstacles, ensuring safe robot operation
- **Model Deployment & Optimization:**
  - Successfully deployed VLN models on NVIDIA Thor and Orin platforms, achieving end-to-end inference optimization
  - Significantly reduced latency through model quantization, operator fusion, memory optimization, and other technical means
  - Completed autonomous following robot functionality development, achieving stable and reliable following performance
  - Successfully completed 42 critical test verifications, system stability and reliability reached mass production standards
  - Embodied Intelligence Algorithm Research & Deployment: Achieved engineering deployment of embodied intelligence algorithms on NVIDIA Thor and Orin platforms
- **Project Achievements:**
  - Independently completed full-stack hardware-software integration from almost ”zero foundation”, including Thor hardware configuration, algorithm video tuning, and all related work
  - Successfully achieved stable following functionality, overcoming the most challenging technical difficulties in the project
  - Established complete deployment processes and verification standards, laying the foundation for subsequent large-scale applications
- **Dec. 18 Demo Rescue:**
  - In an emergency situation where the voice network completely crashed during the demo, leveraged Plan B and complete control environment backup
  - Coordinated the team on-site, stabilized team morale, ensuring the second demo passed successfully
  - Demonstrated excellent project control and emergency handling capabilities under high-pressure conditions
- **Project Planning & Team Building:**
  - Developed Thor project scale-up plan based on successful demo operation
  - Applied for team expansion, planned basic work division, focused on core technology breakthroughs

- Transitioned from ”solo operation” to ”group army charge”, pursuing greater technical breakthroughs and business value

Agent Algorithm R&D (Competition Period) | Freelance | Feb. 2025 - Sep. 2025

- **Competition Achievement:** NeurIPS 2025 Agent Tool-Augmented Reasoning Workshop - CureBench International Agent Evaluation Competition Global 2nd Place (Top 2)
- **Project Background:** Participated in NeurIPS 2025 top-tier conference Workshop, building professional Agent system in biomedical field
- **Technical Architecture:** CureBench + TxAgent RL Framework - ART Training
- **Core Work:**
  - Built medication assistant agent using Agent and Test-time Scaling technologies
  - Integrated biomedical tools (FDA, OpenTargets, PubMed) for agent tool-augmented reasoning
  - Completed initial draft of book ”Self-Evolving Agents - Architecture Practice of Dynamic Memory and Continuous Operation”

Deep Learning Algorithm Engineer | Aptiv Central Electrical (Shanghai) Co., Ltd. | Aug. 2024 - Feb. 2025

- **Project:** General Obstacle Perception Algorithm R&D and End-to-End Mass Production Pre-research
- **Model Deployment Work Hierarchy:**
  - Deploy a single network
  - Deploy multiple networks, optimize performance on a single chip:
    - 2 backbones + 3 heads, multi-task pipeline
    - Shared memory and queues, instance bank
  - Unified adaptation to multiple chip frameworks (GPU+ASIC+FPGA):
    - Multi-chip platform scheduling framework
    - One codebase adapts to multiple chips
- Participated in L2++ mapless end-to-end network design, led network structure design and optimization of general obstacle OCC branch
- Responsible for chip selection, completed benchmark testing of three chip types, designed end-to-end network multi-task scheduling framework
- Established deployment work standards SOP and workflow, independently completed AI model deployment of OD and OCC branches on 2 chip types
- **Asynchronous Scheduling:** Designed asynchronous pipeline, maximizing overlap between CPU and multiple AI acceleration cores (NPU/VP), reducing total completion time
- **Race to Idle:** Compressed 1000ms tasks to 28ms (14ms*2), enabling chips to enter low-power idle state faster
- **System-level Optimization - Multi-task Scheduling Framework:** Designed Pipeline orchestrating multiple model execution sequences in serial, parallel, or pipelined parallel modes
  - Stage A (Stage 1): Feature extraction center, equipped with two parallel machines computing Backbone_1 (BEV) and Backbone_2 (Temporal) respectively
  - Stage B (Stage 2): Perception analysis department, equipped with two parallel machines computing Head_1 (OD) and Head_2 (Map) respectively
  - Stage C (Stage 3): Decision fusion station, responsible for final Head_3 (Predict) computation
- **Deployment & Verification Workflow (Mass Production Standards):**
  - Model export & segmentation: Export PyTorch models to ONNX and segment into independent files by DAG nodes (backbone1.onnx, head1.onnx, etc.)
  - Model compilation: Use hardware vendor toolchains to compile ONNX to binary bin files, complete operator fusion and quantization optimization
  - Phased verification:
    - IO alignment: Ensure bit-by-bit consistency between deployment program and simulation script inputs
    - Single node verification: Compare bin file outputs with PC-side ONNX results

