Yiqiao Qiu

yiqiaoqiu@hotmail.com | 341-732-9006 | LinkedIn | Google Scholar

Computer Vision / Machine Learning engineer with production deployment experience and massive-scale distributed systems engineering background and 7 CV / ML research papers published and 110+ citations.

SKILLS

- Machine Learning / Deep Learning: Efficient Industrial Model Optimization, Continual learning, Model Distillation, Transfer Learning, Supervised / Semi-Supervised Learning, VLM LoRA finetuning
- Computer Vision: Semantic Segmentation, Classification, Object Detection, Super-Resolution, Facial-Landmark Detection, Scene Understanding, Visual Question Answering, Anomaly Detection
- Frameworks: PyTorch, ONNX, NVIDIA DALI, Rust Tokio, AWS (DynamoDB, S3, CloudWatch, CloudFront)
- Programming Languages: Python, Rust, C/C++, Java, Kotlin, MySQL

WORK EXPERIENCE

Amazon Web Services Santa Clara, CA

Software Engineer, AWS DC Network Infra, Scalable Intent-Driven Routing (SIDR) **(Rust, Python, Tokio)** Apr 2024 - Present

Massive-scale distributed system development and operation for AWS Datacenter network fabrics routing control plane

- Built an automated release qualification framework with concurrent workflow orchestration and chaos fault injection to AWS ML fabrics, achieving **15x release speedup**.
- Implemented SIDR daemon logic for route redistribution from Quagga, including inter-process communication, message stream parsing, multi-module **asynchronous programming and OS-signaling**.
- Implemented SIDR security enhancement through message certificate based authentication and verification mechanisms.

XPeng Motors San Diego, CA

Computer Vision Engineer Intern, Autonomous Driving Center **(Python, PyTorch, DALI, ONNX)** Oct 2023 - Mar 2024

- **Training Pipeline Acceleration for Large-scale Perception Models.** Accelerated on-car perception model training by integrating NVIDIA DALI for GPU-based online augmentation on huge-scale image datasets, offloading preprocessing from CPU to GPU via multi-process pipelines; achieved **7x training speedup** and **80% CPU resource reduction**, unblocking faster iteration across perception teams.
- **Multi-task Backbone Consolidation for On-car Deployment.** Merged multiple task-specific models into a unified shared backbone, systematically exploring trade-offs across architectures, FLOPs, and cross-task generalization; reduced on-car model scheduling and memory overhead while preserving per-task accuracy, improving deployment efficiency on resource-constrained automotive compute.
- **Eye-Action Video Classification for Driver Monitoring System (DMS).** Owned end-to-end development of the eye-action recognition pipeline for in-cabin fatigue and distraction detection — covering dataset construction, temporal model design and training, and in-vehicle real-scene validation under varied lighting and head poses. Achieved **99.64% binary classification accuracy** with **30% inference latency reduction** to meet real-time on-car constraints.
- **Simulation-driven Data Augmentation for Long-tail Object Detection.** Replenished an object detection dataset with photorealistic simulation data for rare/long-tail categories; validated the pipeline by training YOLO-X on the augmented dataset and demonstrating consistent mAP gains on underrepresented classes. **Co-author of *Anything in Any Scene*.**

ByteDance Shenzhen, China

Video Algorithms Engineer Intern, Real-Time Communication, Video Group **(Python, PyTorch)** Nov 2021 - Apr 2022

- **Real-time Multi-frame Super-Resolution for Live-streaming Codec Enhancement:** Proposed novel auxiliary modules at the low-level encoder/decoder of a real-time multi-frame Super-Resolution model, leveraging temporal consistency across adjacent frames and residual-aware feature fusion to alleviate video decoding blocky artifacts from aggressive compression without inflating inference latency; achieved **43% PSNR gain improvement** in offline testing, deployed into TikTok live-streaming RTC video engine.
- **Robust Facial Landmark Detection for ROI-aware Bitrate Allocation Livestreaming:** Optimized the landmark model driving ROI-based bitrate allocation on streamers' faces via (1) facial parsing preprocessing for semantic priors, (2) weighted loss with balanced resampling for long-tail poses/occlusions, and (3) an auxiliary global-context branch to stabilize predictions under occlusion and motion blur. Reduced **NME loss by 67%**; unstructured pruning further cut inference time by **20%** to meet RTC latency budget.
- **Training Infrastructure.** Built FFmpeg-based offline augmentation simulating codec artifacts, bitrate ladders, and frame-drop patterns, plus a multi-threaded concurrent I/O queue to hide I/O latency behind GPU compute; reduced training time on large-scale video datasets by **40%**.

DMAI Guangzhou, China

Computer Vision Engineer Intern, DMAI Research Center **(Python, PyTorch)** July 2021 - Oct 2021

- AILA Preschool Learning System Card Recognition: Benchmarked and optimized **lightweight object detection** models (RFB, YOLO-X, YOLO-v5) with data augmentation search and loss tuning to suppress false positives, achieving **99.5% mAP**. Improved classification with open-set loss to resolve **95%** edge-case failures at **99%** precision.