    - Multi-frame end-to-end alignment: Verify complete APP business metrics (mAP, IoU) consistency with golden reference model
  - **Final Integration & Delivery:** Encapsulated as 5 core APIs conforming to standards (Init/Run/Release/GetResult/GetStatus
  - Core optimization metrics: Time latency, throughput, memory bandwidth (reduced usage), power consumption (W4A4), compute utilization (software-hardware integration)
  - Adapted to automotive chips: NVIDIA Orin, Horizon J5/J6, CV3

Model Deployment Lead / Senior Software Architect | Innovusion Intelligent Technology (Shanghai) Co., Ltd. | Jun. 2022 - Apr. 2024

- Recruited and built efficient deep learning deployment team, improved algorithm engineering implementation processes
- Achieved real-time AI LiDAR model inference on Jetson platform, reduced latency by 4x
- Developed ADAS code completion VSCode plugin, improved autonomous driving algorithm development efficiency
- Explored large model inference deployment, successfully deployed Ollama, llama.cpp, vllm, TensorRT-LLM, mlc-llm and other frameworks
- Successfully ran yi-34B model on Mac M1, ran llama3-8b model on Android platform using mlc-llm
- Improved MLOps process based on NVIDIA Drive Sim, generated 100K frame simulation dataset
- Achieved joint training of simulation and real data, improved accuracy by 10 percentage points
- **[Model Deployment]**
  - Performance analysis tool application: Used Nsight Systems and Nsight Compute to locate deployment bottlenecks, solved critical performance issues
  - LiDAR vehicle-side framework construction: Achieved LiDAR point cloud detection and semantic segmentation network deployment iteration, including rangeimage and pillar voxel network real-time inference, designed end-to-end unified detection-segmentation framework
  - NVIDIA technical collaboration: Established direct communication channel with Jetson team, analyzed model bottlenecks and obtained optimization support, reduced overall latency by 75% (one quarter) through task dependency optimization
  - Horizon ecosystem integration: Secured hardware and software development support for Horizon J5 platform, completed full perception algorithm migration
  - Technical challenge breakthrough: Collaborated with SPConv core development team to solve custom sparse convolution operator deployment challenges

Senior Algorithm Engineer | Shanghai Xuehu Technology Co., Ltd. | May 2019 - May 2022

- Built algorithm team from scratch, constructed V2X perception algorithm engineering implementation process
- Designed FPGA hybrid quantization scheme, refined to multiplication and addition operation levels
- Successfully mass-produced 100+ MEC devices, achieved 5M+ RMB in software-hardware revenue, served nearly 10 customers
- **[Challenges & Solutions]**
  - Accuracy loss issue: Proposed hybrid quantization strategy for 30-point accuracy loss caused by quantization
  - Quantization scheme: Implemented PTQ (Post-training Quantization), referenced QAT (Quantization-aware Training) full process
  - Quantization algorithm comparison: Systematically evaluated error distribution characteristics of minmax, KL divergence, and MSE quantization methods
  - Tool development: Independently developed accuracy comparison tool, visualized error heatmaps under different quantization parameters
  - Training-inference consistency: Solved training/inference accuracy deviation from quantization principle formulas
  - Customer trust building: Resolved latency concerns through theoretical derivation and measured data dual verification
- **[Implementation Method]**

  - Core optimization flow: Build hash table → Generate Rulebook → Gather input features into dense matrix M_in → Execute GEMM → Scatter output results

- **[LiDAR MLOPS]**
  - Layered strategy: Key feature extraction layers retain floating-point precision, classification head and post-processing layers quantized to INT8
  - Quantization algorithm: Adopted KL divergence calibration method, used calibration set for data distribution matching
  - Network decomposition: Decomposed complex convolution operations into basic multiply-add operations, adapted to FPGA integer computation architecture
  - Built deep learning deployment process based on ZCU104 IP domain controller, achieved 10x speed improvement
- **[Technical Implementation]**
  - Created V2X solution documents based on projects and existing company solutions, provided on-site explanations, acquired 10+ customers
  - Conducted on-site development at project locations, implemented 10+ pilot projects in various cities across China
  - Promoted FPGA LiDAR acceleration IP adaptation to RoboSense, Hesai, Ouster, Livox and other LiDAR companies
- **[Project Revenue Proof]**
  - Mass-produced 100+ MEC devices based on AMD Xilinx ZCU104, achieved 2M+ RMB in software-hardware revenue
  - Maintained nearly 10 customers with long-term cooperative relationships

Algorithm Engineer | DeepBlue Technology (Shanghai) Co., Ltd. | May 2018 - Feb. 2019

- **Reporting to:** Technical Manager
- **Keypoint-based Object Detection Algorithm Engineering:**
  - Implemented OpenPose-based human keypoint detection algorithm, independently completed C++ inference pipeline
  - Algorithm improvements included Gaussian response enhancement, Heatmap radius adjustment, and intermediate supervision mechanism
  - Innovatively implemented peak-based NMS method replacing traditional IoU approach
- **Model Deployment & Optimization:**
  - Deployed models on P100 GPU, single card supported 28 models for parallel inference
  - Optimized OpenPose inference latency to within 500ms
- **Algorithm Validation:**
  - Built dense scene test set, achieved 99% keypoint recognition accuracy in complex environments
- **Other Achievements:**
  - Obtained 3 invention patent authorizations
  - Developed cloud-based product recognition service based on keypoints

Deep Learning Engineer | Youbang Network Technology Co., Ltd. | Dec. 2017 - Apr. 2018

- **Reporting to:** Technical Manager
- **ADAS System Development:**
  - Developed pedestrian and vehicle object detection algorithms, combined with image segmentation technology to achieve lane detection
  - Optimized networks on embedded platforms such as RK3399, significantly improved detection speed
  - Used TensorFlow Lite for Android model migration and quantization optimization
- **Human Pose Recognition:**
  - Implemented lightweight human keypoint detection based on MobileNet architecture

- Completed algorithm implementation using Caffe2 and TensorFlow frameworks respectively
- Successfully ported to Android devices, supporting real-time inference

---

Education

Bachelor of Network Engineering | North China Institute of Aerospace Engineering | Sep. 2010 - Jun. 2014

Highlights & Publications

Publication:

- Authored book ”Self-Evolving Agents - Architecture Practice of Dynamic Memory and Continuous Operation”, currently completed initial draft and entered publisher’s three-review and three-proofreading stage, planned to be published by Publishing House of Electronics Industry

Competition Achievement:

- NeurIPS 2025 Agent Tool-Augmented Reasoning Workshop - CureBench International Agent Evaluation Competition Global 2nd Place (Top 2)

Selected Projects

[Project: General Obstacle Perception Algorithm R&D and End-to-End Mass Production Pre-research]

- **Project Goal:** Develop general obstacle detection algorithm, achieve open-scene generalized perception based on physical laws (depth distribution and semantic constraints), pre-research end-to-end deployment solution for automotive-grade low-power platform (16W)
- **Core Technical Breakthroughs:**
  - Paradigm shift: From ’learning object shapes’ to ’verifying physical laws’, achieving ’perceiving existence without recognizing shapes’ generalization capability through depth distribution modeling and semantic constraint verification
  - Scene optimization: Enhanced vertical edge gradient constraints in building areas, reduced false detection risk of suspended objects
  - Negative obstacle detection: Utilized geometric features such as height mutations and point density anomalies to achieve detection of negative obstacles like potholes and ditches
- **Core Challenges & Solutions:**
  - Challenge: Contradiction between end-to-end perception (OD+Map+Predict+Plan) computation and 16W power consumption limit
  - Solutions:
    - 1) Model lightweighting: Optimized OD and Map model parameter structures
    - 2) Hybrid quantization strategy: FP16+INT8 hybrid quantization, combined offline initialization with online calibration, accuracy loss $<2\%$
    - 3) Custom operators: Developed key operators such as voxelization and deformable aggregation, improved computational efficiency
- **Key Results:**