PROJECTS and ML PUBLICATIONS

Efficient Vision-Language Models: Training, Model Distillation, Token Compression, and Deployment
PyTorch | Vision-Language Models | LoRA | GGUF Quantization | llama.cpp Feb 2026 - Present

- Proposed **region-aware self-attention distillation (SATS-CRP)** for VLM knowledge distillation (Qwen2.5-VL 32B→3B, LoRA): **distilled class-region pooled self-attention scores of visual tokens within LLM decoder layers**, provided meaningful denoised visual inter-region relationships distillation signal, SATS & LLaVA-KD baseline combined distillation loss method achieved **1%** over baseline method on distilled DriveLM LoRA accuracy
- Fine-tuned Qwen2.5-VL 3B VLM with LoRA (rank 16) on DriveLM-nuScenes; integrated 4 VLM **visual token compression** methods (FasterVLM, PruMerge, PyramidDrop, SATS-CRP) between the vision encoder and LLM, achieving **4× visual token reduction (480→120)** while DriveLM LoRA accuracy **maintained or improved 1%**.
- Systematically profiled the accuracy–compression tradeoff of FasterVLM and SATS-CRP across 4×–16× ratios; **16× extreme compression (480→30 tokens)** is achieved with 2.4% accuracy degradation.
- Quantized the fine-tuned model from BF16 to GGUF Q4_K_M (**3.9**× size reduction) and deployed on a consumer RTX 4070Ti via llama.cpp, achieving **170** tokens/s throughput and **142 ms** time-to-first-token

SATS: Self-Attention Transfer for Continual Semantic Segmentation Sep 2021 – Feb 2023
PyTorch | Vision Transformers | Semantic Segmentation | Knowledge Distillation

- Proposed **self-attention** distillation of visual patch relationships as a lightweight, plug-in technique for **transferring inter-patch relationships in any visual transformer models for knowledge-distillation**, applicable in both continual learning and distillation based visual encoder compression; extended and showed effectiveness to **VLM distillation** (see above project)
- Achieved **state-of-the-art** performance in Continual Learning in Semantic Segmentation and **published in Pattern Recognition** as **1st author**. *SATS: Self-Attention Transfer for Continual Semantic Segmentation*, **53** citations

Agentic RAG System over SQL + Document Corpora Apr 2026
Python | LLM Function Calling | asyncio | Hybrid Retrieval (BM25 + Dense) | SSE Streaming

- Architected an **agentic RAG system** with an LLM orchestrator delegating to specialized **SQL and document sub-agents** via OpenAI-style **function calling** and a typed *Evidence* protocol, enabling **deterministic post-hoc routing fallback** without extra LLM round-trips; validated at **full-pass correctness on a 10-question gold-labeled dev set**.
- Implemented **asynchronous parallel sub-agent dispatch** (`asyncio.gather`) with per-agent timeouts and **graceful partial-result synthesis**; **SSE event streaming** surfaces routing decisions, tool calls, and sub-agent reasoning live to the chat UI.
- Built a **hybrid retrieval pipeline** combining dense embeddings and BM25 sparse search via **Reciprocal Rank Fusion** over **∼900 chunks from 6 SEC 10-K filings**, with a metadata-only section-fetch shortcut for targeted document lookups.
- Designed a **4-mode evaluation harness** (fuzzy-numeric / entity-match / LLM-as-judge / deterministic slot-based component recall) targeting list-coverage silent failures invisible to judge-only scoring.

OTHER RESEARCH PUBLICATIONS *(Sept 2020 - Present) Total Citations: 110*

- Cooperatively designed a novel method in Visual Out-Of-Distribution Detection, revised the full paper. Published in **IEEE Transactions on Circuits and Systems for Video Technology** as **2nd author**, 22 citations. *Classifier-head Informed Feature Masking and Prototype-based Logit Smoothing for Out-of-Distribution Detection*,Link
- Reproduce four baselines for cross-domain text sentiment classification. Published in **Information Processing and Management** as **2nd author**. *Topic Driven Adaptive Network for Cross-Domain Sentiment Classification*,**Link**, 23 citations
- Co-author of *Anything in Any Scene: Photorealistic Video Object Insertion*,Link, 300+ GitHub stars, 11 citations.
- Revised full paper for uncertainty estimation in medical image classification. Published in **MICCAI 2024** as **2nd author**. *Deep Model Reference: Simple Yet Effective Confidence Estimation for Image Classification*,Link
- Revised full paper for class incremental learning with OOD detection, Published in **Neurocomputing Journal** as **3rd author**. *Class Incremental Learning with Task-Specific Batch Normalization and Out-of-Distribution Detection*,Link, 1 citation
- Revised full paper on local background features for OOD detection, under review at Neural Computation as **3rd author**. *Local Background Features Matter in Out-of-Distribution Detection*,Link

EDUCATION

University of California, San Diego La Jolla, CA
Master of Science in Computer Science and Engineering GPA: 3.93/4.0 Sept 2022 – Mar 2024

Sun Yat-sen University Guangzhou, China
Bachelor of Engineering in Computer Science Major GPA: 3.94/4.0 (top 10%) GPA: 3.8/4.0 Sept 2018 - Jun 2022