  - End-to-end inference speed improved 10x, successfully deployed to target platform
  - Hybrid quantization method applied for technical patent
  - Verified feasibility of physics-based perception paradigm in open scenes
- **Forward-looking Research:**
  - Solutions for gradient mismatch issues in mixed precision training
  - Operator fusion solutions under edge-side memory bandwidth constraints

[Project: LiDAR Inference Framework]

- **Challenges & Solutions**
  - **Accuracy Loss Control:** Proposed hybrid quantization strategy for 30-point accuracy loss caused by quantization
  - **Quantization Scheme Implementation:** Built PTQ/QAT full process, supported minmax, KL divergence, MSE three calibration algorithms
  - **Error Analysis Tool:** Independently developed accuracy comparison platform, visualized error heatmaps under different quantization parameters
  - **Training-Inference Consistency:** Solved floating-point/integer computation deviation from quantization principle formulas
  - **Customer Trust Building:** Reduced latency to 1/3 of customer requirements through theoretical derivation + measured data dual verification
- **Technical Implementation**
  - Core optimization flow: Build hash table → Generate Rulebook → Gather input features → Execute GEMM → Scatter output results
  - Quantization algorithm comparison: Systematically evaluated error distribution of three methods across different layers, selected optimal combination strategy
- **Business Value**
  - Mass-produced 100+ MEC devices based on AMD Xilinx ZCU104
  - Achieved 2M+ RMB in software-hardware revenue, served nearly 10 customers
  - Mass production for Shaanxi Heavy Duty Automobile

[Project: NeurIPS 2025 CureBench Medical Agent System]

- **Project Role:** NeurIPS 2025 Workshop Project Lead / Core Developer
- **Project Background:** Participated in NeurIPS 2025 top-tier conference Workshop, built professional Agent system in biomedical field. Utilized Agent and Test-time Scaling technologies to build medication assistant agent.
- **Technical Architecture:** CureBench + TxAgent RL Framework - ART Training
- **Core Technologies:**
  - Integrated biomedical tools (FDA, OpenTargets, PubMed) for agent tool-augmented reasoning
  - Implemented Test-time Scaling technology to improve agent performance during testing phase
  - Conducted agent training and optimization based on reinforcement learning framework (TxAgent RL Framework)
- **Project Results:**
  - Achieved global 2nd place (Top 2) in CureBench International Agent Evaluation Competition
  - Successfully built practical biomedical field agent system

Large Model Fine-tuning & Deployment Projects

- Deployed Baichuan2 model on AWS as OpenAI-compatible format, achieved LoRA fine-tuning
- Used LLaMA-Factory on Alibaba Cloud platform for large model fine-tuning and optimization
- Deployed Start Code model to HuggingFace as API service
- Optimized 8B model performance to equivalent of 671B model level

Academic & Competition Experience

- 2025: NeurIPS 2025 Agent Tool-Augmented Reasoning Workshop - CureBench International Agent Evaluation Competition Global 2nd Place (Top 2)
- Oct. 2021: CVPR Workshop - 3D Object Detection Algorithm (9th Place)
- Multiple Hackathon Awards:
  - Oct. 2023: Baichuan Hackathon - US Traditional Chinese Medicine Customer Acquisition Agent Project (Special Award)
  - Sep. 2023: Google I/O Hackathon - Stable Diffusion Computing Sharing (3rd Place)
  - Jul. 2023: World AI Hackathon - Leadership Tracking & Training (2nd Place)
  - Sep. 2019: AngelHack Shanghai - Leadership Tracking & Training (1st Place)
  - Jul. 2017: Global AI Hackathon - Fake News Detection (Champion)

Technical Community Involvement

- Google Machine Learning Developer Expert, conducted 4 technical lectures annually for 5 consecutive years, impacted 11,874 people